

Evaluating different statistical regression models for industrial energy measurement and verification

I Schuin

 [orcid.org/ 0000-0002-2806-524X](https://orcid.org/0000-0002-2806-524X)

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

Supervisor: Dr J Vosloo

Graduation: May 2020

Student number: 25015869

ABSTRACT

Title: Evaluating different statistical regression models for industrial measurement and verification

Author: I Schuin

Supervisor: Dr JC Vosloo

Keywords: Baseline models; energy efficiency; machine learning; measurement and verification; model evaluation procedure; statistical regression models; model-averaging

The worldwide concern regarding the environmental impact of using fossil fuels as an energy source leads to the increased interest in improving energy efficiency. The energy efficiency of existing facilities can be improved through the implementation of energy saving measures (ESMs). However, to convince different stakeholders involved of the impact of ESMs, the savings need to be quantified accurately. This falls under the field of measurement and verification (M&V).

An important part of the M&V process is the construction of a model to predict energy usage. Different model types are available that fall within different categories. Each of these categories provides different advantages and disadvantages. The black-box based approach of using statistical regression models (SRMs), can offer superior accuracy and ease of use. Also, combining different SRMs through model-averaging can provide even greater accuracy than a single model.

Many studies have focused on evaluating different types of SRMs for energy modelling purposes. These include simple models, such as the well-known linear model, but also the more complex machine learning models. Some studies have also focused on specifically applying different types of SRMs towards M&V.

However, studies tend not to give holistic consideration to the fundamental principles when evaluating models for M&V. Furthermore, there has been minimal evaluation of different types of SRMs for use of M&V within the industrial sector, with only a few isolated case studies reported in literature. There is thus a need for a comprehensive evaluation of different types of SRMs for M&V use within the industrial sector.

This dissertation evaluated four different types SRMs for accuracy and consistency in the M&V process, by applying it to five industrial case studies. The models included a simple linear model and three types of machine learning models. Two model-averaging techniques (simple and weighted) were applied to the four single models to form two new models.

A comprehensive evaluation procedure was developed based on well-known methods and techniques in literature. Different aspects of accuracy and consistency were considered in the procedure, including absolute error, bias, sensitivity to the data sampling process and sensitivity to drivers used. The procedure also accounted for additional complexities introduced by the machine learning models. A scoring technique was applied to evaluate the performance of model types across the five case studies.

The results obtained are verified based on what is expected from literature. The study shows that no specific model type stands out with differences changing from case study to case study. Savings and errors of the different model types tend to fall within the same range. The specific energy drivers used, however, tend to influence the predicted savings and accuracy more than the type of model. This indicates that great care should be taken when considering energy drivers while the different model types can thus be used as verification and to identify anomalies.

ACKNOWLEDGEMENTS

I would like to thank Prof Edward Mathews and Prof Marius Kleingeld for the chance to further my studies at the Centre for Research and Continued Engineering Development (CRCED). I have learnt valuable lessons that I hope to pass on to others.

This dissertation would not have been possible without financing and support received from ETA Operations (Pty) Ltd and its sister companies. A big thank you to everyone involved.

My sincerest gratitude to my study leader, Dr Jan Vosloo. Your leadership has not only influenced this study, but also me.

A special thanks to Dr Andries Gous. Your unceasing willingness to provide guidance and support with this dissertation has been invaluable during the completion of this study.

Thank you to Dr Walter Booysen and Dr Waldt Hamer for their inputs towards this study.

I would like to express my gratitude to all who helped and supported me over the past two years at CRCED, mentioned and unmentioned. It has been an unforgettable experience.

A big thank you to my family for their love and understanding when times were tough. May God richly bless each one of you.

Lastly and most importantly, I would like to thank God for His love and guidance in my life - without you Lord, nothing is possible.

*“Remember them which have the rule over you, who have spoken unto you the word of God:
whose faith follow, considering the end of their conversation.*

Jesus Christ the same yesterday, and to day, and for ever.”

– Hebrews 13:7-8 (KJV)

TABLE OF CONTENTS

ABSTRACT...	I
ACKNOWLEDGEMENTS	III
LIST OF TABLES	VII
LIST OF FIGURES.....	VIII
LIST OF ABBREVIATIONS	XI
LIST OF SYMBOLS.....	XIII
GLOSSARY.....	XV
CHAPTER 1 – INTRODUCTION.....	1
1.1 Quantification of energy efficiency	1
1.2 The M&V process for energy savings calculation.....	2
1.3 Modelling options.....	8
1.4 Statistical regression models	10
1.5 Model evaluation	12
1.6 Model selection.....	13
1.7 Hybrid and ensemble models	18
1.8 Objectives.....	19
1.9 Scope of investigation.....	20

CHAPTER 2 METHODOLOGY FOR MODEL EVALUATION.....	21
2.1 Introduction to general evaluation procedure	21
2.2 Step 1: Data gathering and M&V process parameter selection	22
2.3 Step 2: Data treatment.....	25
2.4 Step 3: Selection of driver sets	28
2.5 Step 4: Selection of hyperparameters.....	32
2.6 Step 5: Model training.....	35
2.7 Step 6: Model evaluation	37
2.8 Step 7: Model-scoring and interpretation of results	45
2.9 Final implementation and summary of procedure	46
2.10 Conclusion of evaluation procedure.....	51
CHAPTER 3 DISCUSSION OF RESULTS AND VERIFICATION	52
3.1 Introduction to results	52
3.2 Selected case studies.....	52
3.3 Removal of outliers.....	62
3.4 Selected driver sets	65
3.5 Savings, errors and consistencies obtained.....	67
3.6 Results of model-scoring	80
3.7 Implications of the results for M&V.....	85
3.8 Conclusion of results	87

CHAPTER 4 CONCLUSION	88
4.1 Summary of study.....	88
4.2 Limitations and recommendations	90
REFERENCES.....	92
APPENDIX A – ADDITIONAL DETAIL REGARDING MODELS	108
APPENDIX B – MODEL CONSTRUCTION & METRIC CALCULATION.....	120
APPENDIX C – ADDITIONAL RESULTS	123

LIST OF TABLES

Table 1.1: Principles of the M&V process	7
Table 2.1: Considerations for choosing the baseline and the reporting period	23
Table 2.2: Different hyperparameter alternatives for each model type	34
Table 2.3: Scoring categories and metrics for final evaluation	46
Table 3.1: Summary of case study I	57
Table 3.2: Summary of case study II	58
Table 3.3: Summary of case study III	59
Table 3.4: Summary of case study IV	61
Table 3.5: Summary of case study V	62
Table 3.6: Selection methods for the different driver sets	65
Table 3.7: Driver sets for each case study	66
Table 3.8: Scoring table for case study I	81
Table 3.9: Scoring table for case study II	81
Table 3.10: Scoring table for case study III	82
Table 3.11: Scoring table for case study IV	82
Table 3.12: Scoring table for case study V	83
Table 3.13: Summary of category winners for each case study	83

LIST OF FIGURES

Figure 1.1: The M&V process	2
Figure 1.2: Different measurement boundary options	4
Figure 1.3: The three periods of interest during M&V	4
Figure 1.4: Quantifying energy savings	5
Figure 1.5: Different measurement frequencies for the same data	6
Figure 1.6: A hypothetical process, X, with different energy carriers and drivers	8
Figure 1.7: Different types of SRMs.....	11
Figure 1.8: Illustration of fitting a linear model	15
Figure 1.9: The structure or <i>architecture</i> of a simple ANN	16
Figure 1.10: Non-linear SVM	17
Figure 1.11: Illustration of KNN for $k = 5$	17
Figure 1.12: Illustration of the working principles of (a) hybrid; and (b) ensemble models.....	18
Figure 2.1: Illustration of the once-off and iterative parts of the process	22
Figure 2.2: Flow diagram for step 1	22
Figure 2.3: Illustration of taking a single variable from a higher measurement frequency to a lower measurement frequency	24
Figure 2.4: Example of (a) time-series plot and (b) scatter plot of energy usage vs a potential driver	25
Figure 2.5: Illustration of the data treatment step.....	26
Figure 2.6: Inspection plots for outlier removal; (a) Time-series plot with flagged points and (b) scatter plot with flagged points and points selected for removal.....	27
Figure 2.7: Illustration of the distinction between inputs and driver sets.....	28
Figure 2.8: Conceptual illustration of PCA	29
Figure 2.9: Illustration of the four driver set selection options	31
Figure 2.10: Illustration of grid-search through combinations of hyperparameters for KNN.....	32

Figure 2.11: K-fold cross-validation illustration for $k = 5$	33
Figure 2.12: Illustration of the hyperparameter choice step	34
Figure 2.13: Example of overfitting of data with a fourth-order order polynomial in contrast to the true trend given by the linear model	35
Figure 2.14: The iterative sampling process to form a training and testing set.....	36
Figure 2.15: Example of the residuals of data leading to (a) high CV(RMSE) and low NMBE and (b) low CV(RMSE) and high NMBE.....	38
Figure 2.16: Example of a density plot showing a normal distribution	39
Figure 2.17: The concept of back-casting to determine savings	41
Figure 2.18: An example of a bar plot comparing the forward and backward-calculated savings for both approaches for each driver set of each of the different model types.....	42
Figure 2.19: A generic example of a density plot comparing the distributions for the different driver sets for each model type	43
Figure 2.20: Scatter plots for the evaluation step.....	44
Figure 2.21: Part 1 of the methodology.....	47
Figure 2.22: Part 2 of the methodology.....	49
Figure 3.1: Process diagram for Facility A	53
Figure 3.2: Average product distribution over the two years	53
Figure 3.3: Time -series plot of energy usage (scaled)	54
Figure 3.4: Time-series plot of the production (scaled)	55
Figure 3.5: Scatter plots of energy usage vs the various drivers.....	56
Figure 3.6: System layout for Facility B	57
Figure 3.7: Representation of a typical furnace for industrial production facility C.....	59
Figure 3.8: Representation of a single chiller system for mine A.....	60
Figure 3.9: The compressor network for mine B	61
Figure 3.10: Time-series plot of the total production showing the DSOD outliers.....	62

Figure 3.11: Scatter plot of the energy usage vs total production showing the DSOD outliers for case study III	63
Figure 3.12: Time-series plot with DSOD outliers for the total ore	64
Figure 3.13: Scatter plot of the energy usage vs the ore usage showing the DSOD outliers and the point removed	64
Figure 3.14: The relationship between (a) the energy produced and the energy supplied for case study II and (b) the energy usage and the evaporator duty for case study IV	66
Figure 3.15: The savings calculated for case study I (daily)	68
Figure 3.16: Weights for the WMA model for case study I (daily)	70
Figure 3.17: The savings calculated for case study I (weekly)	70
Figure 3.18: The savings calculated for case study I (monthly)	71
Figure 3.19: Data distribution for the total production for Year 1 and Year 2	72
Figure 3.20: Savings density distribution for case study I (daily)	73
Figure 3.21: Density plot for the CV(RMSE) for the sensitivity repetitions for case study I (daily)	74
Figure 3.22: Density plot for the CV(RMSE) for the sensitivity repetitions for case study I (monthly)	75
Figure 3.23: Density plot for the NMBE for the sensitivity repetitions for case study I (daily) ...	76
Figure 3.24: NMBE vs CV(RMSE) for the forward models when using all the training data	77
Figure 3.25: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages)	78
Figure 3.26: Difference between the all-data and sample-data approach for the KNN prediction	78
Figure 3.27: The density plot of the CV(RMSE) for case study II	80

LIST OF ABBREVIATIONS

Abbreviation	Description
ANN	Artificial neural network
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BFW	Boiler feedwater
CV(RMSE) ¹	Coefficient of variation of root-mean squared error
DSOD	Dynamic sequential outlier detection
EEM / ECM / ESM	Energy efficiency measure / energy conservation measure / energy saving measure
EKF	Element-wise k -fold (EKF)
ESCO	Energy service company
EVO	Efficiency Valuation Organisation
IPMVP	International performance measurement and verification protocol
KNN	k -Nearest neighbours
LIN	Linear regression
MAPE	Mean absolute percentage error
M&V	Measurement and verification
NMBE ¹	Normalised mean bias error
NNLS	Non-negative least squares
OLS	Ordinary least squares
PCA	Principal component analysis
RMSE	Root-mean squared error
RSA	Republic of South Africa

¹Refer to glossary for definition

SABS	South African Bureau of Standards
SANS	South African National Standards
SMA	Simple model-averaging
SRM ²	Statistical regression model
SVM / SVR	Support vector machine / support vector regression
USA	United States of America
WMA	Weighted model-averaging

²Refer to glossary for definition

LIST OF SYMBOLS

Symbols

A	Analyser
E	Energy used
$\hat{E}$	Predicted energy use
E^-	Energy recovered
F	Volume flowrate
J	Power
k	An integer value
M	Mass flowrate
M	Number of model types
N	Number of data points or periods or repetitions
S	Number of driver sets
T	Temperature
x	Independent/input variable(s)
x'	Derived independent/input variable(s)
y	Dependent/output variable(s)
$\hat{y}$	Prediction of the dependent variable
$\bar{y}$	Mean y -value
z	Generic variable
w	Weight(s)

Greek letters

ε	Error value
λ	Constant parameter of the kernel functions

Subscripts

i	The i -th input variable
j	The j -th period
$true$	The true value
pre	The predicted value
sav	Calculated savings

GLOSSARY

“CV(RMSE)” is an error metric that is a measure of the absolute error between the true and predicted values.

“Driver set” means those variable(s) chosen to form the independent variables in the model for energy usage prediction.

“Hyperparameter” refers to a selection that must be made before model construction that influences the functional form of the model but is not calculated by the regression process.

“NMBE” is an error metric that is a measure of the bias between the true and predicted values.

“Observation” or data point means a singular instance that includes all the relevant variables measured together at that point in time.

“Regression” means the process of calculating the parameters or constants or weights of a model based on a dataset with independent and dependent variables.

“Statistical regression model” refers to any model constructed through the process of regression.

“Training” with respect to models refers to the application of the regression process or model construction.

“Training dataset” refers to the dataset used to train the model of which all the observations influence the final model constructed.

“Validation dataset” means the dataset used to test the model of which the observations were not included in the training dataset.

CHAPTER 1 – INTRODUCTION

1.1 Quantification of energy efficiency

The worldwide energy consumption is increasing due to population growth and economic development [1], [2]. This leads to an increasing scarcity of fossil fuels, the most used energy source [3]–[5]. There is also increased pressure from various legislative bodies to decrease the use of fossil fuels due to the environmental damage associated with its use [6]–[8].

Energy efficiency is seen as one of the most promising ways to improve environmental sustainability by decreasing energy usage [9]–[11]. There is a need not only to ensure energy efficiency with the construction of new facilities (greenfield projects), but also to improve the energy efficiency of existing facilities (brown field projects). These outcomes can be reached through the implementation of energy saving measures (ESMs) [12], also referred to as energy conservation measures (ECMs) [13], [14], energy efficiency measures (EEMs) [9] or interventions.

The implementation of ESMs often requires capital to be invested [10], [15], [16]. Convincing stakeholders of the potential positive impacts of ESMs could ease the process of obtaining the necessary capital [17]–[19]. To this end stakeholders can be provided with the necessary information to objectively evaluate an ESM, including savings that are quantified in a reliable and assured manner [20], [21].

Savings achieved is not only an indicator for investors, but also to those implementing the projects, such as energy services companies (ESCOs). This is especially of concern when the remuneration of such companies is dependent on the savings achieved [20], [22]. There is therefore a need to quantify savings that are accurate with a low degree of uncertainty.

Unfortunately, energy efficiency savings cannot be directly measured and therefore mathematical and other calculation techniques are needed to quantify energy efficiency savings [13], [23]. Quantifying the savings achieved by an ESM or the “avoided energy use” is a subject of the field of measurement and verification (M&V) [19]. M&V forms an integral part of energy management [24] and has become a well-studied field [25]. Rigorously applying the fundamental principles of in the M&V process could provide the confidence required by various stakeholders [26].

1.2 The M&V process for energy savings calculation

Many protocols and standards have been set up to describe the basic principles of M&V and ensure that a standard process is followed when performing M&V analyses. These include:

- The International Performance Measurement and Verification Protocol (IPMVP) [27]
- The American Society of Heating, Refrigeration, and Air-Conditioning Engineers' (ASHRAE's) Guideline 14 [28]
- ISO 50015-2014 [29]
- The South African National Standard (SANS) 50010 – Measurement and verification of energy savings [30]

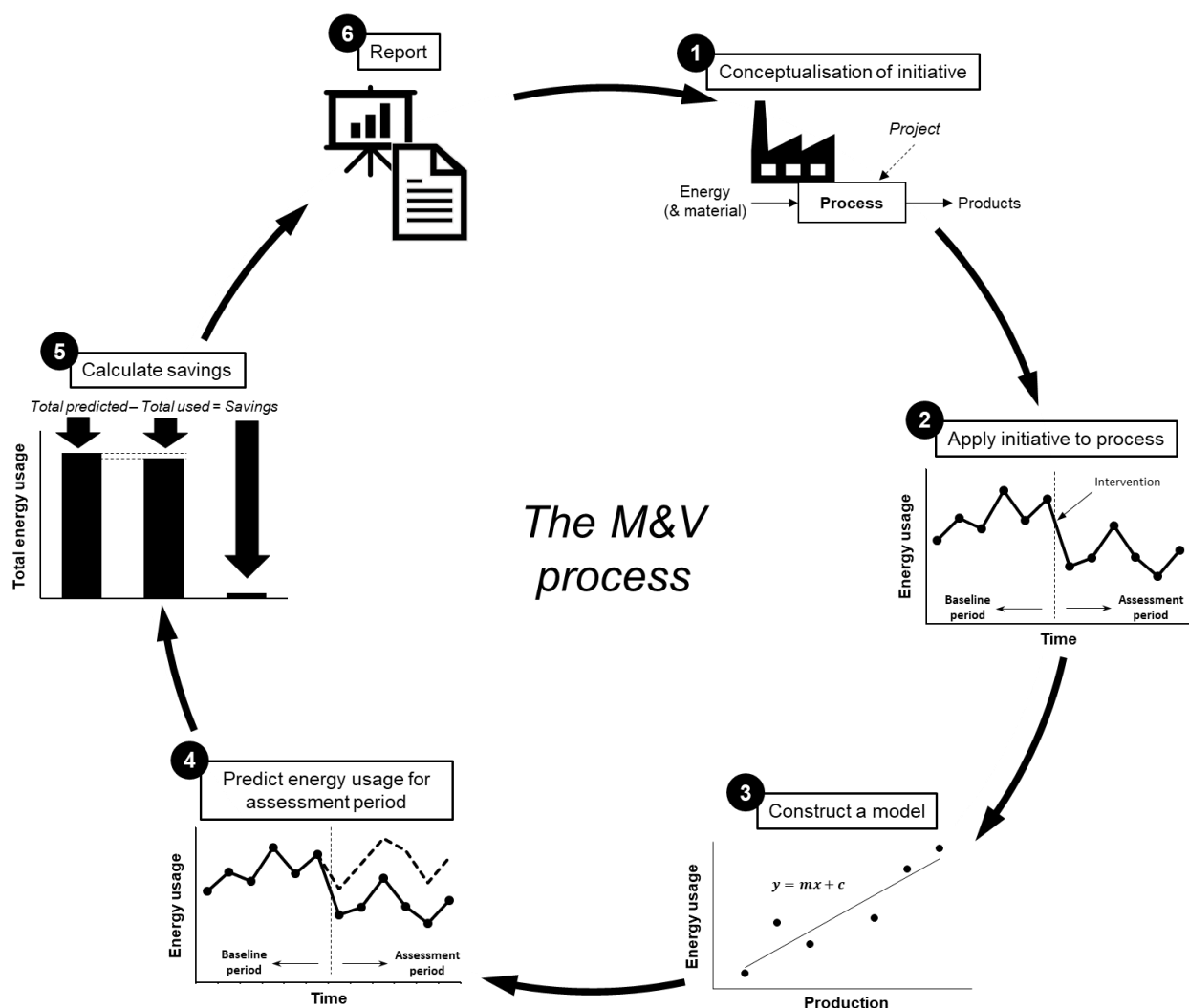


Figure 1.1: The M&V process

Although these standards were set up for different M&V applications, there are common generic steps that form the standard M&V process. Six generic steps were determined from these standards that are presented in Figure 1.1. It involves (1) identification of and ESM that is to be applied, (2) construction of a baseline model, (3) application of the ESM, (4) prediction of the energy usage in the assessment period, (5) calculating the energy savings and (6) reporting.

Step 1 and 2: Identifying and applying an ESM as an initiative

An ESM is identified to be applied to a certain process or facility. Examples of these include lighting changes, improved process control, load management and combining heat and power [12]. Formalising this process and informing different stakeholders of ESM before it is implemented can help to ensure that each step in the M&V process can be sufficiently planned before implementation [27].

Step 3: Constructing a baseline model

Once the ESM has been applied, a model must be constructed through which to predict what the energy usage would have been had the ESM not been implemented [31]. This can be either a physical or a statistical model or a combination of both, which is discussed in more detail in section 1.3. However, before this can be done there are two parameters influencing the process that must be decided on:

- (i) The measurement boundary;
- (ii) The baseline period.

The measurement boundary is the part of the facility that will be included in the evaluation and is an important consideration in the M&V process [12]. There are various options of the measurement boundary can be chosen [12], [27]. An example of different measurement boundary options for a hypothetical facility is given in Figure 1.2. The measurement boundary should be chosen to include the entire effect of the ESM while avoiding interactive effects [32]. This can be done by expanding the measurement boundary to include a larger portion of the facility [27].

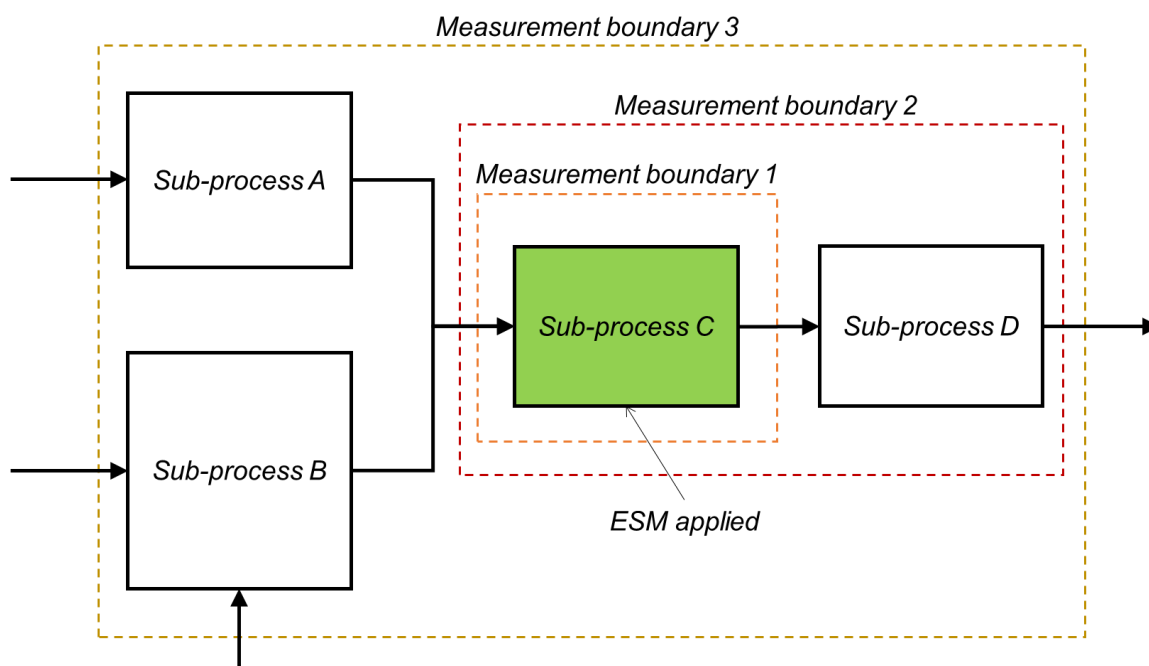


Figure 1.2: Different measurement boundary options

There are three time periods of interest for M&V process. These are (1) the baseline period, (2) the implementation period and (3) the reporting period. These periods are illustrated in Figure 1.3.

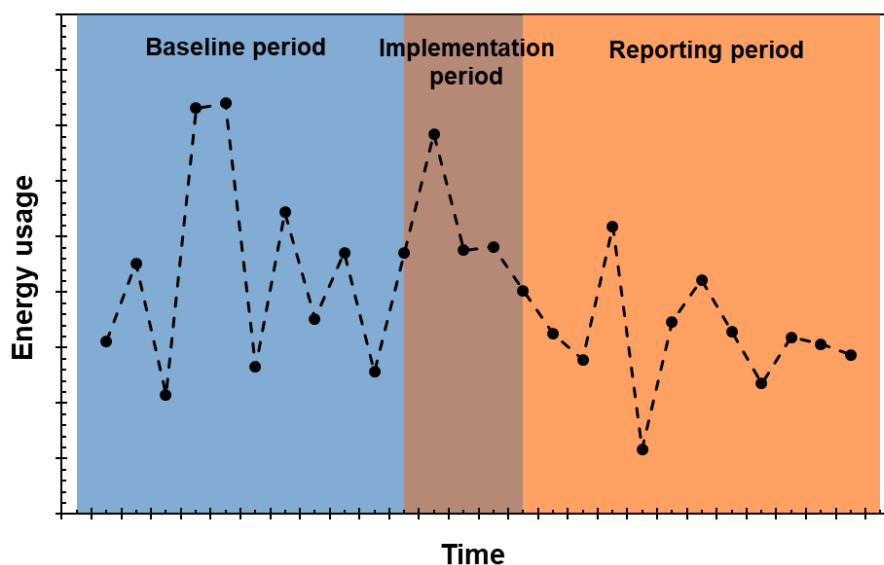


Figure 1.3: The three periods of interest during M&V

The baseline period is the period on which the model construction is based [27], [32]. It should therefore be representative of the conditions before the implementation of the ESM [31]. Provision should be made to insure that enough data is included to represent the different operating modes [20], [32]. Many authors thus suggest a 12-month period to insure that seasonal effects are included which often influences the operation of the system [20], [33], [34].

The implementation period represents the period of change during the implementation of the ESM before stable operation is achieved [20]. This period is normally excluded from both the baseline and the reporting periods. The reporting period is the period after the implementation of the ESM for which savings achieved need to be quantified [31]. The IPMVP recommends that the user of the savings report should determine the length of the reporting period [27]. The reporting period can thus be decided on prior to implementing the ESM or when savings calculation is done.

Once the measurement boundary and the baseline period have been decided on, the model can be constructed. For this, a choice must first be made about the modelling option that will be used. The model must then be calibrated (physical model) or trained (statistical model) so that it represents operation prior to the ESM being implemented [12], [27].

Step 4-6: Apply the ESM, predict energy usage for the assessment period, calculate energy savings and reporting

The model constructed in step 3 is used to predict what the energy usage would have been in the reporting period if the ESM had not been implemented. This is called the baseline [27] and is illustrated in Figure 1.4 along with true energy usage over the period. These two sets of values can be used to calculate the savings according to (Eq. 1) [35], [36]:

$$\sum_j^N E_{save,j} = \sum_j^N (\hat{E}_{pre,j} - E_{true,j}) \quad (\text{Eq. 1})$$

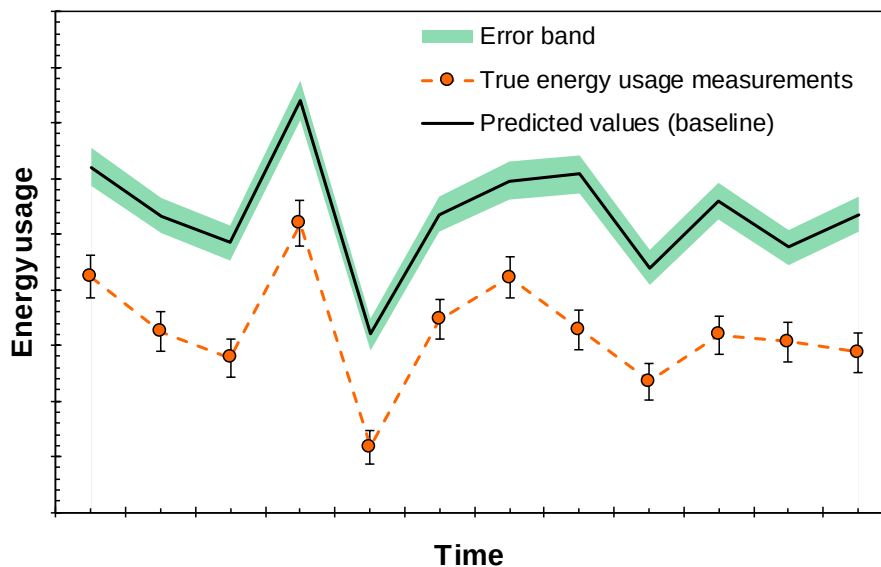


Figure 1.4: Quantifying energy savings

The total true energy usage for the j -th period, $E_{true,j}$, is subtracted from the total predicted energy usage for the period, $\hat{E}_{pre,j}$ and the sum of the savings for each period. Both the predicted values and the measured values have uncertainties associated with it [37], [38]. Therefore, the final savings will also have a measure of uncertainty associated with it.

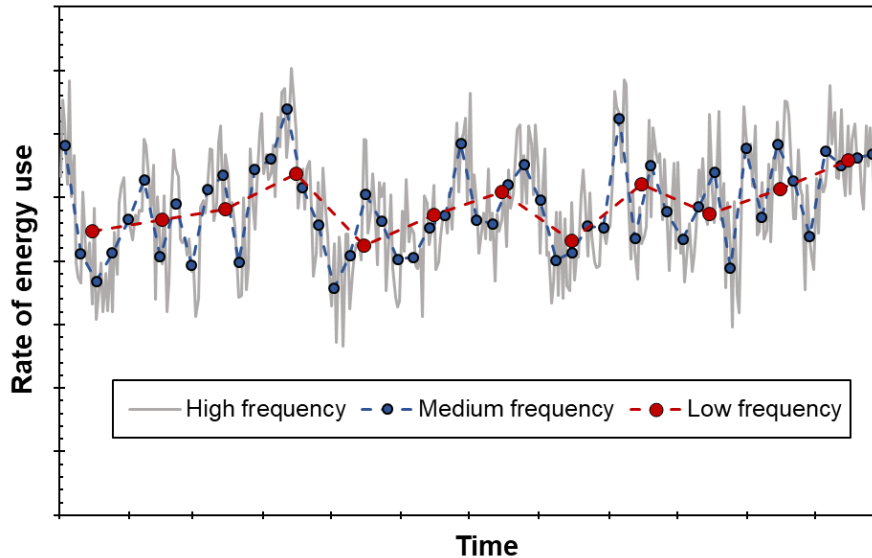


Figure 1.5: Different measurement frequencies for the same data

It is important to note that a particular measurement frequency is associated with the true energy usage (e.g. every two minutes, once a day, once a month, etc.) [27]. An illustration of high, medium and low measurement frequencies for the same data is given in Figure 1.5. The more regular measurements are taken, the higher the measurement frequency is. This is also referred to as the data resolution [25] and it determines the length of each individual period during the savings calculation. A lower measurement frequency will have less noise while a higher measurement frequency will have greater variation in the data.

The predictions made for the baseline must be of the same frequency as the measurement frequency to ensure fair comparison and should be considered during model construction. If specific parameters are measured at a lower frequency (e.g. electricity is measured every two minutes, but production is determined daily), all the parameters measured must be adjusted to that frequency (in this case, daily).

Once the savings achieved by the ESM has been quantified, it should be reported in the required manner [23], [27], [31]. How the reporting should be done is determined by the specific standard or governmental law that should be adhered to. For example, SANS 50010 first only required management of uncertainty [39]. However, later it required uncertainty to be quantified [30].

The standards mentioned and other literature have compiled certain fundamental principles that should be applied during the application of the M&V process. These are presented in Table 1.1.

Table 1.1: Principles of the M&V process

Principle	Reference	Description
<i>Accuracy</i>	[20], [27]	The quantified savings should be within acceptable limits from the actual savings
<i>Completeness</i>	[27]	All the effects of the ESM should be considered
<i>Conservativeness</i>	[20], [27]	Savings should be underestimated where uncertainties arise
<i>Consistency</i>	[27]	Principles and the process itself consistently applied across ESMs
<i>Relevancy</i>	[27]	All the relevant parameters should be measured or known
<i>Transparency / Audibility</i>	[20], [27], [40]	All steps in the process should be traceable and clearly communicated
<i>Independence</i>	[20]	The process should be administrated by an independent party as far as possible

The principles listed help to create a reference framework that M&V practitioners can refer to when decisions need to be made in the process. Consistently applying these principles to multiple M&V evaluations help to ensure a fair comparison between projects. However, this implies that the principles should also be consistently applied during each of the individual steps of the M&V process.

The steps discussed form the basis for the M&V process, but certain additional steps or adjustments can be made [27], [40]. Each standard will provide guidance towards applying the fundamental principles in different ways. Each application of the process will differ based on the type of ESM, the type of facility, the measurement boundary selected, among other factors. [32]. For example, the IPMVP, one of the most reference standards in literature, lists four options [23], [26], [27], [37], [41], [42]:

- Option A: Retrofit isolation – Key parameter measurement
Isolating a part of the facility to reduce the number of variables that must be monitored and eliminate the effect of other ESMs implemented. Only key parameters are measured while others are estimated.
- Option B: Retrofit isolation – All parameter measurement
The same as Option A, except that all influencing parameters are measured.
- Option C: Whole facility
Consideration is given to the whole facility or a major section of it. The effect of the ESM on the entire section is determined and positive and negative effects of non-ESM changes

made are thus also included. The IPMVP suggests this method where savings are 10% of total energy use or larger.

- Option D: Calibrated simulation

Using simulation software to determine either baseline and/or true energy use. The simulation must be calibrated to approximately match the pattern of energy measured for the measurement boundary. This option is useful to separate the effects of different ESMs.

Any of these options can be used to predict energy use, depending on the given scenario. As mentioned, the choice depends on the various factors influencing the M&V process. However, the final option chosen will determine the modelling option used.

1.3 Modelling options

There are qualitative and quantitative methods for predicting energy usage. Qualitative methods are based on the opinions of experts and can therefore be controversial [43]. However, quantitative methods are based on a physical or statistical model of which the uncertainty can be quantified [14], [44]–[46].

A quantitative model is based on a predictable relationship between the energy usage and independent variable(s) [27], [47]. These variables are referred to as the energy driver(s) because it, conceptually speaking, drives the energy usage [27]. These independent variables can potentially be anything that influences the energy usage process or facility. For example, it is common in the building industry to use environmental conditions (such as temperature) and the occupancy of the building [13], [48].

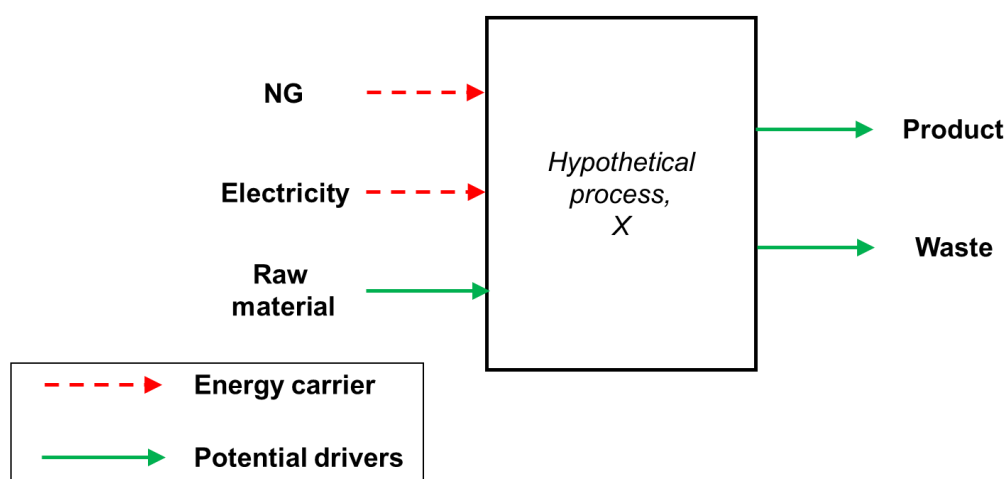


Figure 1.6: A hypothetical process, X, with different energy carriers and drivers

As another example, consider Figure 1.6 where a hypothetical production process is shown. The process consists of certain input and output streams (energy or material). All the streams providing some form of energy to the process forms part of the total energy, in this case, natural gas (NG) and electricity. These are referred to as energy carriers and none of the variables associated with it should be used as drivers [40]. All the remaining variables can be used as potential drivers. Different combinations of the energy carriers and drivers can be used to predict the energy usage of the process.

Over the years different options have been developed and used to predict the energy usage, each having different advantages for different applications. These options are classified in one of three categories [14], [25], [49], [50]:

- (i) White box models (referred to as an engineering method);
- (ii) Black box models (referred to as statistical or artificial intelligence (AI) methods);
- (iii) Grey box models (referred to as combination models or hybrid methods).

White box models are based on simulations using known physical relationships and fundamental principles [19], [50], [51]. These are typically expected under Option D of the IPMVP because of the use of simulation. Black box models are based on statistical learning between the energy drivers and the energy usage from historical data [50]. Here, the actual relationship is unknown, but is approximated by a mathematical function or a group of mathematical functions. Grey box models are obtained when combining physical simulations and regression [52], [53]. This is done when some of the relationships between variables cannot be approximated or are poorly approximated by physical laws [54]. These relationships are then predicted using black box models, while the others are based on fundamental physical principles.

White box models require expert knowledge to set up, can be costly and tend to be less accurate [50], [51], [55]–[57]. However, it has the advantage that other, similar facilities can be used to calibrate the simulation for the baseline, if baseline data is not available [14], especially for greenfield projects. Thus, white box models are the most relevant when facility design specifications are available [49].

The disadvantages associated with the requirement for expert knowledge and cost also apply to some extent to grey box models [50], which can also suffer from computational inefficiencies [52]. However, grey box models have shown promising levels of accuracy in some instances [58].

Black box models can be either regression, absolute or intensity-based [59]. Absolute and intensity-based approaches are not suitable when there is a high degree of variability in operating conditions or significant baseload [60]. Due to its simplicity and speed of calculation [61], regression techniques are the most popular method for establishing baselines in M&V [50], [60]–[62].

1.4 Statistical regression models

Statistical regression can be described as the process of obtaining an approximation of the relationship between two sets of variables, usually a set of dependent variables, y , and set of independent variables, x , by changing some parameters using known data [63], [64]. A statistical regression model (SRM) is thus any mathematical model (or function) that approximates this relationship to within a certain error, ε , through the process of regression. This can be expressed mathematically as,

$$y = f(x; w) + \varepsilon \quad (\text{Eq. 2})$$

where w is the set of parameters or weights that are changed during the process of regression.

The field of M&V has looked into applying SRMs in four different sectors – residential, commercial, industrial and education and research [19], [46], [65], [66]. Much work has been done on energy modelling and M&V for the residential, commercial and education and research sectors, but not so much for the industrial sector. A review done by Wang and Srinivasan showed that 17%, 33% and 42% of energy modelling studies focus on the residential, commercial and education and research sectors respectively [50]. However, the industrial sector can differ substantially from the residential and commercial sectors and in 2012 accounted for 54% of delivered energy use in the world [46], [67].

Predicting energy use does not only allow calculation of savings (M&V) but further allows the evaluation of energy efficiency, detection of faulty systems and control amongst other things [50], [68]. Therefore, to understand the work done on SRMs, the broader field of energy modelling must be considered.

Many different types of SRMs have been used in energy modelling. This includes linear regression (LIN), non-linear regression, the machine learning methods and the hybrid and ensemble methods [25]. These four classes of SRMs are shown in Figure 1.7.

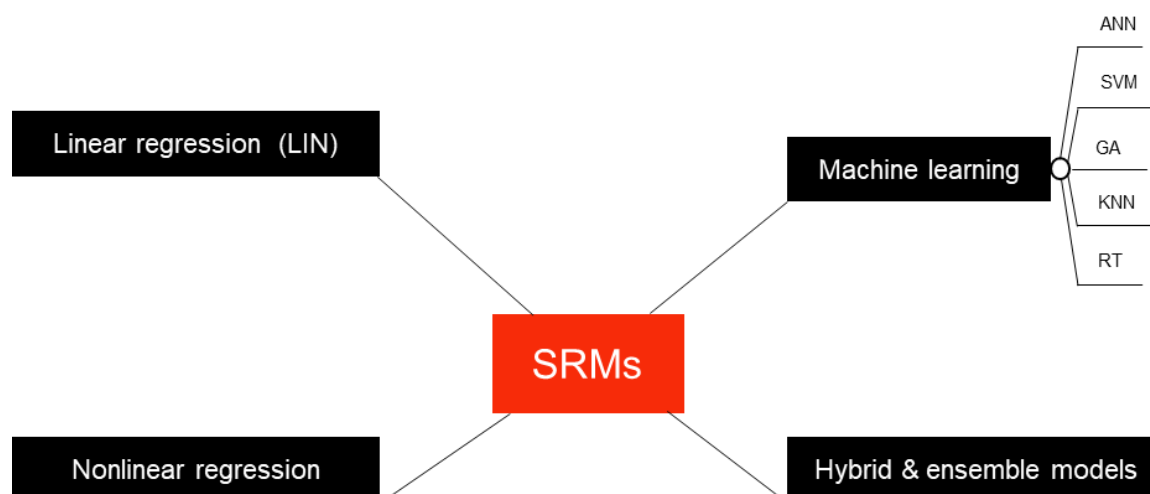


Figure 1.7: Different types of SRMs

Linear regression (LIN) is a common technique used in many different areas of study [69]–[72]. It assumes that there is a linear relationship (or relationships) between the dependent and independent variables [73], [74]. This can be expanded to non-linear regression where the relationship is assumed to be non-linear.

Over the past couple of decades, machine learning models have also come to be frequently used [63]. Models in this category include artificial neural networks (ANNs), support vector machines (SVMs), genetic algorithms, *k*-nearest neighbour (KNNs) and regression trees. Any of the models in the different categories mentioned here can be combined to form hybrid or ensemble models [75].

Machine learning models have become widely used in the residential and commercial sectors [76], [77]. ANNs have become a popular choice because of their ability to account for non-linear dependencies on multiple independent variables. Other models in this category such as SVMs, regression trees and genetic algorithms, are also commonly used as it holds significance for the field of energy modelling [49], [51], [52], [78].

One of the main drivers of advances in energy modelling is the progress made in computer technology [79]. The increased availability of open-source machine learning packages in popular data processing software like Python and R makes it easy for practitioners to set up more complex models [14], [46], [80].

This has given rise to a revolution in M&V referred to as M&V 2.0 [17]. M&V 2.0 refers to advanced, automated M&V that tend to use higher measurement frequencies due to the increased availability of data [19], [25]. It also allows multiple models to be set up using different model types

and different combinations of drivers. Consequently, it has increased research towards the use of machine learning models for M&V purposes. A need arose from these multiple modelling options to evaluate the models effectively to choose the best model for the required purposes.

1.5 Model evaluation

When looking at model evaluation, the fundamental principles of M&V given in Table 1.1 in section 1.2 can be considered. Completeness and relevancy are dependent on data quality and the choice of variables [81]. Factors like transparency and independence are determined by the holistic approach to the M&V process. Therefore, only accuracy, conservativeness and consistency are dependent on the model developed.

Most studies are concerned with model accuracy [12], [25], [82]–[84]. A more accurate model will reduce the uncertainty of the final savings calculation. Three sources of uncertainty exist during the M&V process – (1) sampling, (2) measurement (data capturing) and (3) modelling [27], [46].

Sampling uncertainty is based on the completeness of the data collected and the final data selected to represent the process. Measurement uncertainty involves the accuracy of the measurements taken during data collection and the amount of noise present in the data. The remaining uncertainty is the modelling uncertainty which arises from simplifications and general imperfections of the model [85].

Reducing measurement uncertainty may have a small impact and is potentially unnecessary [86]. Therefore, this study only focuses on uncertainties originating from the model and the data sampling process. Both these uncertainties are assumed to be independent of measurement uncertainty.

Usually, the accuracy of models is evaluated by calculation of an error metric for a test dataset and through use of some visualisation technique [20], [77], [84], [87], [88]. Different studies employ a variety of statistical metrics, which have different advantages and disadvantages. However, in many cases these metrics deliver similar rankings between models.

Models have also been evaluated for sensitivity, which relates to model consistency (i.e. a model that is less sensitive to changes can be said to be more consistent). Different authors have subjected models to different types of sensitivity analyses. Granderson *et al.* performed a sensitivity analysis by applying their selected models to many commercial buildings in different climate zones without varying the inputs used for each model [19]. Gallagher *et al.* investigated

the sensitivity of their selected models to measurement frequency [14]. Other studies have looked at the influence of different seasonal periods or months of the year [83], [89], [90]. Although, these studies also considered model accuracy, each study only looked at a specific aspect of consistency.

Many studies have looked at optimising a single model type for a specific case study based on accuracy [55], [56], [87], [88]. Studies that do compare multiple model types, usually also consider a certain specific case study that are not aimed at M&V [77], [83], [91]. Few studies have evaluated different model types for both accuracy and consistency. There is thus a need for a comprehensive model evaluation technique to evaluate multiple model types in the context of M&V.

The evaluation of the multiple model types for M&V purposes, especially machine learning models, have been limited in the industrial sector. Studies in this area have only looked at isolated case studies. Therefore, there is also a need for more practical examples of applying and evaluating multiple models, particularly the newer machine learning models, for M&V in the industrial sector.

1.6 Model selection

As mentioned, several studies have focused on the accuracy of different model types in the non-industrial sectors. The results of these studies can be used for the initial screening of the models available.

Carstens *et. al.* presents the case for applying Bayesian analysis during the M&V process rather than the traditional frequentist approach (given in most standards like ASHRAE and IPMVP) [37]. They argue that Bayesian analysis offer a better representation of uncertainty while being simple and robust to outliers. They also argue that the inclusion of a prior (an initial assumption of the saving) as required by the Bayesian approach is an advantage, although they acknowledge that it is often seen by others as a draw-back. Bayesian analysis requires more thinking from the M&V practitioner(s) [86].

Heo & Zavala described a Gaussian process modelling framework for M&V [92]. They demonstrated that unlike LIN, Gaussian process models can capture complex non-linear and multivariable interactions. However, Gaussian process models are developed under a Bayesian setting, so priors are still required [93]. Therefore, machine learning models are a promising alternative because they can capture non-linear behaviour without the requirement of a prior.

ANNs generally gain focus in many energy modelling applications [52]. ANNs have been used for many purposes, including the accurate estimation of wind turbine power curves [94]. Its use has yielded high accuracies, achieving errors in the range of 1.3% to 5.1% in the residential sector [51], [55], [56], [61], [95], 0.01% to 21.0% in the commercial sector [13], [57], [68], [81], [87], [96]–[100], 0.5% to 11.2% [48], [77] in the education and research sector and 0.02% to 21% industrial sector [14], [46], [101], [102]. However, ANNs appear to be more sensitive [46]. In a comparison done by Platon *et al.* between ANNs and case-based reasoning, overall, the ANNs were more accurate but were more sensitive to the length of time forecasted [81].

In the study by Yildiz *et al.* in the commercial sector, the simpler linear models performed well in comparison to more advanced machine learning options [77]. However, the best performing model was an ANN, showing the lowest mean absolute percentage error (MAPE) and root-mean squared error (RMSE). In another case, Li & Huang found better accuracy and precision with LIN than with ANN [68]. In a comparison done by Zhang *et al.*, the ANN tended to perform worse than the other models, although all models were comparable [13]. Therefore, when there is a simple relationship between the input variables and the predicted field, ANNs will not necessarily produce better results than LIN [103].

Catalina *et al.* obtained errors between 0.67% and 3.04% for different residential case studies with LIN [104]. For the study done in the industrial sector by Therkelsen *et al.*, single variable LIN models performed better than multivariate linear regression models based on statistical validity tests [59].

In the comparison done between SVM and different types of ANNs by Li *et al.*, SVM performed slightly better than the ANNs [99], [100]. Gallagher *et al.* compared the accuracy and sensitivity of popular machine learning models with that of LIN models for a single industrial case study [46]. The SVM models' performance was the most erratic in terms of average savings across measurement frequencies, while LIN was the most consistent across all measurement frequencies. KNN models showed the lowest error across all measurement frequencies.

Massana *et al.* concluded that linear models are easier to implement and use than machine learning models, but that machine learning models can be advantageous for obtaining a higher accuracy [77]. Thus, there is no decisive best performing algorithm, as noted by Gallagher *et al.* [46].

Many other models exist besides those already mentioned [105]. For example, Yildiz *et al.* have suggested auto-regressive models that appear to work well for electricity load forecasting [77],

[106]. However, it (like many other well-performing models) is a time-series model, meaning that each successive prediction is dependent on the measurement of the predicted field made directly prior to it. This makes it unsuitable for M&V where predictions in the reporting period must be representative of the old system and may not be influenced by the new system.

Of the models mentioned, LIN, ANN and SVM are popular and performed well in previous studies. Therefore, these three models are also chosen for the current study. Although regression trees are also popular, it tends to perform worse than other models [46], [48], [77] and will not be included. Given the promising results obtained by Gallagher *et al.* with KNN, it is also added to the study.

Although machine learning models are more complex, the M&V practitioner does not have to understand the inner workings of the model, but only needs to evaluate the model(s) according to the chosen criteria [14]. However, a brief overview of these models is presented to allow the reader to understand the concepts mentioned in this work better.

Linear regression (LIN)

As mentioned, LIN is one of the most well-known and used techniques in the academic world. It tries to fit some linear relationship between the dependent variable and each of the independent variables [107]. This is illustrated in Figure 1.8.

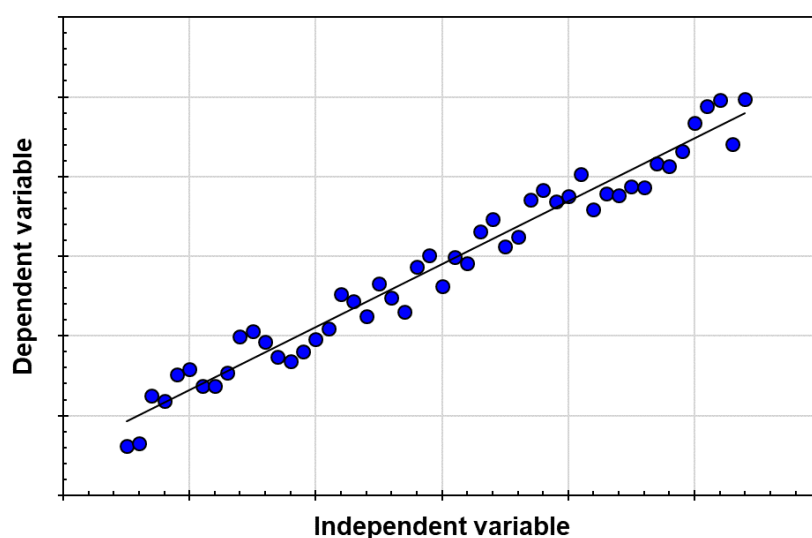


Figure 1.8: Illustration of fitting a linear model

LIN is often divided into simple linear regression and multiple linear regression to distinguish between a single and multiple independent variables [74], [108]. While different algorithms exist, these algorithms apply to both simple and multiple linear regression [107], [109]. However, each

of these algorithms tries to distribute the residuals evenly between positive and negative values [107]. The residuals refer to the difference between the true and predicted values for each of the data points provided in the training set. This can be calculated using (Eq. 3):

$$\text{Residual} = y_i - \hat{y}_i \quad (\text{Eq. 3})$$

Artificial neural networks (ANNs)

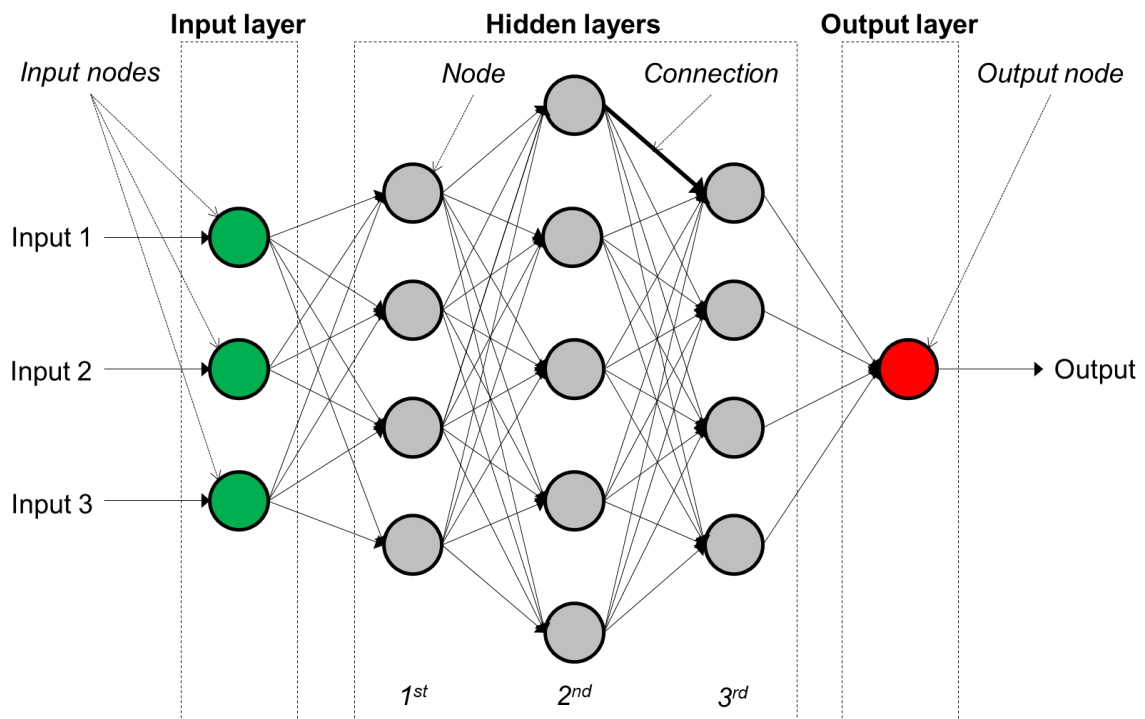


Figure 1.9: The structure or architecture of a simple ANN

Although an ANN represents a complex mathematical function, its structure can be represented in a relatively simple manner, as with the example given in Figure 1.9. An ANN consists of nodes that are organised in layers [13], [50], [51], [94]. Each node in a layer is connected to every other node in the layer that follows. Three groups of layers form the layout or architecture of the network:

- (i) An input layer;
- (ii) Hidden layer(s) (single or multiple);
- (iii) An output layer.

Each node represents a mathematical function (known as an activation function) and each connection has a weight associated with it [50]. The number of inputs is equal to the number of independent variables. The exact calculation procedure is given in Appendix A1.

Support vector machines (SVMs)

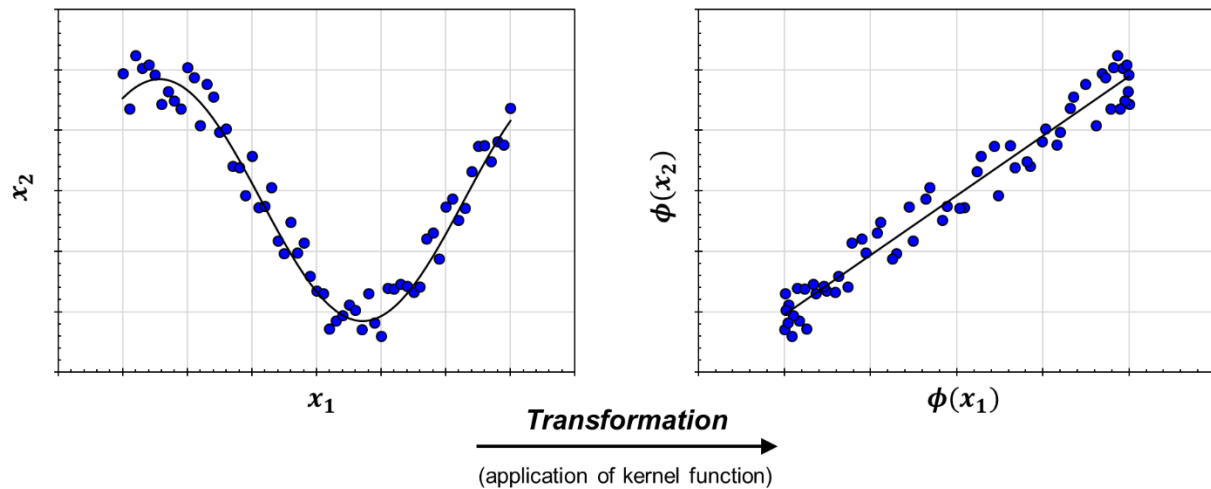


Figure 1.10: Non-linear SVM

SVMs entail transforming a potential non-linear regression problem to a linear one by applying a mathematical function known as a kernel function [110]. This type of “mapping” is represented in Figure 1.10 where variables x_1 and x_2 are transformed to new variables $\phi(x_1)$ and $\phi(x_2)$. After this is done, the problem becomes a LIN problem. Examples of the different types of kernel functions and a simple discussion on the mathematics used are provided in Appendix A1

K-nearest neighbours (KNNs)

The application of the KNN algorithm is conceptually simple. An example is given in Figure 1.11.

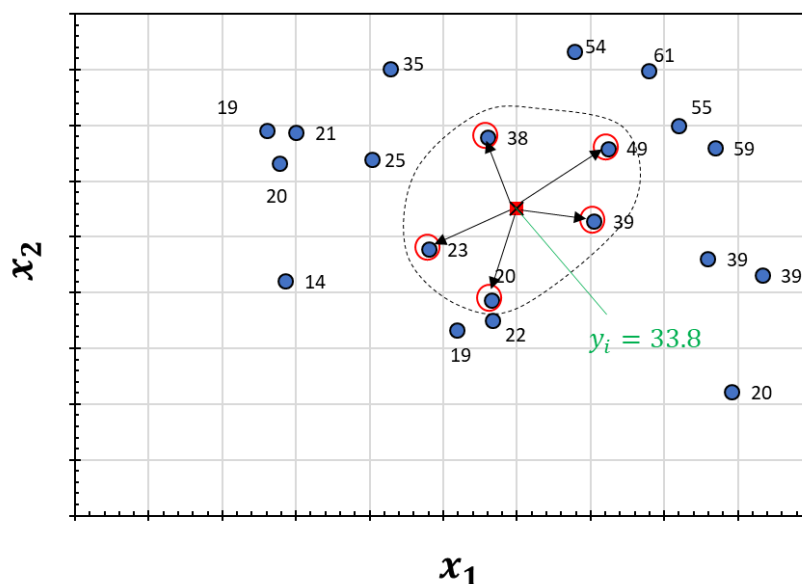


Figure 1.11: Illustration of KNN for $k = 5$

The algorithm calculates all outputs directly from the training dataset provided [111]. To predict the output for any new data point, the average is taken of the k data points nearest to the desired point [63], [112]–[114]. In the example provided, the five nearest data points were used (i.e. $k = 5$) and the result is 33.8. The algorithm becomes more complex when the averaging is done on a weighted basis by using a function also known as a kernel function [114]. The details for this process are presented in Appendix A1.

For each of these models, certain parameters must be chosen that are not calculated by the regression process. These parameters are known as hyperparameters, which are, for the LIN model, the use (or absence) of a constant or intercept term. Other examples of these include the number of hidden layers for the ANN and the kernel functions for both the SVM and KNN models. A complete list of hyperparameters for the chosen models is given in Appendix A2.

1.7 Hybrid and ensemble models

There also exist hybrid and ensemble models, which uses single models in combination, as illustrated in Figure 1.12 [50]. The term hybrid model is used when models are combined in series (Figure 1.12(a)) while an ensemble model combines the results of different models used in parallel through some integration technique (Figure 1.12(b)). Many examples exist of model type combinations and more complex models that are created through it to yield more accurate results [52].

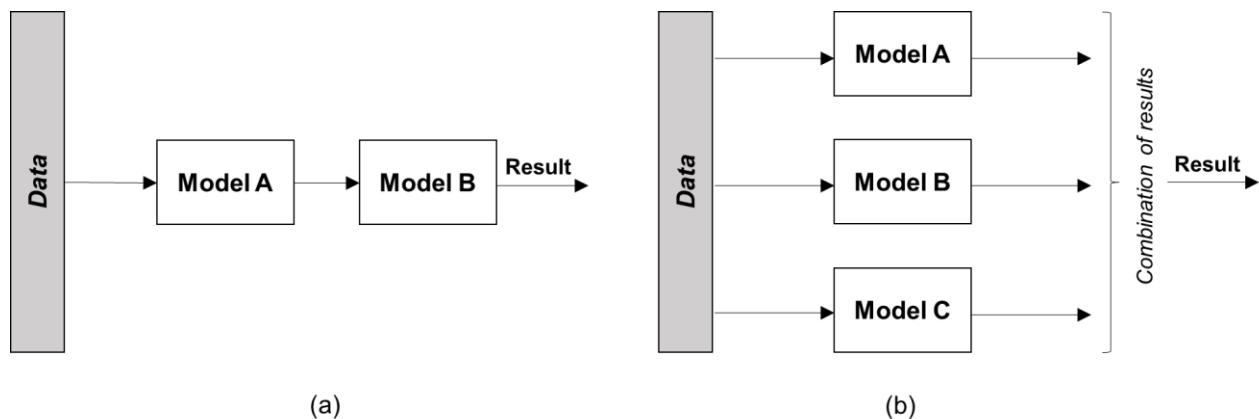


Figure 1.12: Illustration of the working principles of (a) hybrid; and (b) ensemble models

Many of the models investigated by Granderson *et al.* were combination-type models [19]. Da Costa e Cruz proposed a method for combining several models using a regression tree approach where certain branches are activated (turned on or off) based on certain conditions [115]. A well-studied hybrid model is change-point models [77], which is a simple combination or a regression

tree and LIN and relates to the approach followed by Da Costa e Cruz. Mittal *et al.* found that their proposed hybrid model reduced error variance in comparison to independent models [88]. A review on some hybrid models is given by Ahmad *et al.* [78].

For ensemble models, the models can either be the same type of models trained with different data or different types of models [50]. The most common integration technique is averaging of the outputs of the different models. Fan *et al.* combined the outputs of eight popular regression algorithms using a genetic algorithm for a commercial building in Hong Kong [116]. They obtained good results with a MAPE between 2.32% and 2.85%. Jovanović *et al.* analysed three different integration methods using a university campus as a case study – simple (model) averaging (SMA), weighted (model) averaging (WMA) and taking the median [117]. Their results did show notable differences between the integration approaches. However, in both studies the use of an ensemble model yielded better results than the single models tested by them. A review of ensemble models is given by Wang and Srinivasan [78].

1.8 Objectives

Based on the literature reviewed, there is a need for a more comprehensive model evaluation technique for comparing model types for the purpose of M&V. Also, research into the application of different models for the M&V purposes has been very limited in the industrial sector. There is also a need in particular to consider model-averaging for use in the industrial sector, given the potential improvement in accuracy that can be obtained from it.

The aim of this study therefore is to comprehensively evaluate the use of different models regarding accuracy and consistency for various industrial M&V case studies including model-averaging. This will provide a more holistic view of the suitability of the use of different models, specifically the newer machine learning models, in the industrial sector.

The objectives of this study are:

- (i) To develop a more comprehensive model evaluation procedure, focusing on different model types for application within the M&V context by giving consideration to the applicable fundamental principles of M&V;
- (ii) To apply the developed evaluation procedure to multiple industrial case studies;
- (iii) To determine whether the different types of SRMs perform consistently across multiple case studies within the industrial sector;
- (iv) To test the suitability of the use of model-averaging in the industrial sector;

1.9 Scope of investigation

In order to convince stakeholders of the benefit of ESMs, savings need to be quantified in a reliable manner. This can be done through diligent application of the fundamental principles of M&V within the M&V process. A critical part of the M&V process is the construction of a model to predict the energy usage within the reporting period.

Although white box and grey box models exist, the black box approach of using SRMs provide high accuracy and user-ease. Various types of SRMs exist. These model types must be evaluated for accuracy and consistency and various studies have done so within various sectors in M&V. However, limited evaluations of different types of SRMs within the industrial sector has been done for M&V, specifically the use of ensemble models.

Four model types were chosen based on the results of previous studies. These are the LIN, ANN, SVM and KNN model types. Two model-averaging techniques, namely SMA and WMA, are also included so that six model types are investigated in total. In order to evaluate these model types for accuracy and consistency, a comprehensive evaluation procedure will be developed. This will be used during the application of the different models to five selected M&V case studies within the industrial sector.

Chapter 1 explained the background for the use of SRMs in the field of M&V and the relevant research done on it in the broader field of energy modelling. In Chapter 2, the development of the model evaluation procedure is given together with details about the five chosen case studies. Chapter 3 presents the results of the evaluation of the models according to each case study and a discussion of the general observations drawn from it. Conclusions from the study and recommendations for future studies are given in Chapter 4.

CHAPTER 2 METHODOLOGY FOR MODEL EVALUATION

2.1 Introduction to general evaluation procedure

In Chapter 1, six models were chosen for evaluation within the industrial M&V sector. Four of these models were single models (LIN, ANN, SVM and KNN) and two were ensemble models (SMA and WMA). The chapter also motivated why these different models should be evaluated within an M&V process, specifically for accuracy and consistency. To do this, a general evaluation procedure must be developed which will be described in this chapter.

Different evaluation procedures have been developed and reported in literature. As mentioned, most studies focus on model accuracy while some have taken consistency into account. Although these procedures differ in their approach and purpose, there are certain steps that are common to most procedures. Based on observations from literature, a seven-step evaluation procedure was developed:

Step 1: Data gathering and selection of measurement boundary and data periods [14], [19], [36], [68], [82], [118]–[120]

Step 2: Data treatment [14], [19], [68]

Step 3: Selection of driver sets [14], [82]

Step 4: Selection of hyperparameters for each model type [14], [46], [102], [118]

Step 5: Model training [14], [19], [46], [68], [82], [102], [118]–[120]

Step 6: Model evaluation [14], [19], [46], [68], [82], [102], [118]–[120]

Step 7: Model-scoring and interpretation of results [14], [19], [82]

The purpose of this study in part is to investigate the influence of different measurement frequencies on the results. Daily, weekly and monthly frequencies were chosen as used by Gallagher *et al.* [46]. These frequencies are created in step 2 of the procedure. Steps 3 to 6 are then iteratively applied to each measurement frequency, as illustrated in Figure 2.1. Step 7 is then applied by considering the results of the different measurement frequencies collectively.

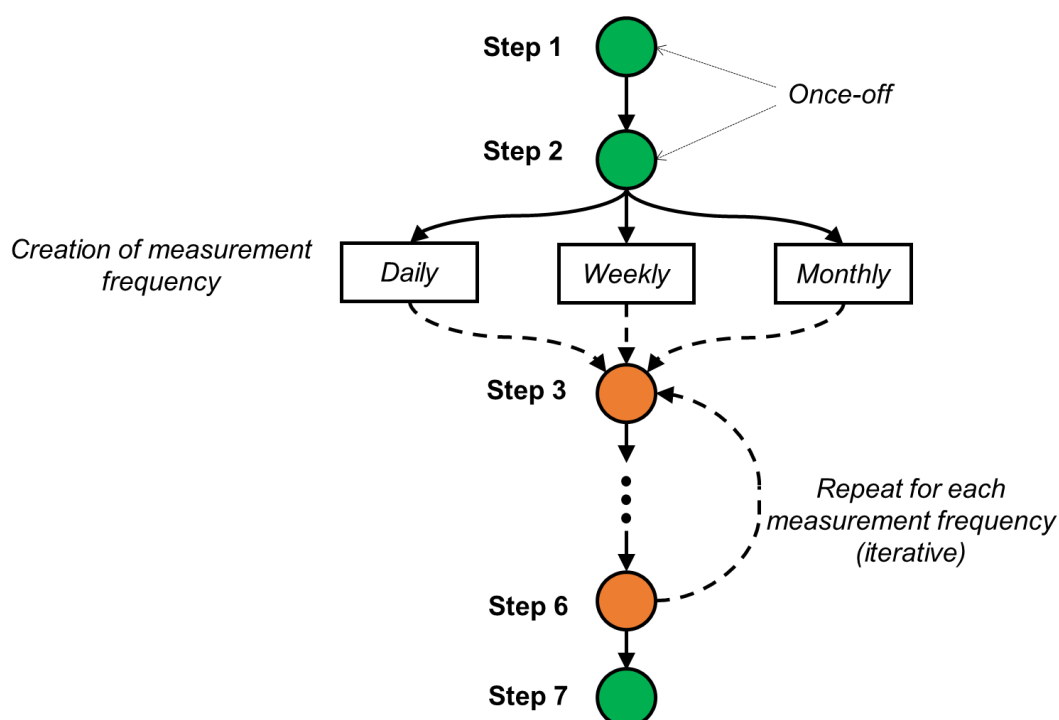


Figure 2.1: Illustration of the once-off and iterative parts of the process

2.2 Step 1: Data gathering and M&V process parameter selection

The purpose of the first step is to do the initial preparation of the datasets needed for the rest of the procedure. It involves (1) gathering information about the ESM, (2) selecting the general parameters needed for the application of the M&V process, (3) gathering data and creating plots, (4) preparing variables and (5) splitting the data into separate sets. The flow diagram showing these sub-steps is given in Figure 2.2.

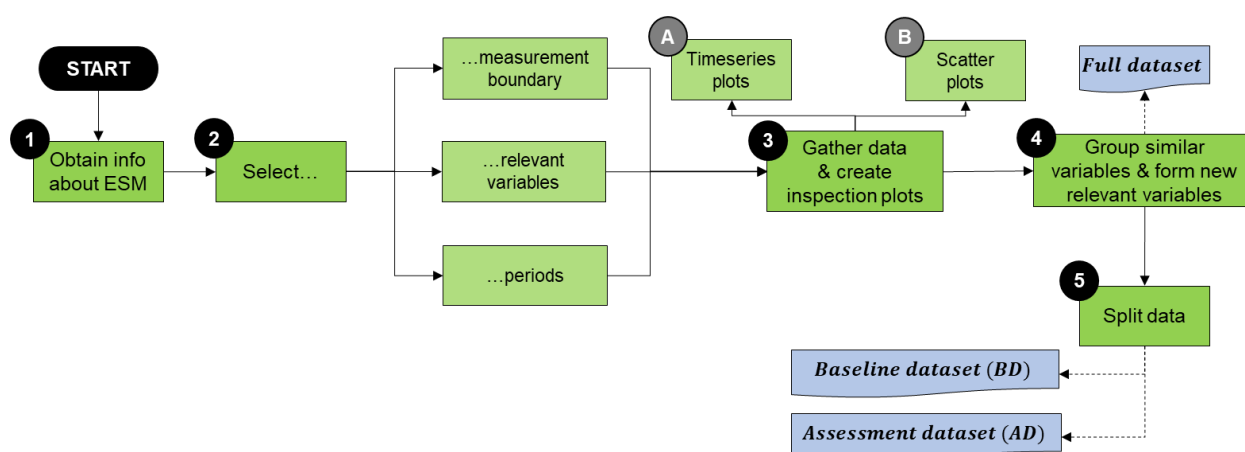


Figure 2.2: Flow diagram for step 1

The first sub-step is to obtain as much information about the ESM as possible. This will include details like the part of the process to which the ESM was applied, what ESM was and when the ESM was applied. This will help to ensure that the measurement boundary and the data periods (baseline and reporting) are correctly chosen and all applicable data sources are considered.

Considerations for choosing the measurement boundary, the baseline and reporting periods, were discussed in section 1.2 as part of the general M&V process. These are summarised in Table 2.1 along with other, additional considerations. Once each of these parameters has been selected, the data for the variables that fall within the measurement boundary can be obtained for the applicable periods.

Table 2.1: Considerations for choosing the baseline and the reporting period

Period	Reference	Consideration
Measurement boundary		Encapsulate the entire ESM while avoiding interactive effects.
Baseline period	[20], [27]	The baseline period must represent all the operating modes of the facility.
		Only periods of known energy use should be included if possible. If not possible, the energy use can be estimated through simulation.
	[20], [27]	A 12-month baseline period is suggested to include the entire range of production and weather.
Reporting period	[27]	Where interactive effects of other ESMs must be avoided, shorter baseline periods can be chosen.
	[27]	The user of the savings report should determine the length of the reporting period.
	[20]	A popular time period is 12 months so that it is comparable to the baseline period.
Additional considerations	[27]	Only choose adjacent periods for the baseline and reporting when the ESM can easily be turned “on and off”. Otherwise, allowance must be made to remove the transition period (where possible) in order to avoid inclusion of unstable operation that affects model quality negatively.

In industrial applications, there are various data sources available [60], [121]. Examples of these include log sheets and different forms of databases [14], [20]. These values can come from periodic analyses taken by practitioners or online measurements from permanent sensors installed at the facility.

The M&V practitioner can use a variety of sources when compiling the final dataset. However, all these sources must be adjusted to the same measurement frequency, as explained in Chapter 1. This can be done by averaging or summing over the shorter data periods in the higher frequency that fall within the longer data periods of the lower frequency. An example of this is given in Figure 2.3.

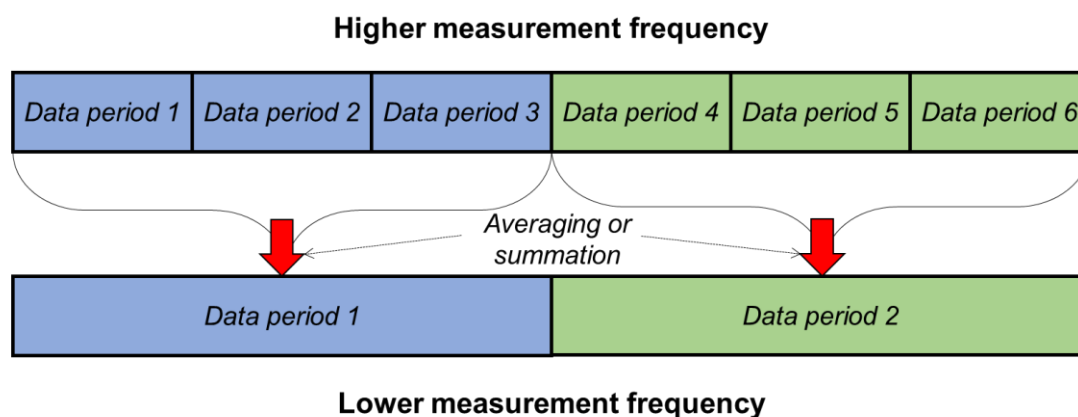


Figure 2.3: Illustration of taking a single variable from a higher measurement frequency to a lower measurement frequency

It is suggested that averaging be used for potential drivers, while summation should be used for variables that form part of the total energy use so that savings calculations can be explicit. Once this is done and the final dataset (*FD*) is constructed it can be split into the baseline (*BD*) and assessment (reporting) (*AD*) datasets respectively according to the time periods chosen.

The M&V practitioner must include all the potentially relevant variables in the final dataset used. However, if any of the variables used were influenced purposefully by the ESM, that variable must not be included in the dataset, e.g. if the pressure setpoint was reduced in a compressed air system to reduce energy usage, pressure cannot be included. Variables must be identified either as (1) energy drivers (input variables) or (2) energy carriers and grouped accordingly. Also, any non-measured variables that might be relevant must be calculated. This might be cooling duties, production totals or the average operating temperature for similar process equipment included in the measurement boundary.

After all the relevant variables have been grouped, different inspection plots can be created. These include (1) time-series plots of each of the variables and (2) scatter plots of each of the energy usage against each of the potential drivers. For the scatter plots, a simple linear model is also fitted and the resulting R^2 is calculated. An example of each of the plots is shown in Figure 2.4.

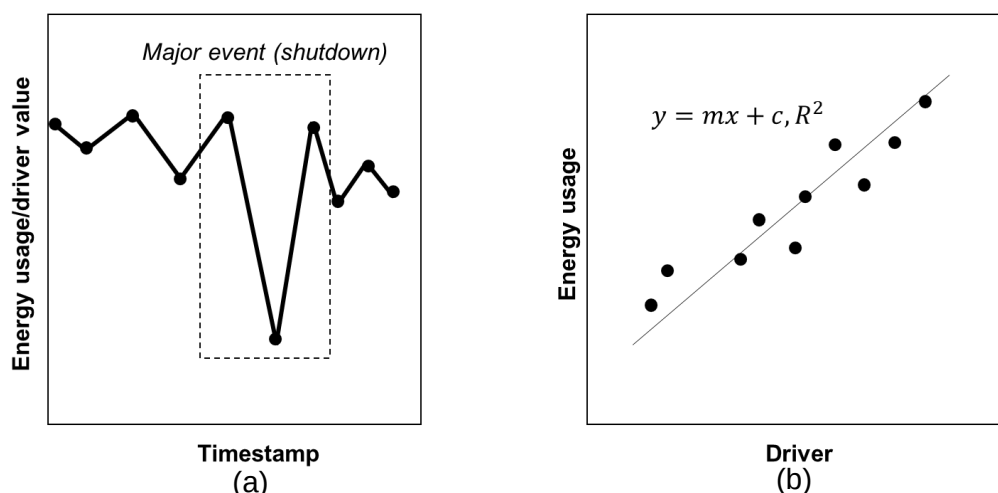


Figure 2.4: Example of (a) time-series plot and (b) scatter plot of energy usage vs a potential driver

The plots are constructed based on the full dataset. From the time-series plot, major events (such as shutdowns or upsets) can be identified. Any general patterns in the operation of the process can also become visible. Inspection of the scatter plot gives a general indication of the linearity of the process with the help of the R^2 -value. More detail regarding the significance of the R^2 -value is given in section 2.7.

Each of the sub-steps mentioned here assists in creating a final dataset with relevant variables for the representative period. However, the quality of the raw data gathered may be poor, which will be transferred to the variables included in the final dataset. In this event, the accuracy of models constructed from the dataset can be influenced negatively [122]. To prevent this, a data treatment step should be applied.

2.3 Step 2: Data treatment

Data treatment involves, in most procedures, detection of outliers and the scaling of data [14], [19], [68]. This improves the quality of data and enhances model training in later steps. The flow diagram for this step is given in Figure 2.5.

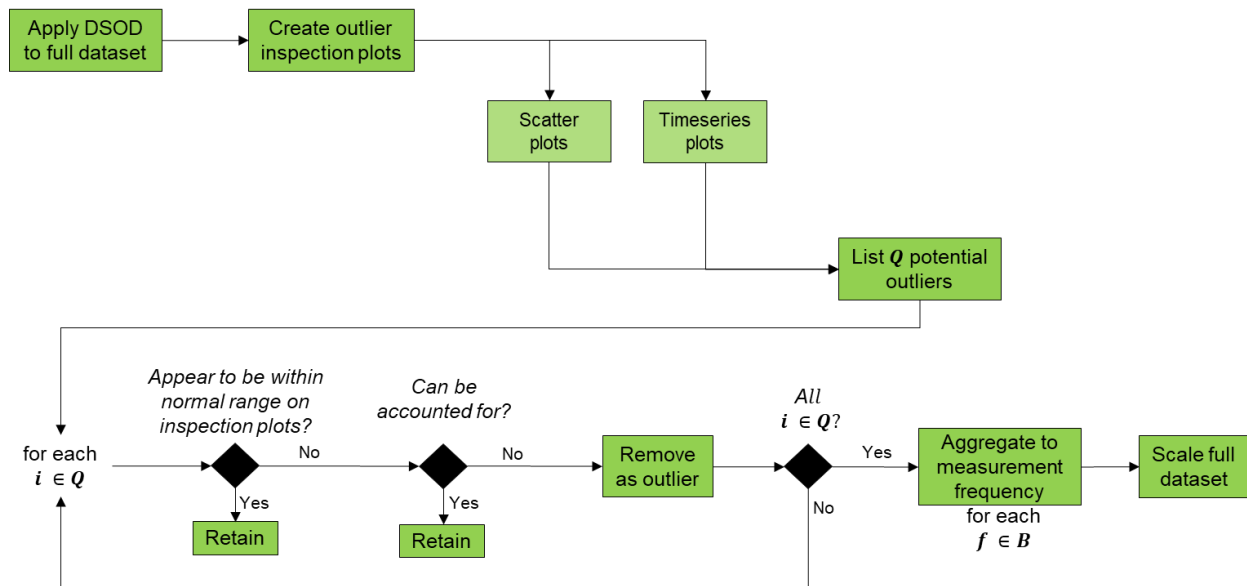


Figure 2.5: Illustration of the data treatment step

Much has been done to detect outliers by considering each variable individually (univariate approach), but also by considering interactions between variables (multivariate approach) [123]–[125]. Although univariate approaches can be useful, improved results can be obtained when taking multivariate interactions into account [125]. Yet critique has been expressed against the reliability of traditional multivariate detection methods, such as the Mahalanobis distance and hat matrix leverages [123], [124].

Zhang *et al.* developed a new outlier detection method for sequential multivariate data referred to as dynamic sequential outlier detection (DSOD) [126]. They demonstrated improved performance of the method over the optimal subset scheme and the principal components scheme. Data for M&V regression is always multivariate (there is always at least one energy driver and the energy usage). Further data is recorded for industrial facilities sequentially. Having considered these reasons, their approach was adopted in this study to flag potential outliers.

Booyesen *et al.* pointed out that a simple, structured, intuitive approach can deliver similar results to complex statistical procedures [127]. Also, not all inconsistent data points should be treated as outliers as it is not necessarily incorrect [112]. Rather, inclusion of certain inconsistent data points can add variation to the data to improve model applicability.

Therefore, the DSOD method is only applied to flag any potential outliers. These points are then highlighted on a time-series plots of all applicable variables along with a series scatter plots of energy usage against each potential driver. Examples of these are given in Figure 2.6.

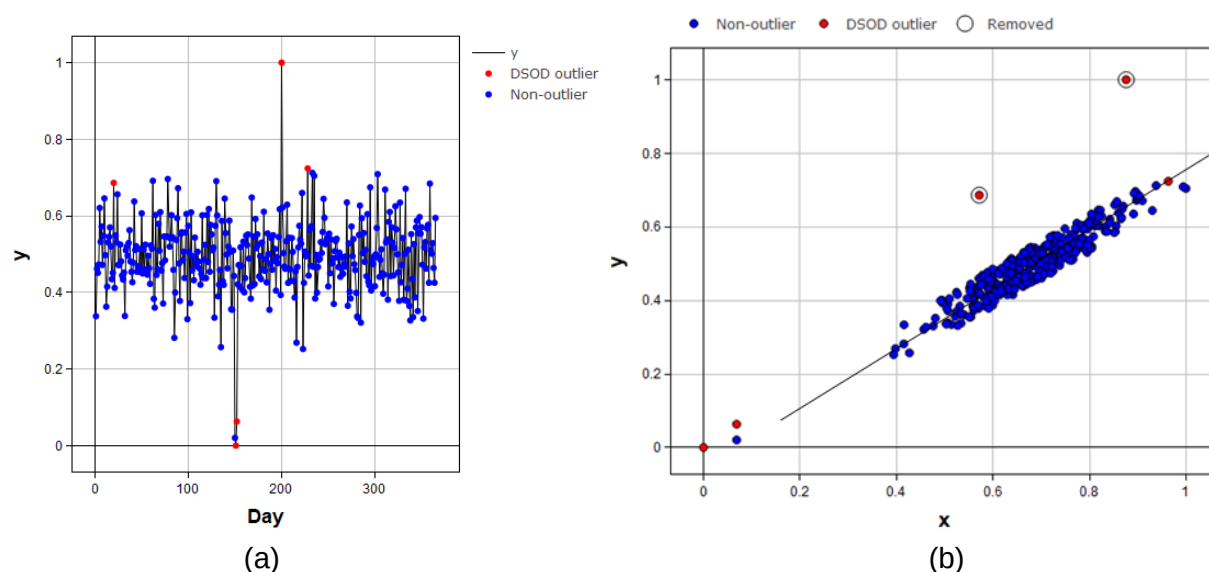


Figure 2.6: Inspection plots for outlier removal; (a) Time-series plot with flagged points and (b) scatter plot with flagged points and points selected for removal

Each flagged data point (the red dots) is then inspected intuitively. If the flagged point appears to lie within the normal operating ranges when collectively considering the plots, the point is retained. In the example in Figure 2.5(b), three of the five flagged points lie close to what appears to be the standard operating line in the scatter plot, and these points will be retained. If the point seems to fall outside the standard operating ranges, an attempt will be made to account for it through information obtained about the ESM like major events at the facility, etc. If the point can be accounted for, it is retained, otherwise it is identified as an outlier.

Outliers are removed from the raw dataset, which in turn is set up at the highest possible measurement frequency. Furthermore, all observations where data is missing for one of the variables should also be removed as it cannot be used during modelling. After that the dataset can be aggregated to the lower measurement frequencies for investigation by summing the data points in each period.

It is common to scale or normalise the data before modelling [84]. When scaling the data, all variables are adjusted to have the same minimum and maximum values. This is done to prevent variables with greater numerical values from dwarfing the effect of variables with smaller numerical values [128].

It thus allows suitable comparison between variables and further helps to improve calculation efficiency of the machine learning models that uses kernel- or activation functions [48], [100],

[103], [128]. A popular method is to scale values between zero and one using (Eq. 4) [100], [128], [129], which will also be employed in this study:

$$z_{i,scaled} = \frac{z_i - z_{min}}{z_{max} - z_{min}} \quad (\text{Eq. 4})$$

The use of this method creates unitless variables. Scaling is done after the removal of outliers so that the range of retained data points are not limited by outliers with very high or low values. It is also important to apply the process to the complete dataset in order to ensure that the relationship between values in the baseline period and the reporting period stays the same. However, the normalisation must be done separately for each measurement frequency to be investigated as the values in lower measurement frequencies will be of a higher order.

The sub-steps described here prepare the dataset for model training by improving data quality and normalising the variables. However, before models can be trained, other selections must be made to determine the functional form of the models. The first of these are the drivers to be used for the energy-use prediction.

2.4 Step 3: Selection of driver sets

Unlike in the residential, commercial and research and education sectors, the type of energy drivers in the industrial sector vary much from case study to case study. There is a need to differentiate between variables that are potential energy drivers and those used within the regression process. The term inputs will be used for the former and the term driver sets for the latter. An example of this distinction is given in Figure 2.7.

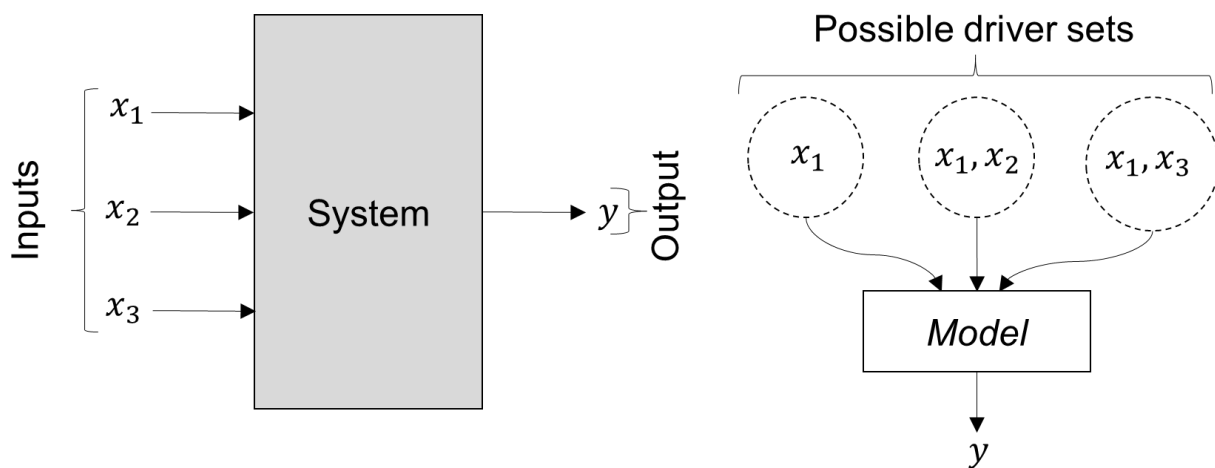


Figure 2.7: Illustration of the distinction between inputs and driver sets

In the example, the system has three inputs referred to as x_1 to x_3 . Any combination of these can be used as a set of inputs to the SRMs and forms the possible driver sets, including the use of any of the inputs on its own. However, one problem that must be avoided when selecting inputs is multicollinearity. Multicollinearity refers to linear relationships that may exist between various independent variables that lead to various potential complications [130].

One of the most popular techniques to counteract this effect within the field of energy modelling is principal component analysis (PCA), which has yielded favourable results [77], [81]. It entails the transformation of a set of variables to a new set that accounts for most of the variability by including effects from each respective variable [131]. The new variables created are referred to as principal components. To illustrate this concept, an example is given in Figure 2.8.

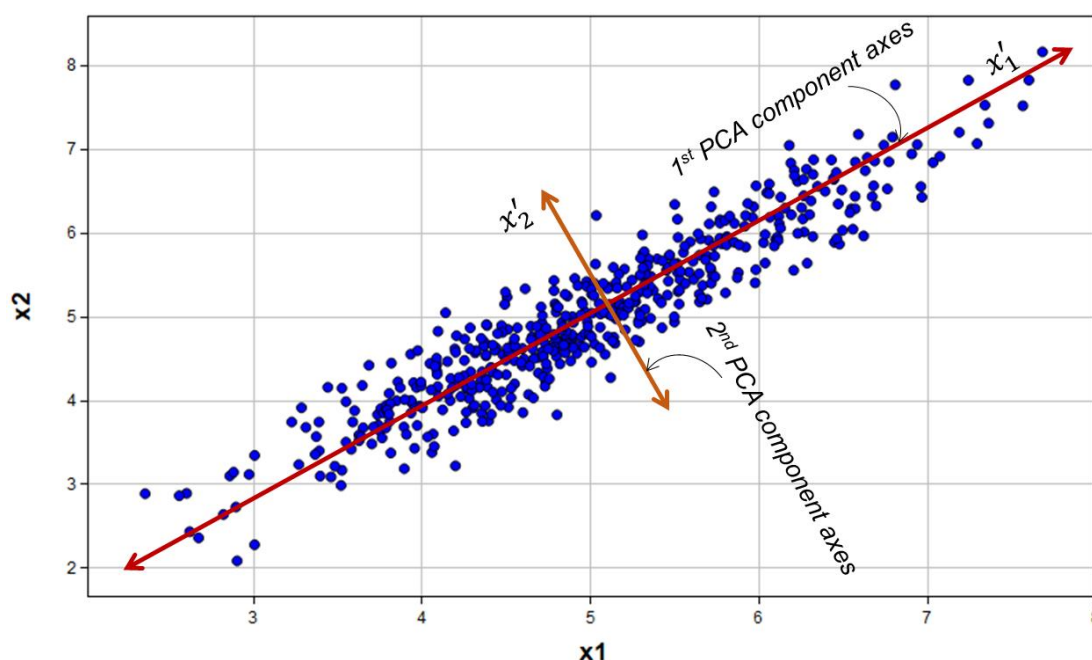


Figure 2.8: Conceptual illustration of PCA

It shows multivariate dataset with two variables, with a strong linear relationship between them. Each data point has a coordinate on the axes of x_1 and x_2 . A new dataset can be created by “rotating” the axes so that the two new variables, x'_1 and x'_2 are entirely independent of one another. Each data point now has a new set of coordinates along these new axes. It is noted that the explanation is only a conceptual illustration of how PCA transforms a dataset but does not represent the actual process of the algorithm.

This example is for a set of two variables, but the PCA algorithm works the same for any multivariate dataset. The algorithm produces the same number of variables as contained by the

original dataset. For example, if the dataset has five variables, then the PCA dataset will have five variables. Another algorithm is then used to select the number of PCA components to be used, starting with the one that captures the most variance. In the example given here, it is the first PCA component, x'_1 . For this study the element-wise k -fold (EKF) algorithm will be used to select the number of PCA components [132].

When applying the PCA algorithm to the dataset, only the measured inputs should be included. It is unnecessary to include calculated variables as they do not add additional variance to the model. The energy usage should also not be included, as this will create an inherent relationship between PCA variables and the total energy usage, even when the total energy use is entirely independent of the inputs. The PCA algorithm should also be applied to the entire dataset so that, as mentioned before, the relationship between the baseline period and the assessment period remains the same.

Although PCA has yielded good results in the energy modelling sector, its use within the M&V industry is uncommon [81]. Further, each M&V practitioner may decide on different drivers depending on the case study. Selecting multiple driver sets will increase the number of modelling options available, which can all be included in the final evaluation to determine the sensitivity of different model types to the inputs selected. For this procedure, four techniques were chosen, including PCA, to choose four driver sets. These are:

- (i) **Technique for driver set A:** An intuitive selection of the single most relevant variable or combining several variables into one (for example by summation)
- (ii) **Technique for driver set B:** Using all the measured potential drivers (not calculated)
- (iii) **Technique for driver set C:** The method given by Gallagher *et al.* [14] to select the most relevant inputs from the measured variables
- (iv) **Technique for driver set D:** Using PCA with the EKF method to select the number of PCA components [133]

Each of the four selection options is illustrated in Figure 2.9.

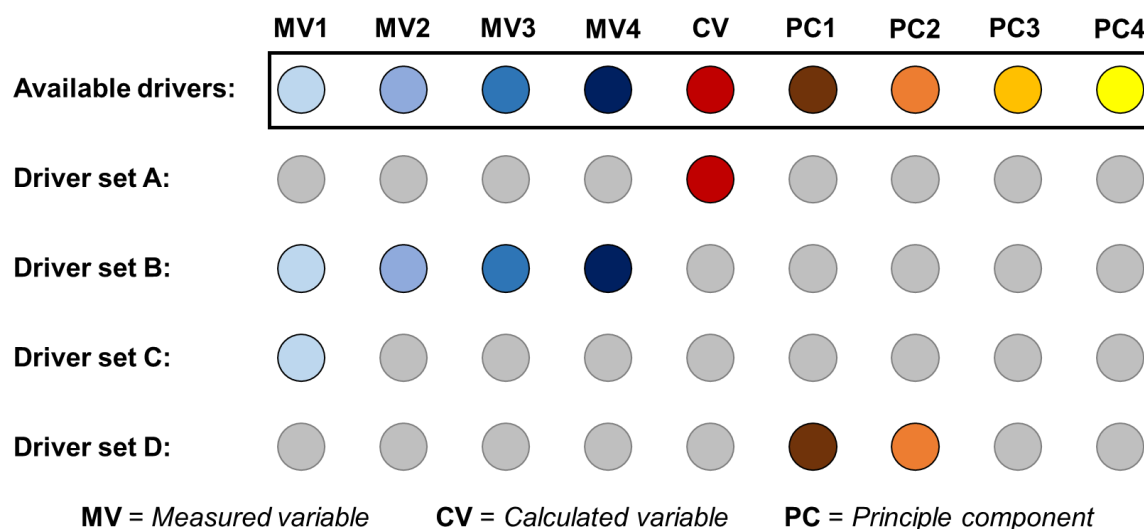


Figure 2.9: Illustration of the four driver set selection options

The diagram represents each of the different drivers available for a hypothetical process. It includes four measured variables (MVs), a calculated variable and the four principal components created through the application of the PCA process. Each of the selection options select different drivers from the available list.

For the intuitive selection of a single variable, the calculated variable is chosen (driver set A). This might be the total production if more than one of the MVs are production variables. For the second driver set, all the MVs are included (driver set B). The procedure given by Gallagher *et al.* [14] might select, for example, the first MV (driver set C). Finally, application of the EKF algorithm might select the first two principal components for the PCA driver set (driver set D).

These methods are only applicable where there is more than one potential energy driver. As with PCA, only the measured variables are used for driver sets B and C because the calculated variables do not add additional variance. The inclusion of driver set B ignores the risk of multicollinearity. This can be done because the goal of its inclusion is to determine the sensitivity of model types towards the inputs provided during model evaluation rather than to construct an optimal model. For this option, it is important not to include any derived variables determined by summation or averaging in step 1.

Once PCA has been applied and the driver sets selected, the dataset finally includes all the variables necessary for model training. However, before the models can be trained for the predictive purposes, the hyperparameters of the different models must be selected.

2.5 Step 4: Selection of hyperparameters

As explained in Chapter 1, hyperparameters (details in Appendix A2) are selections that influence the functional form of the model. They are not the parameters or “weights” changed during the training process, although they do influence the final number of parameters. Its choice also influences the final model accuracy. Therefore, it is critical to choose the best hyperparameter combination possible.

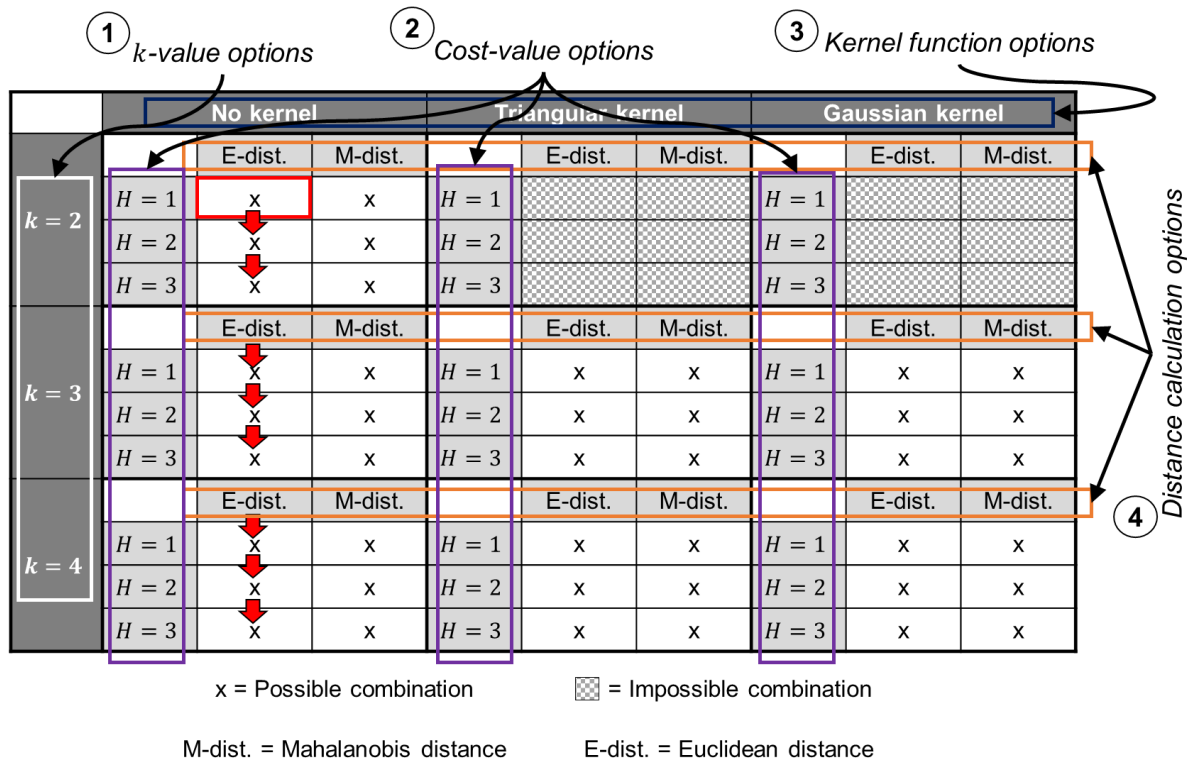


Figure 2.10: Illustration of grid-search through combinations of hyperparameters for KNN

Most hyperparameter selection techniques are based on some structured form of “trial-and-error”. Once such common technique is referred to as grid-search [77]. This concept is illustrated in Figure 2.10, using KNN as an example.

There are four different types of hyperparameters which must be chosen in this case. For each of these hyperparameters, there are different options which can be chosen from. These options can be used in different combinations and can be represented in a grid. Not all combinations are possible, e.g. for the triangular and gaussian kernel functions, k must be greater than two.

A simple form of the grid-search algorithm involves training and evaluating the model for each combination of hyperparameters. Only a portion of the total data available is used to train each

model option which is referred to as the training set. The remaining portion is referred to as the testing or validation set and is used to calculate the error through some metric (a standard calculation technique). The combination with the smallest error value is then chosen. This approach is referred to as cross-validation [61], [134]. A common split during cross-validation is 70% for training and 30% for testing [77], [101], [135], [136], although an 80%-20% [94] and other similar splits [55], [56], [102] have also been employed.

The problem with cross-validation is that a larger training set means a smaller testing set and vice versa. In order to effectively increase the size of both the training and testing set, an improved version of cross-validation known as k -fold cross-validation has come into use [14], [46], [88], [118]. The concept behind this technique is illustrated in Figure 2.11.

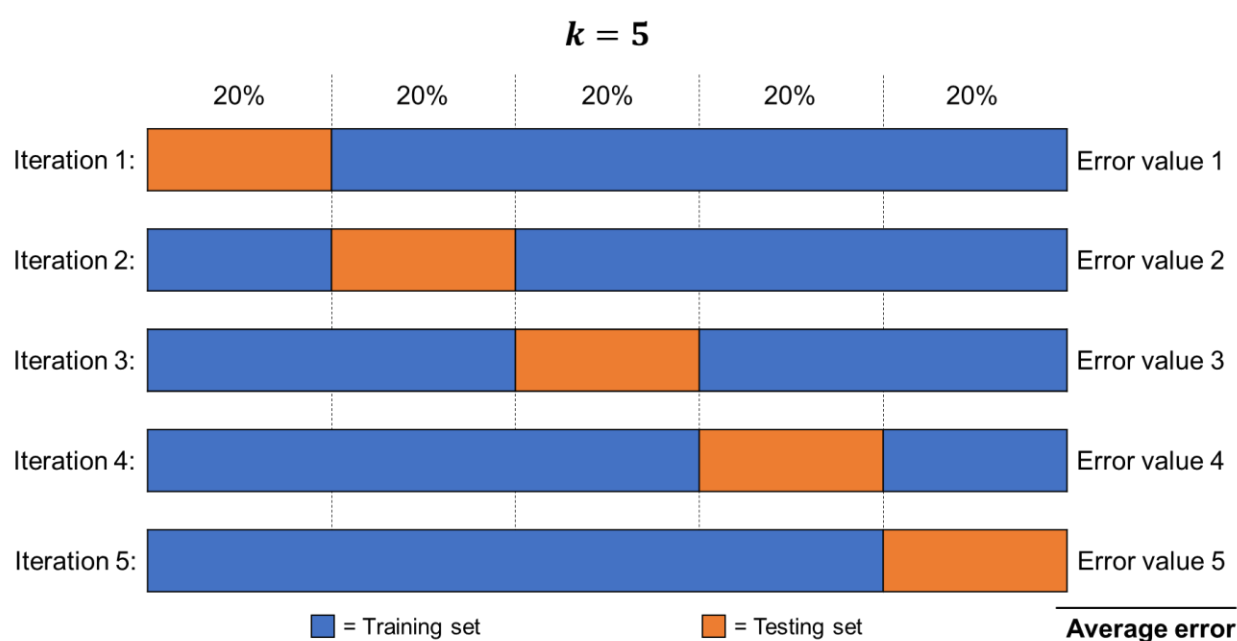


Figure 2.11: K-fold cross-validation illustration for $k = 5$

During k -fold cross-validation, the dataset is split into k equally sized datasets, normally after undergoing random shuffling. For the example given in Figure 2.11, k is chosen as five. An iterative process with five repetitions is then followed. For each iteration, a different portion is reserved as the testing set and the remaining datasets are combined to form the training set. The chosen error metric is calculated for each iteration and the average error for the different iterations is then calculated at the end.

The user must choose the number of equally sized datasets. In this study, five sets were used to obtain an 80%-20% split for each iteration. MAPE was chosen as the error metric. The formula for its calculation is given in Appendix B2. The method is applied separately to each model type and repeated for each of the driver sets. Based on the standard M&V process, this step is based solely on the baseline dataset. The different alternatives for the different hyperparameters that will be used in this study are given in Table 2.2.

Table 2.2: Different hyperparameter alternatives for each model type

SRM	Hyperparameters	Alternatives
LIN	Intercept	With intercept / without intercept
ANN	Number of neurons	1:11 (in a single layer)
	Activation functions	Hyperbolic tangent
SVM	Kernel function	Linear, Radial, Sigmoid
	Lambda	1, 2, 3
	Cost	1, 2, 4, 10
KNN	k -value	1, 2, 3, 4, 5
	Calculation method	Euclidean, Mahalanobis
	Kernel functions	Unifrom, Triangular
	Lambda	1, 2, 3

The hyperparameter selection must be completed for each combination between the respective model types and the different driver sets. Generally, if there are M model types and S driver sets, the selection process must be repeated for $M \times S$ times. In this study, there are 24 combinations (six model types and four driver sets). This can be represented using a flow diagram, as done in Figure 2.12.

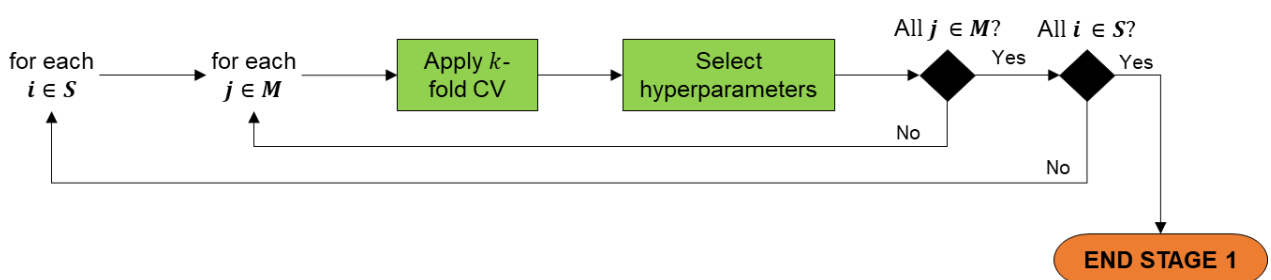


Figure 2.12: Illustration of the hyperparameter choice step

Once the hyperparameters have been selected, all the steps necessary to prepare for final model training will have been completed.

2.6 Step 5: Model training

Training machine learning models using more common software like Microsoft Excel is not practically feasible, especially if multiple repetitions are involved. The R-programming language has become a popular language for statistical processing due in part to the many external packages available [137]. Training these models in the R-programming environment further eases the processes of prediction and evaluation of the models. Therefore, model training was completed using different packages available for the R-programming language. The exact packages used are provided in Appendix B1 with additional details on the model training, specifically model-averaging.

The standard approach for machine learning applications is to use only a portion of the data available for training and the rest for testing, as explained for cross-validation in section 2.5. This is done to determine whether a model suffers from overfitting. Overfitting occurs when a model predicts the training data very well but fails to predict new data well. The concept is conceptually illustrated in Figure 2.13.

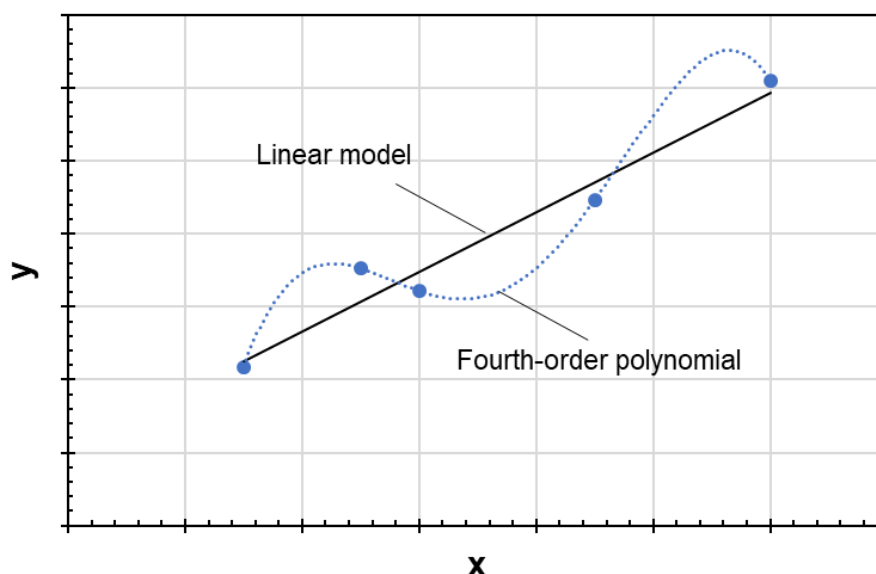


Figure 2.13: Example of overfitting of data with a fourth-order order polynomial in contrast to the true trend given by the linear model

The graph shows a random dataset with five data points used to fit a fourth-order polynomial model. Each of the points in the dataset is predicted perfectly. However, the true trend is more likely to be linear. Thus, new data points further away from the existing data points will be over or underestimated. Therefore, some data is excluded during model training so that it can be used to test the model's general applicability.

ANNs are especially susceptible to overfitting [83], [138]. The chances are higher for ANNs with larger architectures because more weights can be changed to fit the model to the data, as with the example between the polynomial and the linear model given in Figure 2.13.

In contrast to the machine learning approach, the standard approach in many M&V applications is to use all the data available for training [12], [20], [62], [121]. This method is used because the models considered in these cases are usually simple models like LIN models or variants thereof (e.g. change-point models). These models cannot be overfitted because it cannot capture complex non-linear trends.

The machine learning approach can be applied to these simpler models, but the standard M&V approach cannot be used for the more complex machine learning models. This implies that the machine learning approach (or sampling approach) must be used. However, to illustrate the effect of the different approaches, the models are also trained using all the data. These approaches are respectively referred to as the sample-data and all-data approaches.

When using the sample-data approach, the final model constructed for a specific model type and driver set will differ based on the data sampled for model training. In order to obtain a more general or average result, a procedure is adapted from Jain *et al.* where the model is trained N times by resampling 70% of the data each time with replacement [118]. This is illustrated in Figure 2.14.

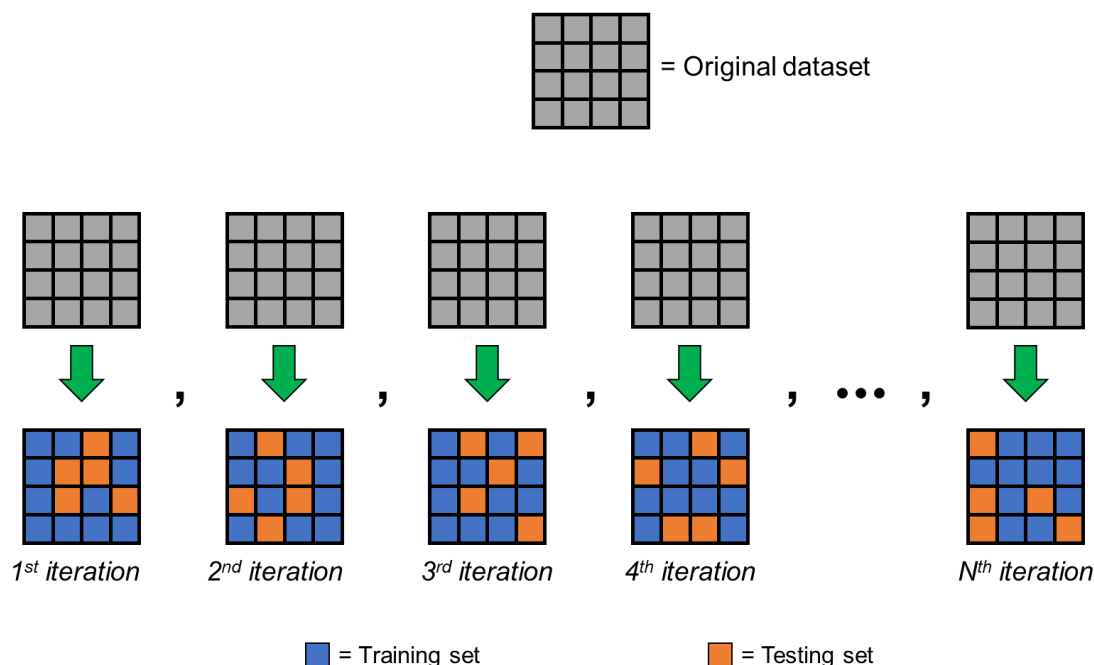


Figure 2.14: The iterative sampling process to form a training and testing set

The example shows a dataset with 16 data points, represented in a square grid. During each iteration, 11 data points (70%) are sampled from the original dataset to form the training set (blue), while the remaining five data points form the testing set (orange). The data points sampled differ in each iteration, but each pair of iterations has certain common data points in the training set. For example, the training sets for the first and second iteration both include the data point in the top-left corner and other data points also agree.

The training set from each iteration is used to train the model so that N individual models are constructed for each specific model type using each specific driver set. These are referred to as the sample models. Each of these models will then be used for predictions (of either the testing set or the assessment dataset) and the average predicted value for each data point will be determined at the end. For this study N is set to 100. When both the all-data models and the sample models has been created, these models can be evaluated

2.7 Step 6: Model evaluation

As discussed in Chapter 1, the models are to be evaluated for (1) accuracy and (2) consistency, of which both are referred to as model attributes. However, as explained in section 1.5, there are different ways to evaluate models for each of these attributes.

The standard approach of calculating different errors was chosen to evaluate accuracy. Two popular metrics that is often used in combination are the coefficient of variation for root-mean squared error (CV(RMSE)) and the normalised mean bias error (NMBE), because they highlight different aspects of model accuracy [25], [35], [84]. CV(RMSE) provides information on the absolute-error sizes, i.e. how closely the true and the predicted values are to one another, irrespective of the direction of the errors. The NMBE predicts the bias of the model – whether the predicted values tend to be greater or less than the true values. The calculation of each chosen metric is based on the residuals across the entire prediction range. The formulas for the calculation of these metrics are given in Appendix B2.

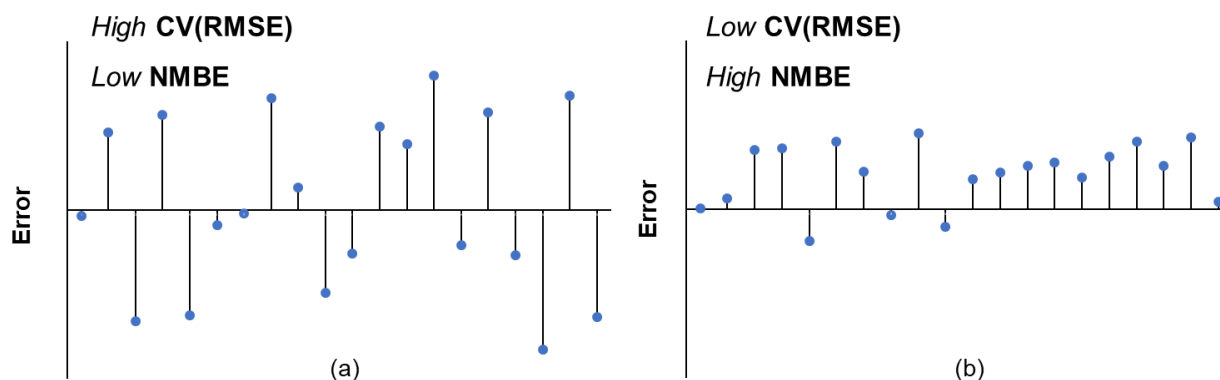


Figure 2.15: Example of the residuals of data leading to (a) high CV(RMSE) and low NMBE and (b) low CV(RMSE) and high NMBE

The conceptual difference between the CV(RMSE) and NMBE can be demonstrated by visualisation of the residuals for a dataset, as done in Figure 2.15. If the residuals are close in magnitude and are equally distributed between positive and negative, the NMBE, and thus the bias, will be small (see Figure 2.15(a)). In this case, it will be more likely that residuals in the final prediction will cancel each other out during savings calculations. However, if the absolute magnitudes of the residuals are significant, the CV(RMSE) will be high even if the NMBE is small. If the magnitudes of the residuals are lower, but tend to positive rather than negative values, the CV(RMSE) will be lower than in the first case, while the NMBE will be higher (see Figure 2.15(b)).

The second model attribute to be evaluated is consistency. As explained in Chapter 1, the consistency of the model relates to the model's sensitivity. Two aspects that influence the sensitivity of the model were included:

- (i) Sensitivity to the data sampling process;
- (ii) Sensitivity to the driver set used.

During the sampling approach to model training, a set of 100 different values is created for the savings and each of the error metrics. The model's sensitivity towards the data sampled can be determined by calculating the standard deviations for each of these value sets. The standard deviation is a measure of the spread around the mean of normal distribution. This is illustrated in the density plot given in Figure 2.16.

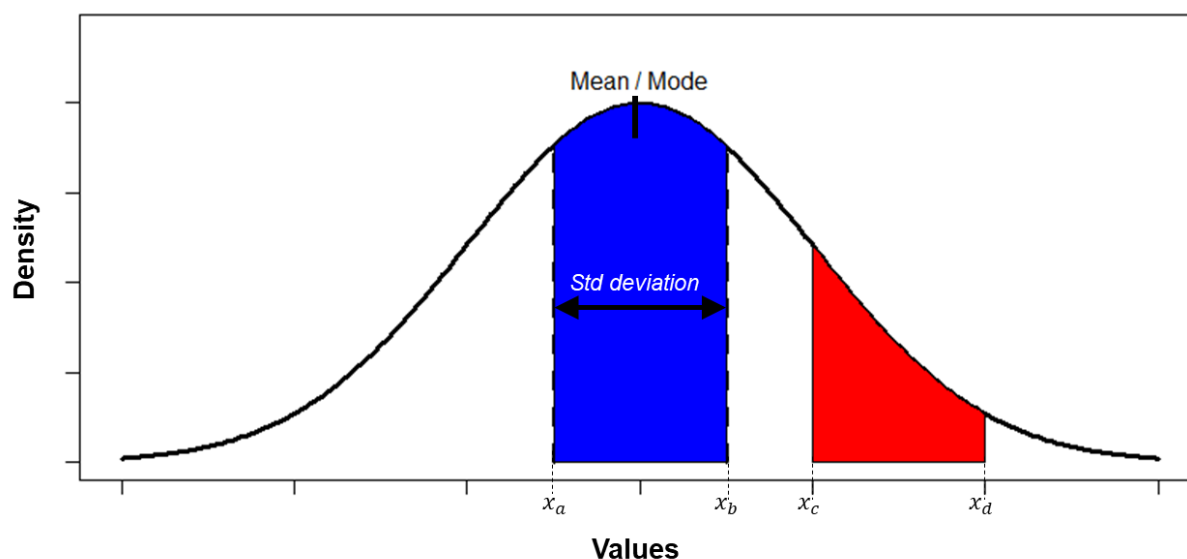


Figure 2.16: Example of a density plot showing a normal distribution

A density plot represents the distribution of data points for a series of values over the range provided. The greater the portion of values that are within a certain distance of a certain value, the higher the metric's density at that value. The portion of values that occur between two values is proportional to the area under the curve between the values. In Figure 2.16, for example, a much greater portion of the values in the series falls between x_a and x_b than between x_c and x_d , even though the range included in each case is identical.

Most random sampling processes create a normal distribution where the values are distributed equally on both sides of the mean [139]. This is characterised by a bell-shaped density curve, as shown in Figure 2.16. The peak of the curve occurs at the mean (or average) of the series. The standard deviation is the distance between x_a and x_b and contains 68% of the data. Thus, the wider the density curve is, the greater the standard deviation. The greater the standard deviation, the greater the model's sensitivity to the sampling process because the distribution of values is greater. The formula for calculating the standard deviation is given in Appendix B2.

The standard deviation can also be determined for the series of savings calculated using the different driver sets. Although the series only includes four values (one for each driver set), the standard deviation still gives an indication of the spread of the savings calculated. If the different savings calculated lie closer together for a certain type of model, the model is less sensitive to the driver set selected.

A third measure of consistency is also determined, namely the difference between the “forward” and “backward”-calculated savings. This test assumes that there is supposed to be a de-savings (negative savings) in the baseline period that is proportionally equal to the normal savings calculated. This concept is illustrated in Figure 2.17.

The diagram in the figure shows the two variations for steps three to five of the standard M&V process described in section 1.2. On the left-hand side the standard M&V process is given where construction of the model is based on the baseline period (step 3) and the energy usage is predicted for the reporting period (step 4). The savings is then calculated in the usual manner (step 5). On the right-hand side the “backward” savings or de-savings calculation is shown. Here, the data from the reporting period is used to construct the model (step 3) while the energy usage is predicted for the baseline period. Thus, the role of the baseline and reporting periods are switched around. The terms for the savings calculation are then also switched around to calculate the de-saving (step 5). The de-savings need to be adjusted then to the according to the length in the baseline period. The savings and de-savings can then be compared to determine the consistency of the model type investigated by calculating the percentage difference between the two results.

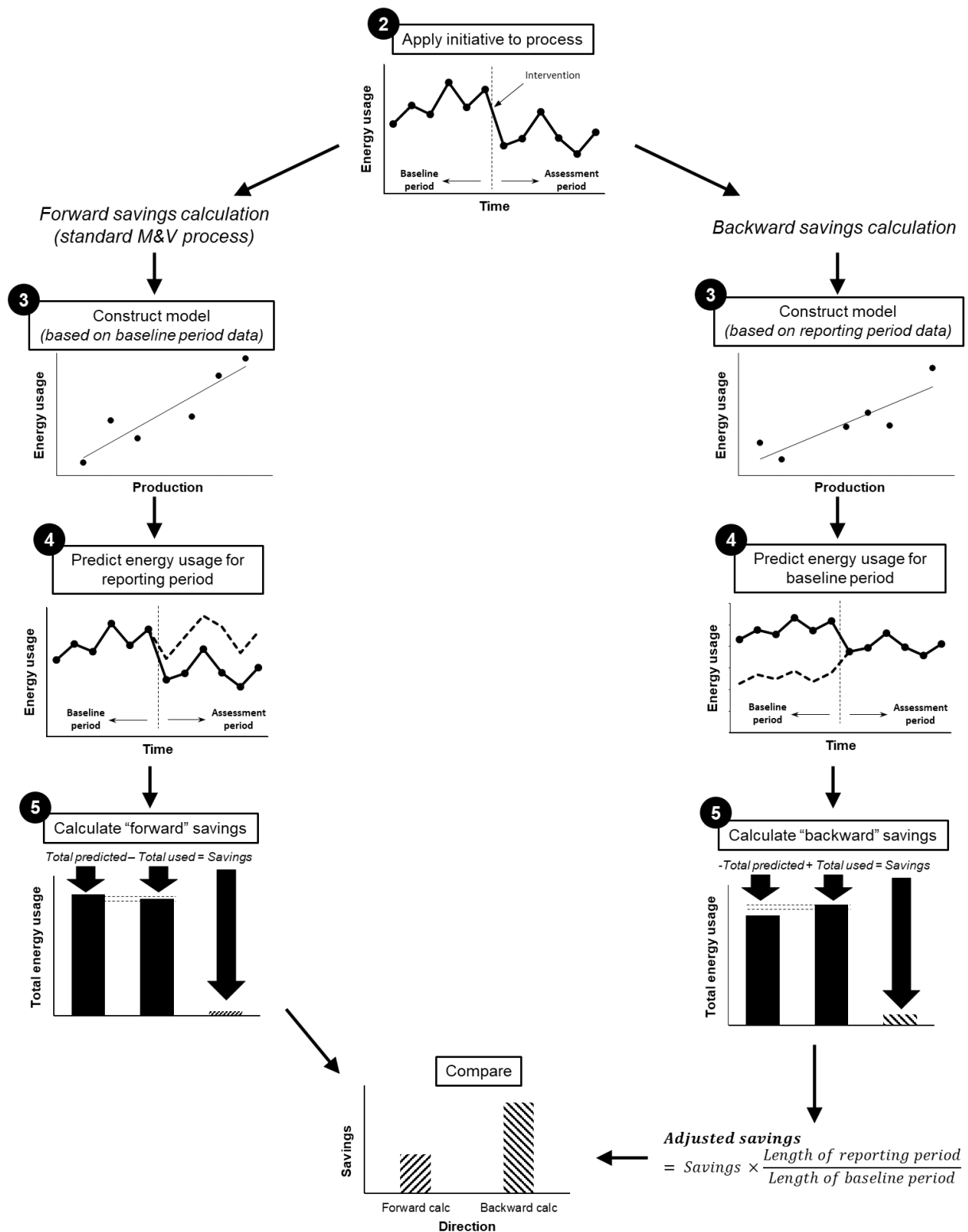


Figure 2.17: The concept of back-casting to determine savings

The results obtained from the different approaches can be effectively visualised to obtain different insights regarding the different model types. A series of plots are created to do this. These include:

- (i) A bar plot comparing the forward and backward-calculated savings for both approaches;
- (ii) Density plots for the savings, CV(RMSE) and NMBE;
- (iii) Scatter plots of NMBE vs CV(RMSE) for both the all-data and sample-data approach.

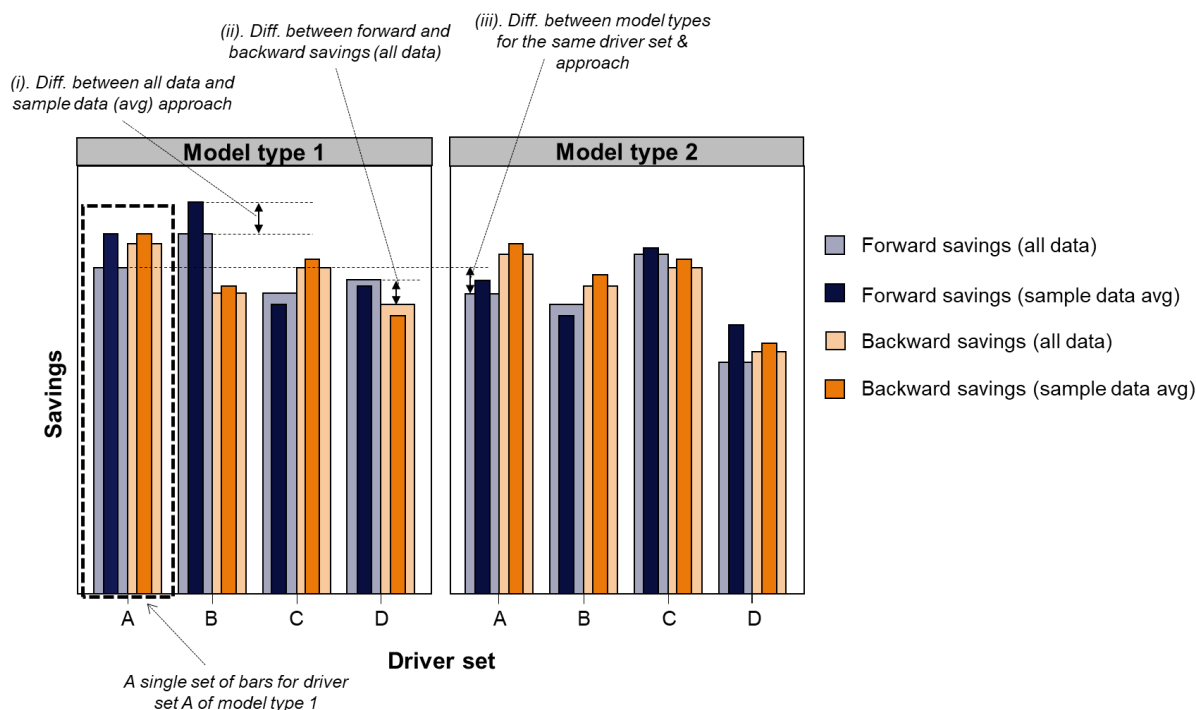


Figure 2.18: An example of a bar plot comparing the forward and backward-calculated savings for both approaches for each driver set of each of the different model types

An example of the suggested bar plot is given in Figure 2.18. The example is set up for two generic model types, type 1 and type 2. Each model type is shown in a separate block while sharing a common y-axis. The block splits the bars into four sets that correspond to the different driver sets. A set contains two pairs of bars – a pair of blue bars on the left for the forward-calculated savings and a pair of orange bars on the right for the backward-calculated savings. For each pair, the value calculated using the sample-data approach is shown as a narrower bar in a darker shade in front of the wider showing the savings calculated using the all-data approach.

This type of plot shows three things (as illustrated in Figure 2.18):

- (i) The differences between the all-data and sample-data approach for each driver set in each direction for each model type;

- (ii) The differences between the forward and backward-calculated savings for each approach for each driver set for each model type;
- (iii) The differences between model types for each driver set and each approach in each direction.

This visualisation represents a lot of information on a single plot. As such, it can be used as a tool to determine whether a certain model type stands out from the rest or whether values appear to be distributed randomly. It also provides a general overview of the consistency of the different model types. The closer the savings of the different for options for a certain model type are to one another, the more consistent the model type is.

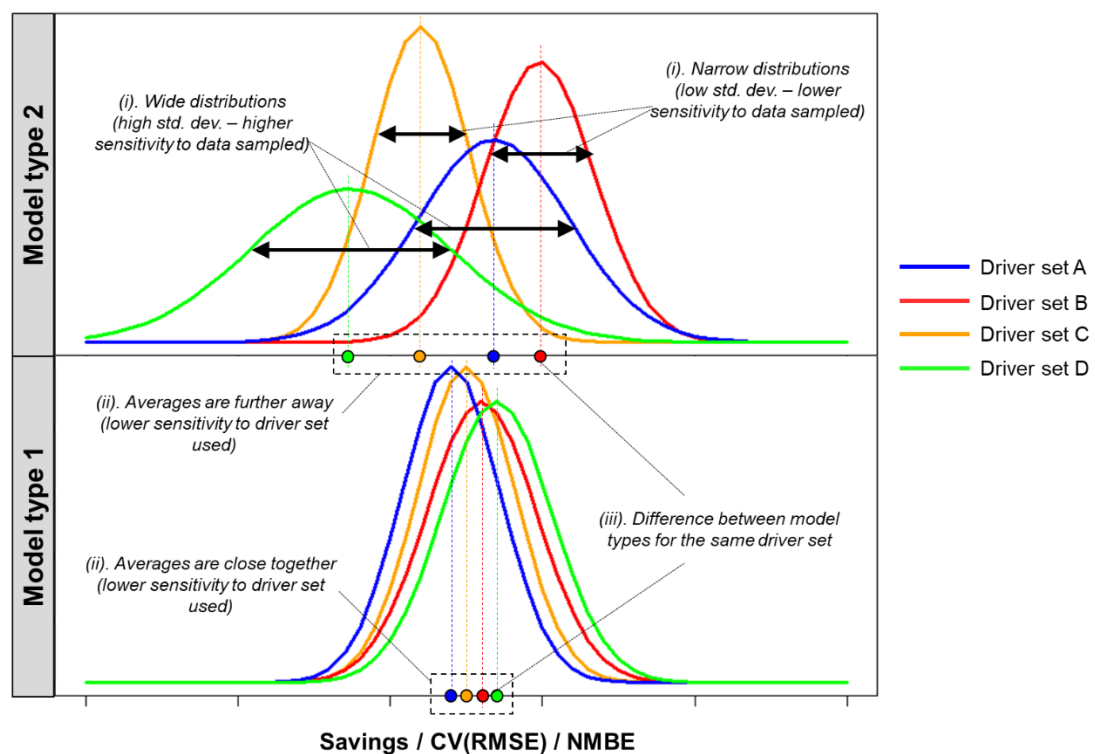


Figure 2.19: A generic example of a density plot comparing the distributions for the different driver sets for each model type

A similar example is provided for the density plots to be constructed in Figure 2.19. Again, it is set up for two generic model types in different blocks sharing a common x-axis. The density curves for the set of values generated from the sample iterations are shown for each driver set for each model type.

Three different aspects can be gathered from the plot (as shown in the figure):

- (i) The distribution width of each driver set for each model type thus indicating the sensitivity to the sampling process;
- (ii) The spread of the average savings for the different driver sets of a model type thus indicating the sensitivity towards the driver set used;
- (iii) The differences in the average savings between model types for a specific driver set.

The plots thus provide a way to compare both the averages and sensitivity of different models for different driver sets. These distribution plots are only set up for the forward-savings (and error calculations) as the purpose of the backward approach is only to determine the model consistency and is captured in the bar plot shown in Figure 2.18.

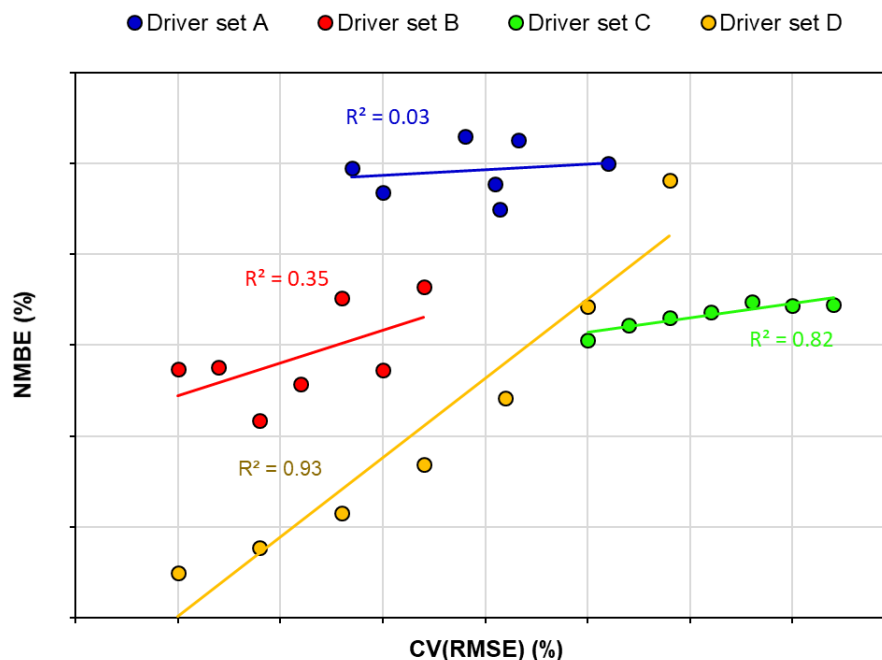


Figure 2.20: Scatter plots for the evaluation step

A scatter plot is created of the NMBE vs the CV(RMSE) for each model type, as done by Gallagher *et al.* [46]. This is to determine if any relationships exist between the error metrics for any of the model types. An example of this plot is shown in Figure 2.20.

Many different possibilities exist for the relationship between the two metrics. These possibilities can be classified under four categories – (1) no relationship, (2) weak to moderate linear relationship, (3) strong linear relationship and (4) a non-linear relationship [62], [74], [139]. Each of these is illustrated in Figure 2.20 using driver sets A to D.

If the plotted points are scattered regularly along both axes around a central mean, then it can be said that there exists no relationship (shown with driver set A). This is because a two-dimensional normal distribution is formed which is characteristic of independent variables [139]. Therefore, the R^2 -value is typically very low (<0.1). For a weak to moderate linear relationship, the R^2 -value will fall between 0.1 and 0.5. This is true for driver set B. For any R^2 greater than 0.5 the relationship can be considered as a strong linear relationship and specifically as a very strong relationship if the value is greater than 0.8 [140]. A non-linear relationship cannot necessarily be identified based on R^2 alone. Rather, it should be identified based on intuitive inspection of the plot. This is true for driver set D which shows an exponential relationship between the two variables (although the R^2 -value is high).

The use of plots can help the practitioner to identify patterns that might otherwise have been missed based on the inspection of the values of the various metrics. However, the use of plots alone does not provide enough information for model evaluation, because the information that can be gathered from these plots still need to be interpreted in a structured manner to evaluate the model type.

2.8 Step 7: Model-scoring and interpretation of results

Determining a single best model based on both accuracy and consistency will always be subject to some degree bias. Based on the assumption that these attributes are independent of one another, there will always be some form of scoring involved that will give more merit to one of the attributes. One such example is a matrix scoring method developed by Botes [62]. Even within each individual attribute, determining the best model type will involve some level of bias towards a certain aspect contributing towards that attribute.

In an effort to eliminate as much bias as possible, the method used here does not aim to determine an overall best model (winner). Rather, models are evaluated in separate categories that highlight different aspects of the model. Each category consists of different metric(s) calculated for each model, using each driver set separately. These categories are indicated in Table 2.3 along with the metrics used.

Table 2.3: Scoring categories and metrics for final evaluation

Category	Metric
Accuracy	CV(RMSE) NMBE
Sensitivity to driver set used	Standard deviation of savings across driver sets
Sensitivity to sampling process	Standard deviation of savings for sensitivity repetitions Standard deviation of CV(RMSE) for sensitivity repetitions Standard deviation of NMBE for sensitivity repetitions
Consistency test	Consistency difference between forward and backward-savings

A metric winner is determined for each driver set in each case study by finding the model with the minimum value for that metric. Once the metric winners are obtained, the category winner is determined as the model that achieved the most metric wins. To avoid the complication of having to decide between ties, no category winner is selected when two or models achieved the same number of metric wins. This simple equal-weight system is used to avoid bias introduced by a scoring system.

It is recognised that even with the equal-weight system used here, there is still a level of bias. Equal merit is given to each driver set and to the different metrics used within the categories for accuracy and the sensitivity to the sampling process. However, the simplicity of the method used results in simplicity of the interpretation of the results.

As mentioned, the standard M&V procedure is to use all the available data for model training and base errors and other metrics on these models. However, as explained earlier, given the purpose to evaluate the effect of different model types and avoid overfitting for the ANN model type, all the metrics used in the final evaluation was based on the sampling-data approach. Also, except for the consistency test, only the values in the forward direction were used as this standard approach for M&V, while the backward calculation was introduced solely to test the consistency.

Once the different category winners are determined, it should be determined if the results can be explained by knowledge about the ESM, the system to which it was applied and the different model types. Even though purely statistical metrics can be calculated, the intuitive nature of M&V is continuously recognised in literature and should thus be applied in the evaluation procedure.

2.9 Final implementation and summary of procedure

The entire evaluation procedure discussed here was developed using the R-programming language, and the benefits of using such a system were explained. However, the developed

procedure remains generic and thus can be transferred to any digital environment with the ability to perform the statistical methods given. The procedure is summarised in this section to provide clarity and to place each of the different steps into context.

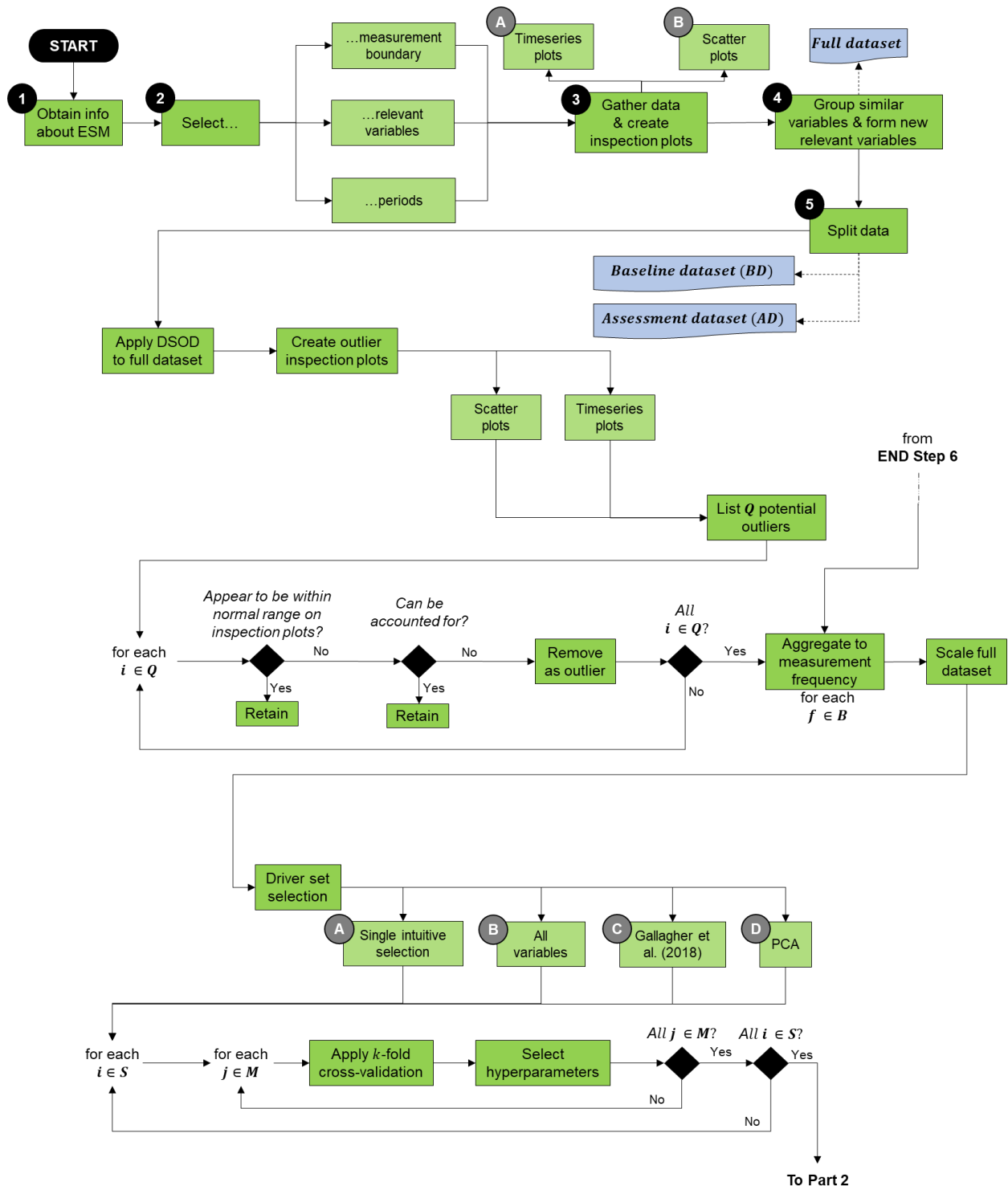


Figure 2.21: Part 1 of the methodology

The first part of the procedure is presented in Figure 2.21. The M&V practitioner must possess enough knowledge of the ESM to apply the intuitive steps included in this procedure, and should therefore first obtain as much information as possible regarding the ESM. Based on this knowledge, the practitioner can select a relevant measurement boundary, all relevant variables and both the baseline and reporting period. The applicable data for the periods and measurement boundary chosen can then be gathered.

If the data includes variables of different measurement frequencies, the variables at higher frequencies must then be aggregated to lower frequencies as part of the data gathering step. Different time-series and scatter plots can then be constructed to identify major events that occurred at the facility. Similar variables can then be grouped together and new potentially relevant variables can be calculated from others (such as duties, totals, etc.). After the data has been split to form the baseline and reporting datasets, step 1 will have been completed.

Step 2 entails treatment of the data. The DSOD method is applied to flag potential outliers. These points are then indicated on scatter and time-series plots used for inspection. Each potential outlier is evaluated separately to determine whether it falls within the normal range of operation in the inspection plots or whether it can otherwise be accounted for. If this is so, the data point is retained but is otherwise removed as an outlier from the dataset. After this step, the dataset can be aggregated to form new measurement frequencies as determined by the practitioner. However, each of these new datasets is scaled independently and the process is continued using scaled dataset for the highest measurement frequency. This scaled dataset is then used for the driver set selection in the next step.

Four different driver sets are selected in step 3. This includes (A) an intuitive selection of a single variable, (B) selecting all the measured variables (C) choosing a single variable through the method given by Gallagher *et al.* [14] and (D) selecting the appropriate number of PCA variables with the EKF algorithm. Different hyperparameters are then selected for each combination between each of the S driver sets selected and the M model types investigated. The selection of the hyperparameters concludes the preparation for model training, which forms the first step of the second part, as given in Figure 2.22.

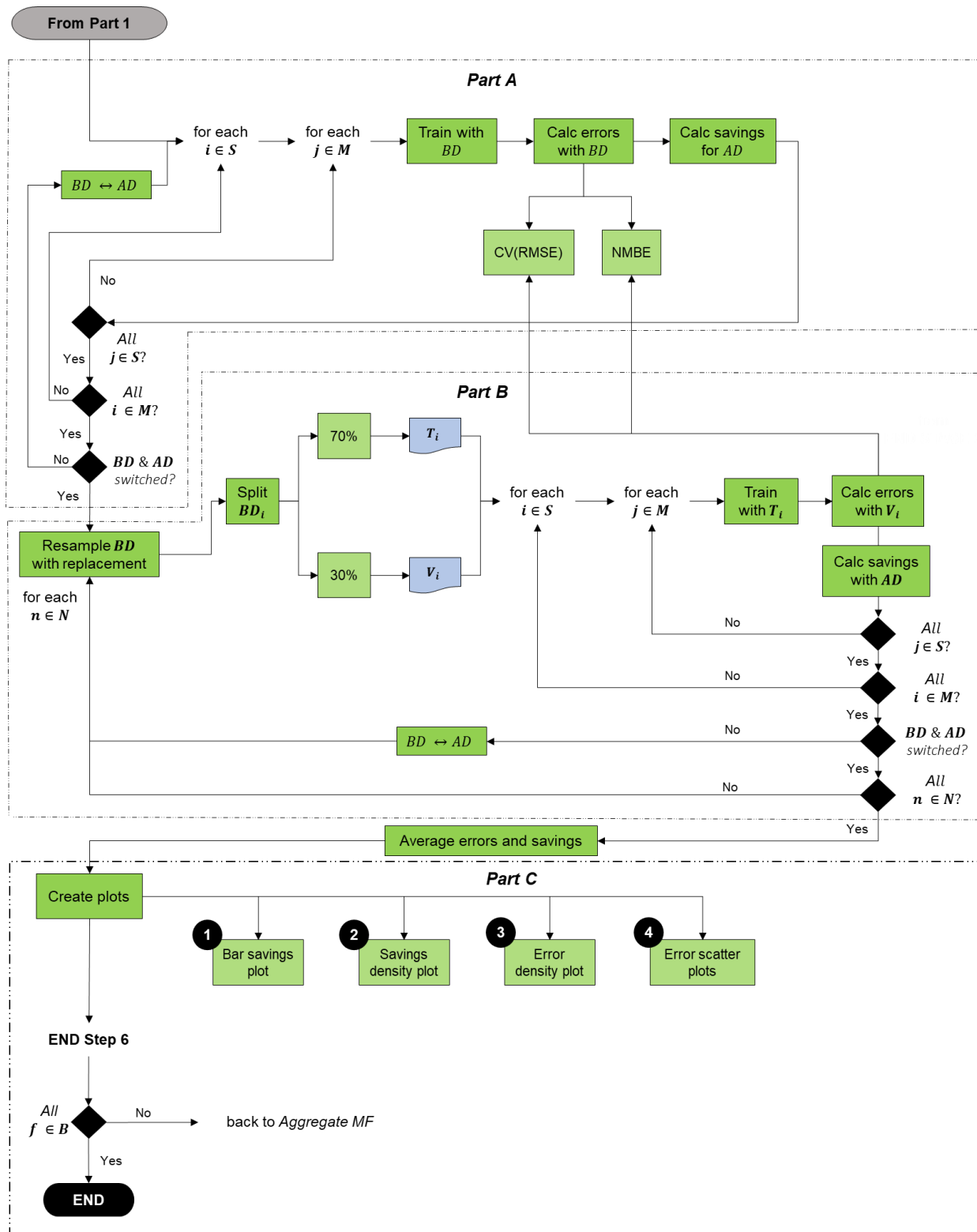


Figure 2.22: Part 2 of the methodology

For the second part of the procedure, models are trained using both the standard M&V approach and the sampling approach from the field of machine learning. These are referred to as the all-data approach and the sample-data approach respectively. The all-data approach is applied first in the procedure. This approach is repeated for each model type using each of the driver sets selected. After repeating the process for each combination, the applicable savings and error metrics are calculated, which forms part of step 6. After all the models have been trained using the different model types and driver sets, the baseline dataset and the reporting dataset are switched around to calculate the applicable de-savings for each combination. This forms part A.

The sample-data approach is done in a similar manner. However, it is repeated N times to obtain an average result. For each repetition n , the baseline data is split into a training set (T_i) and a validation set (V_i) according to a 70%-30% ratio. A separate model is trained for each model type using each driver set within the training set. The error metrics are then calculated based on the predictions for the validation set and the savings based on the entire reporting period. The baseline and reporting datasets are again switched around and the sampling process is repeated to calculate the de-savings. Only when all these calculations are completed does the procedure continue to the next iteration of the sample-data approach. This repetitive process forms part B.

A series of plots are created using the different results that were obtained. These include a bar plot of the various savings calculated and density plots for the savings calculated in the forward direction, the CV(RMSE) and the NMBE using the sample-data approach. The final plots are scatter plots of the NMBE against the CV(RMSE) for both approaches to determine if any relationships exist between the metrics. The plots created assist with the visual evaluation of the accuracy and consistency. The consistency can be evaluated based on the differences between the forward and backward-calculated savings, the standard deviations of the density curves and the standard deviation across the average savings calculated for each driver set for a certain model type. This forms part C in Figure 2.22.

Steps 3 to 6 are repeated using each of the scaled datasets created for each of the **B** measurement frequencies to be investigated and complete the explicit part of the process. The results can then be compared to determine differences introduced due to different measurement frequencies.

Finally, the models are scored within four different categories. As part of the interpretation of the results, an intuitive explanation is sought for the category winners based on the knowledge obtained throughout the procedure.

2.10 Conclusion of evaluation procedure

This chapter discussed the seven-step procedure developed to evaluate different model types according to accuracy and consistency for M&V purposes. The steps in the procedure were developed from what has been observed in literature and provided for the standard M&V approach as well as for the machine learning approach of cross-validation. Through this provision could be made to evaluate the models according to accuracy and consistency according to different metrics in a simple model scoring technique.

CHAPTER 3 DISCUSSION OF RESULTS AND VERIFICATION

3.1 Introduction to results

To answer the research questions, five real-life case studies of M&V initiatives applied in the South African industrial sector were selected and the developed evaluation procedure was applied. This chapter discusses the results of the selected case studies according to the steps given in Chapter 2. The results of most of the case studies highlighted similar findings. To illustrate and discuss these findings, focus is placed on the first case study. Other significant findings are also highlighted.

3.2 Selected case studies

The case studies selected for this study included three case studies from industrial production facilities and two case studies from the mining industry. Due to the confidentiality of the data, only critical details regarding the layout of the facilities/mines and the initiatives applied are presented.

In most of the case studies, there were several ESMs applied or changes in the ESMs occurred during different stages (making it difficult to identify a single transition period). However, the purpose of the study is not to determine the most accurate saving for a single ESM, but to determine the effect of using different models in the M&V procedure (be that for a single ESM or multiple ESMs).

Rather than trying to isolate a single ESM or remove the effect of the transition period, a simple year-on-year assessment was done for each case study. Given that multiple events and changes are characteristic of industrial systems, retaining these multiple ESMs will thus maintain a true representation of the dynamics of industrial data. The baseline and reporting years are consistently referred to as year 1 and year 2 for all the case studies. These years are fiscal years for the facilities or mines considered and not calendar years.

Case study I

The first case study is for an ESM applied to an industrial production facility, referred to here as Facility A. The facility introduced several ESM projects, of which the greatest was a change in a critical setpoint on the facility. The measurement boundary was selected as the whole facility to encapsulate all the ESMs. The highest measurement frequency available was daily (due to the

frequency of measurement of certain variables). Figure 3.1 provides a process diagram of the facility with the applicable measurement locations.

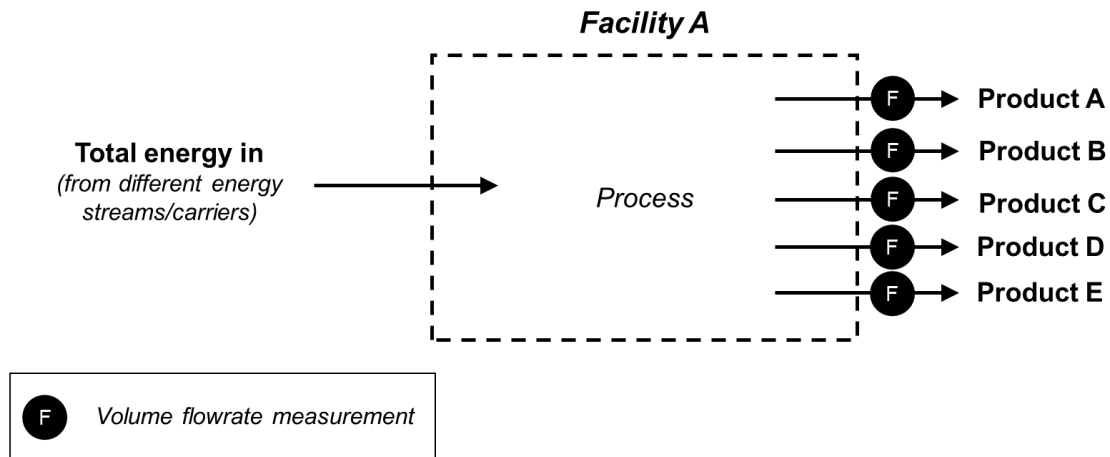


Figure 3.1: Process diagram for Facility A

The energy used by the facility comes from a variety of energy carriers, including the raw materials used to make the products. The energy content of each energy stream into the facility was added together on a non-weighted basis to calculate the total energy supply. The energy is used by the facility to produce five products. These products are referred to as product A to E and the normal volume flowrate (F) of each is measured. A plot of the product fractions against the total energy usage based on a daily frequency over the two years of interest is given in Figure 3.2.

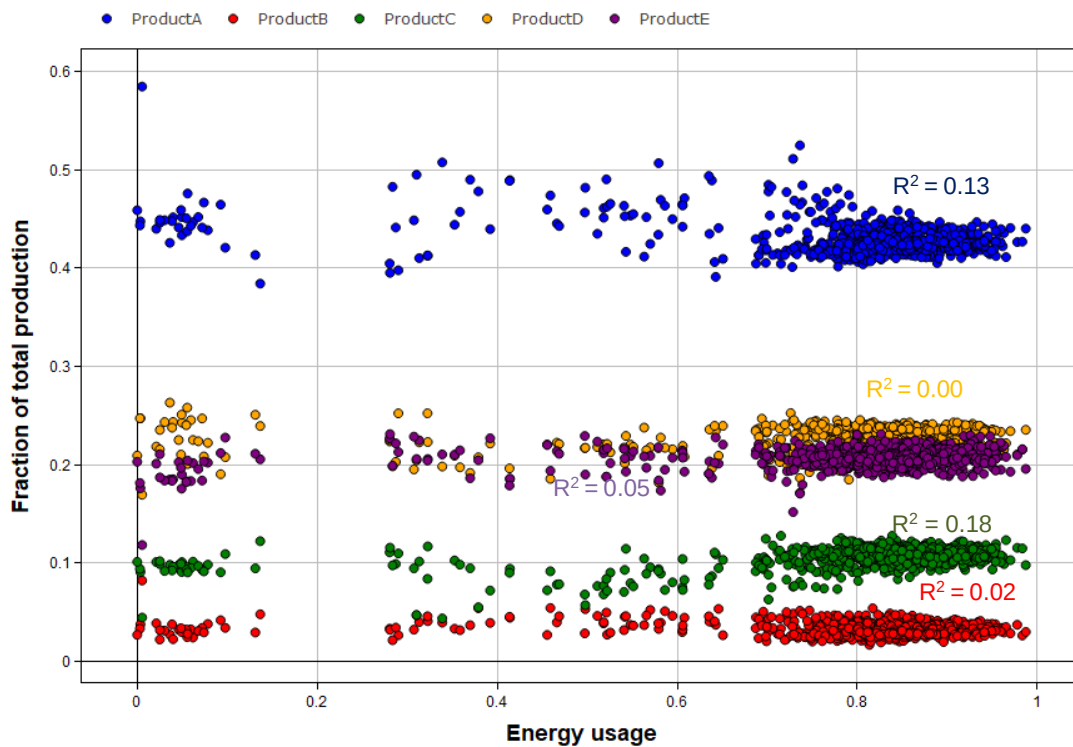


Figure 3.2: Average product distribution over the two years

The plot shows that the fraction of each product is relatively constant in comparison to one another. There is a clear separation between the fractions of each product, except between product D and E. Product A is produced most and thereafter product D, product E, product C and then product B.

Variation within each product fraction occurs within a certain band across the entire range energy usage. In many cases, a certain amount of energy usage corresponds to many different product distributions, as illustrated by the vertical line in the plot. The product distribution is thus independent of the energy usage which is confirmed by low R^2 -values (shown in the plot) and thus the energy usage is independent of product distribution. Differences in the distribution are caused by process changes and can be considered random with respect to energy usage.

Figure 3.3 and Figure 3.4 respectively show the variation in the energy usage and the production in time-series plots. The most noteworthy events seen in these plots are the multiple shutdown periods. Also, of note is the high variation in the production of product B seen in Figure 3.4.

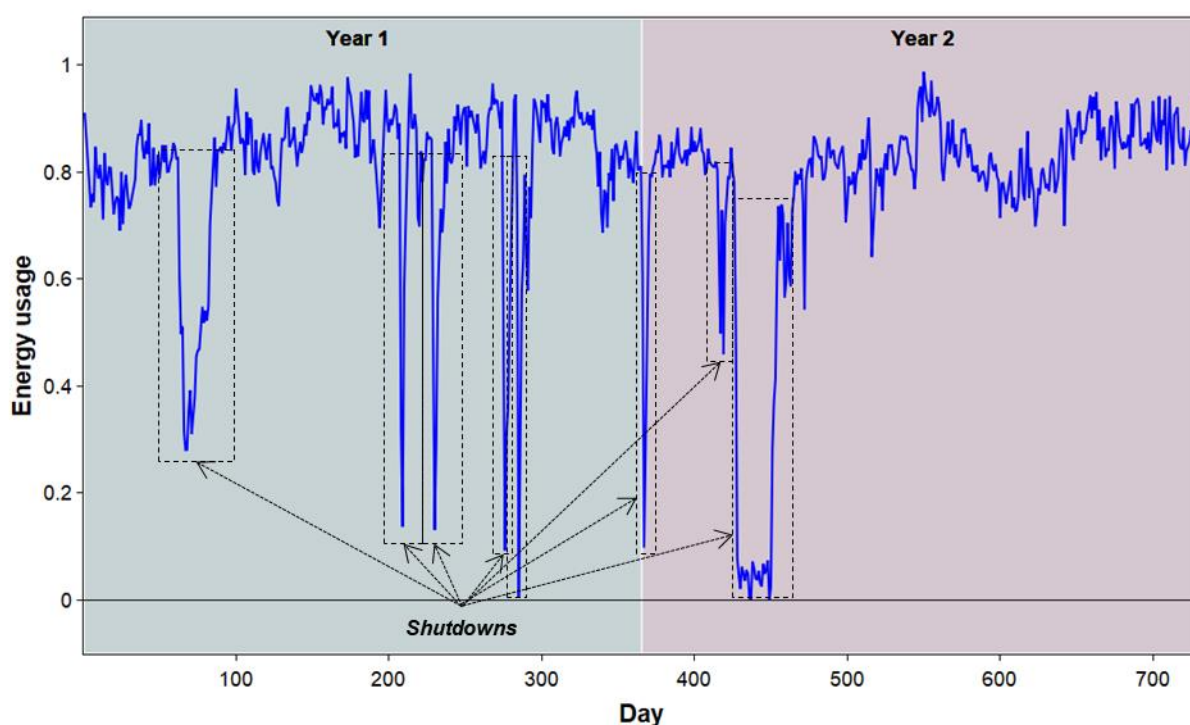


Figure 3.3: Time -series plot of energy usage (scaled)

From Figure 3.3 multiple shutdowns can be seen where the energy usage is reduced substantially; these include routine and non-routine shutdowns. The shutdowns help to add additional variation in the data beyond the variation in the more constant energy usage periods that constitute normal operation.

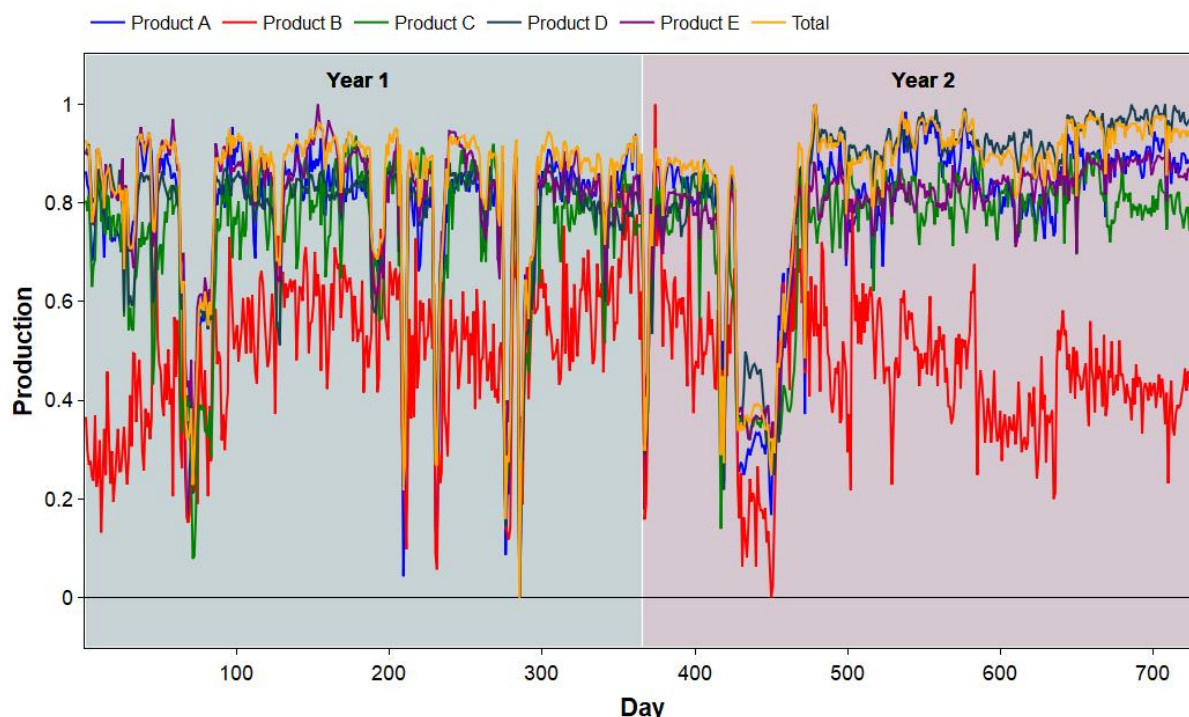
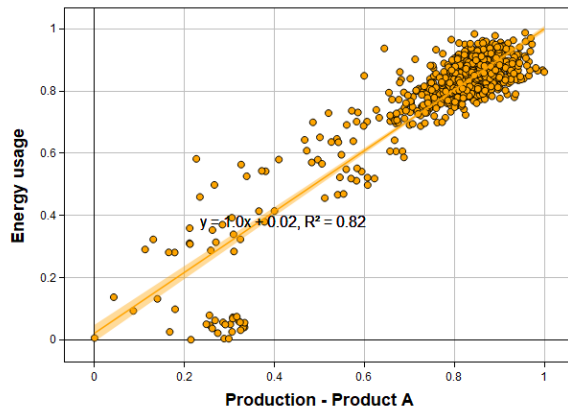


Figure 3.4: Time-series plot of the production (scaled)

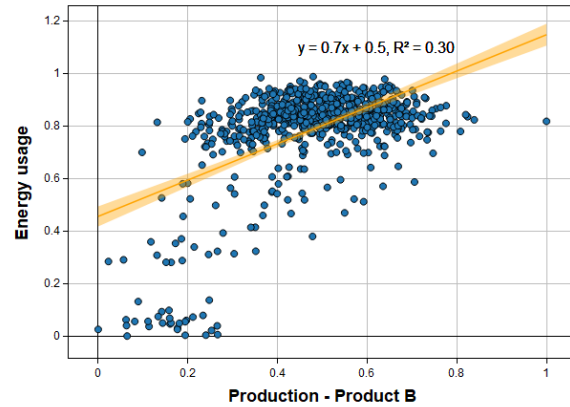
Figure 3.4 shows the scaled production values for each product as well for the total production. The same reduction seen in the energy usage can be seen in the total production during shutdown periods. The same type of pattern is also seen for each individual product except for product B.

For product B, more significant day-to-day changes can be seen. It should be noted that these changes are relative to the maximum production of product B because of the scaling applied. These relatively large changes are to be expected since product B makes up the smallest portion of the total production. A smaller relative change in another product will thus cause a greater relative change in product B and vice versa.

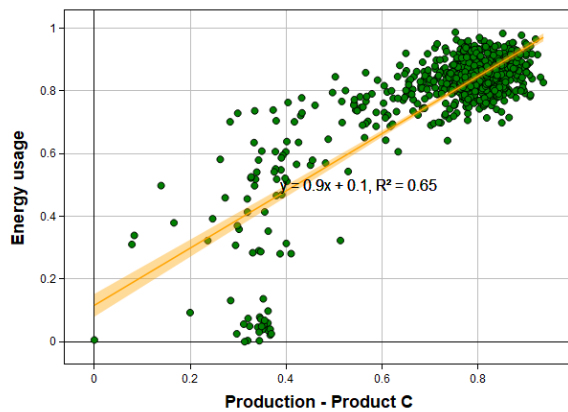
Figure 3.5 shows the relationships between the total energy usage and the production of each product as well as the relationship between the energy usage and the total production. The energy usage shows strong linear relationships with product A, product E and the total production (based on high R^2 -values).



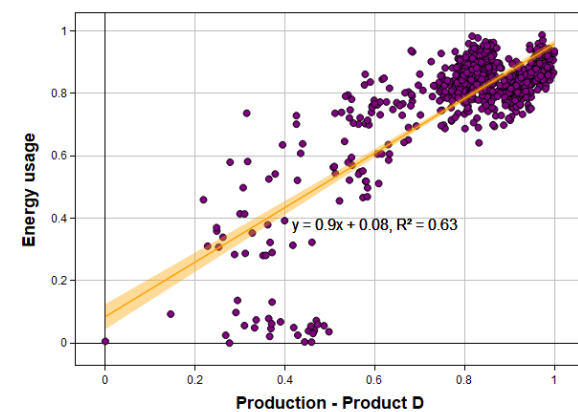
(a)



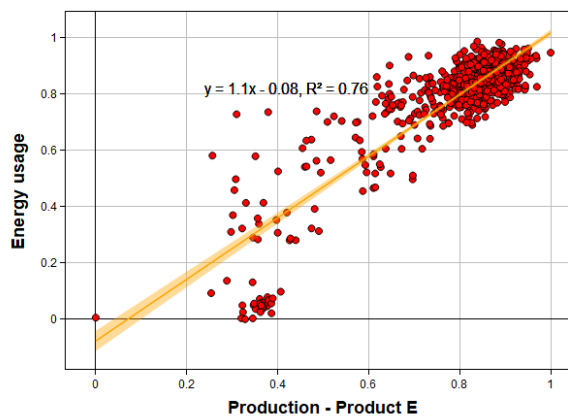
(b)



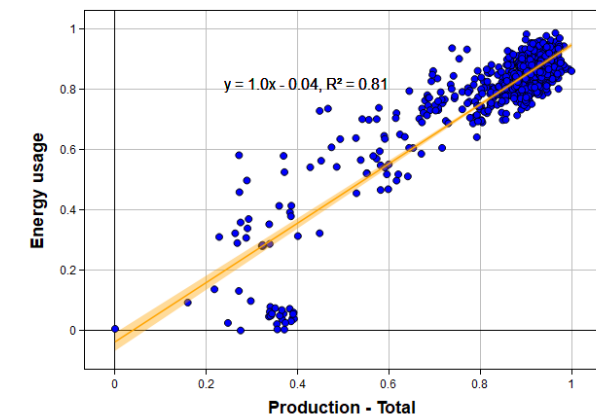
(b)



(d)



(e)



(f)

Figure 3.5: Scatter plots of energy usage vs the various drivers

Except for the plot for product B, the plots show a higher density of points at higher energy use (>0.6) and higher production (>0.6). The remaining points are scattered randomly along and around the fitted linear trend. These points are the result of the various shutdowns. The areas with the higher densities of points represent the day-to-day variation in energy supply, with a degree of randomness being caused by process changes. The weak linear relationship between the energy usage and product B is due to the greater influence of random variations on its production, as explained for Figure 3.4.

A summary is given in Table 3.1 of the applicable parameters for case study I. It provides a short system description, the potential energy drivers, the highest measurement frequency and a description of the ESM applied. The potential drivers include each of the five products and the calculated total production. Such a summary is given after the overview of each case study.

Table 3.1: Summary of case study I

Property	Specification
<i>System description</i>	Full (industrial production) facility
<i>Potential drivers</i>	Product A to E; Total production (calculated)
<i>Highest measurement frequency</i>	Daily
<i>ESM applied</i>	Several, including change in a critical setpoint on the facility

Case study II

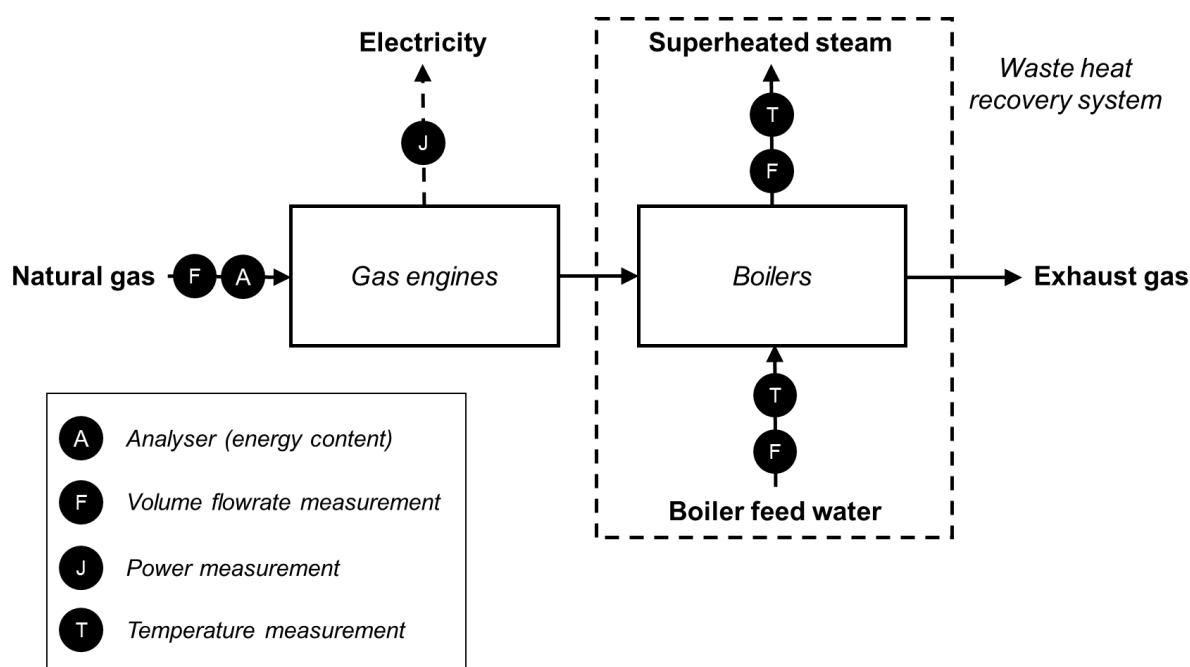


Figure 3.6: System layout for Facility B

Case study II entailed the improved operation of a waste heat recovery system on gas engines. A layout of the system is given in Figure 3.6.

The gas engines receive natural gas of which the normal volume flow (F) and the heating value are measured. The heating value is measured using an on-line analyser (A). A heat exchanger was installed in the year prior to the baseline year to use the waste gas to produce steam from boiler feedwater (BFW). Because of the locations of the measurement points, all the gas engines are viewed as a single system, including all the boilers.

The system was operated in a more stable manner during year 2. Thus, more energy was recovered from the system rather than using less energy. In this case, the terms in the formula for calculating the “savings” must be switched around because the total energy recovered from the system is predicted, which can be thought of as “negative” energy use.

$$\sum_j^N E_{save,j} = \sum_j^N E_{meas,j}^- - \sum_j^N E_{pre,j}^-$$

E^- : Energy recovered

$\sum_j^N E_{save,j}$: Sum of energy saved from period i to N

$\sum_j^N E_{meas,j}^-$: Sum of measured energy recovered from period i to N

$\sum_j^N E_{pre,j}^-$: Sum of predicted energy recovered from period i to N

(Eq. 5)

Table 3.2: Summary of case study II

Property	Specification
<i>System description</i>	Gas engines with a waste heat recovery system
<i>Potential drivers</i>	Natural gas: flow and heating value; BFW: flow and temperature; Total energy input (calculated)
<i>Highest measurement frequency</i>	Daily
<i>ESM applied</i>	Improved (more stable) operation of the system

Case study III

Facility C is a ferrochrome production facility with several furnaces. A representation of a single furnace is given in Figure 3.7 with the available measurements indicated.

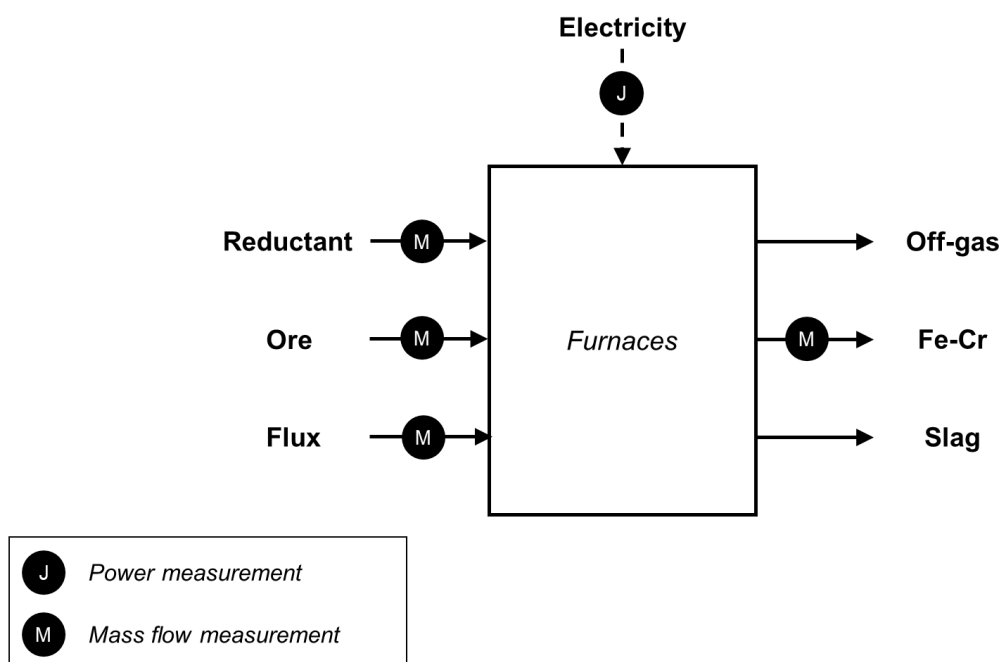


Figure 3.7: Representation of a typical furnace for industrial production facility C

The furnace receives reductant, ore and flux as raw material. Energy is supplied in the form of electricity. The main product of the process is ferrochrome (Fe-Cr) with off-gas and slag as by-products. The type of flux used was systematically changed at each furnace in the baseline year. The purpose of the M&V procedure was to determine whether a change in flux type produced an energy saving. As the ESM was applied to each of the furnaces, the measurement boundary was selected across all the furnaces. The furnaces were viewed as a single “virtual” furnace by adding together similar streams across all the furnaces.

Table 3.3: Summary of case study III

Property	Specification
<i>System description</i>	Ferrochrome furnaces
<i>Potential drivers</i>	Mass flow: reductant, ore, flux and ferrochrome product
<i>Highest measurement frequency</i>	Daily
<i>ESM applied</i>	Change in flux type

Case study IV

Mine A used four chillers to supply the different systems and operations with cold water. The ESM applied was increased operation of the more efficient chillers. A representation of a single chiller is given in Figure 3.8.

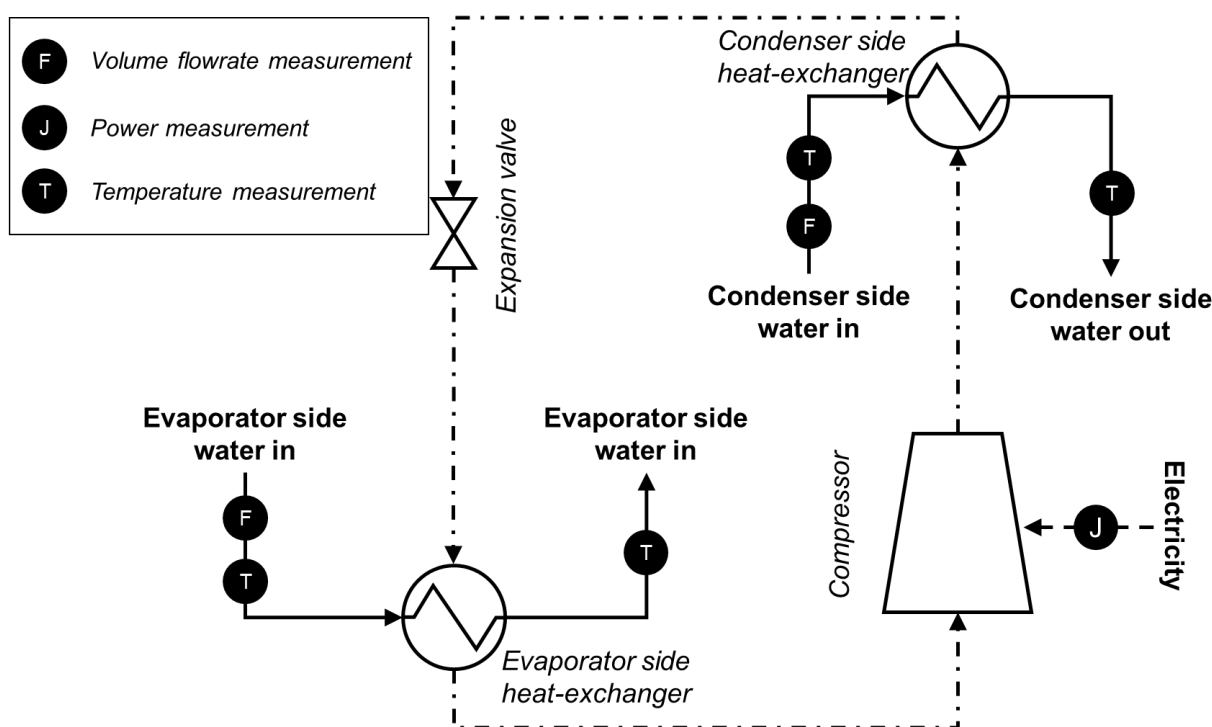


Figure 3.8: Representation of a single chiller system for mine A

The diagram illustrates the main components of the chiller – a compressor, an expansion valve and the evaporator and condenser side heat-exchangers. The chiller cools incoming process water on the evaporator-side through heat exchange with cooler refrigerant circulated in a closed-loop inside the chiller. This circulation is provided by the compressor using electricity as an energy source. The compressor also increases the pressure, which changes the thermo-physical state of the refrigerant so that its temperature increases. It can then transfer the heat picked up on the evaporator side to condenser water at the condenser heat-exchanger. The pressure of the refrigerant is then reduced over the expansion valve to reduce the temperature and complete the cycle.

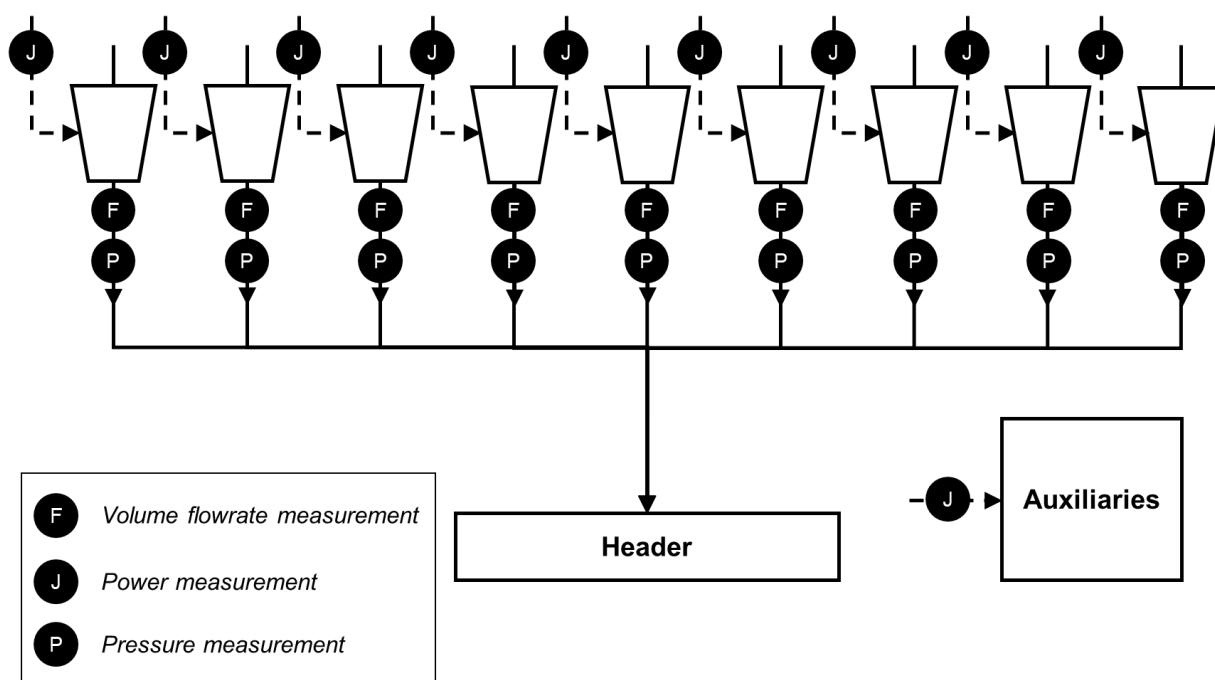
Similar to the previous case study all four chillers were treated as a single unit, by taking the sum of extrinsic variables (flow and power) and the weighted average of intrinsic variables (inlet and outlet temperatures) across the chillers.

Table 3.4: Summary of case study IV

Property	Specification
<i>System description</i>	Four industrial chillers
<i>Potential drivers</i>	Evaporator: flow, temperature in and out; Condenser: flow, temperature in and out; Total cooling delivered (calculated)
<i>Highest measurement frequency</i>	Daily
<i>ESM applied</i>	Increased operation of the more efficient chiller

Case study V

For case study V, the compressor network of mine B was used. A layout of the system is given in Figure 3.9.

**Figure 3.9: The compressor network for mine B**

The network uses nine compressors as well as auxiliaries. Each compressor compresses ambient air to supply a flow of high-pressure air to the system header. This header in return supplies the different processes of the mine. Similar to the previous case study, the mine used the more efficient compressors more than in the baseline year. Here, total flow, the average pressure (weighted) and the total ore retrieved from the mine was used as potential variables.

Table 3.5: Summary of case study V

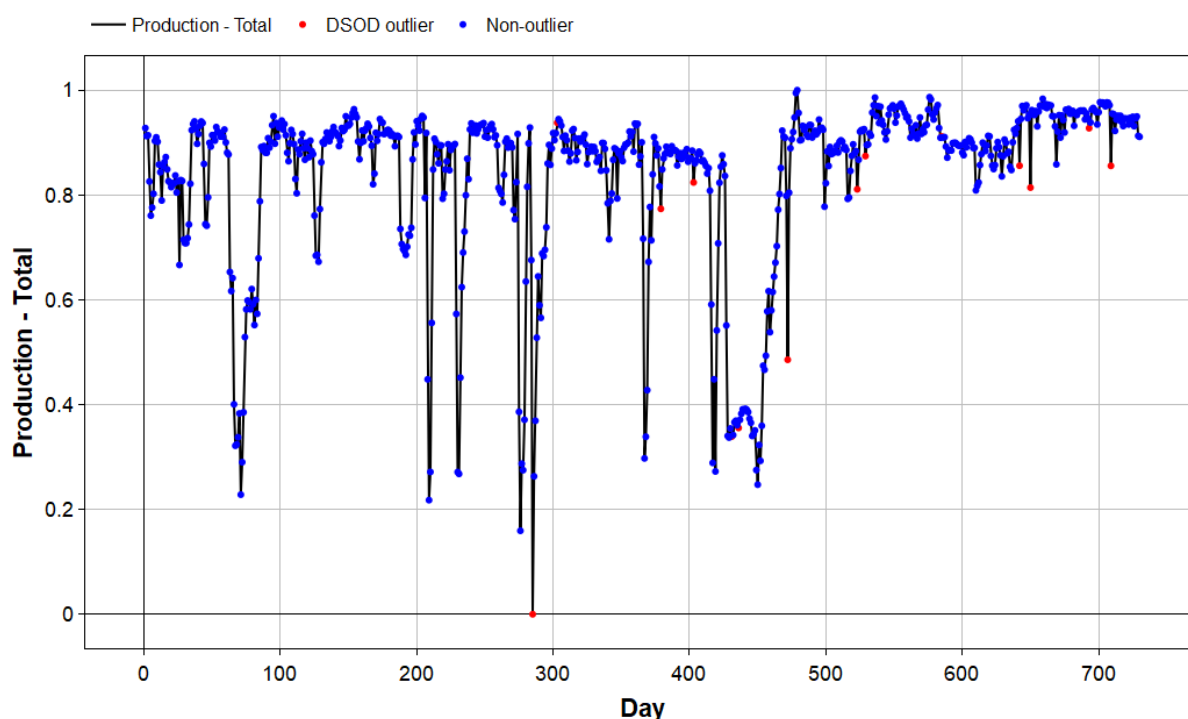
Property	Specification
<i>System description</i>	Compressor network with nine compressors
<i>Potential drivers</i>	Total flow; Total mine production (ore removed); Flow power (calculated)
<i>Highest measurement frequency</i>	Daily
<i>ESM applied</i>	Improved utilisation of more efficient compressors

3.3 Removal of outliers

As suggested in Chapter 2, the DSOD method was applied. The method makes use of several user defined parameters and the chosen values for these are given in Appendix C2.

Because the DSOD method highlights points based on unexpected paths followed (in a multivariate space), upsets or irregular operation is highlighted. However, these points are not necessarily wrong. It is possible that the measurements are correct and that these points are true but represent less likely operating conditions of the process.

To illustrate the process and the argument presented here, two of the selected case studies are used. Consideration is first given to case study I. The time-series plot of the total production is shown in Figure 3.10, while the scatter plot of the energy usage vs the total production is given in Figure 3.11. In each case the DSOD outliers are indicated in red.

**Figure 3.10: Time-series plot of the total production showing the DSOD outliers**

Eleven flagged data points can be seen in the plot. Each of the flagged points occur at a sudden drop in the production. Not all the data points occurring at sudden drops are flagged by the algorithm because the algorithm does not consider each change in isolation but takes into consideration changes before and after a point under investigation. If the change that occurs is relatively large in comparison to the changes immediately before and after it, it is more likely to be flagged. Also, a point is more likely to be flagged if the change occurred in all the variables considered.

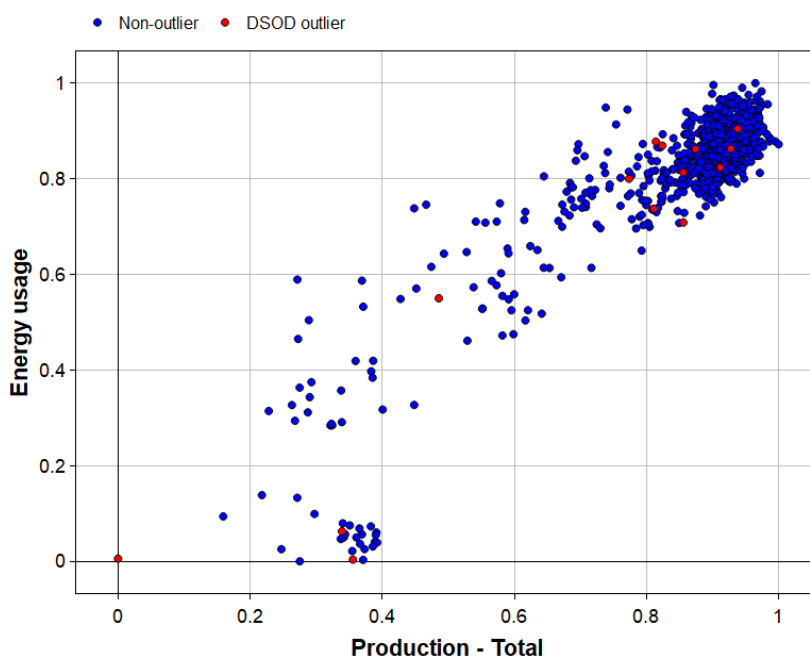


Figure 3.11: Scatter plot of the energy usage vs total production showing the DSOD outliers for case study III

The scatter plot shows the total energy usage of the facility is shown on the y-axis and the total production on the x-axis. It has already been mentioned that there is a strong linear relationship between the two variables. The points flagged by the DSOD algorithm still fall close to the normal linear trend. This suggests that the flagged data points represent less likely events but are still representative of the system. Thus, all the points are retained because no reason can be found to remove them.

Next consideration is given to case study III. The time-series plot for the ore supplied is given in Figure 3.12, while the scatter plot of the energy usage vs the ore supplied is given in Figure 3.13.

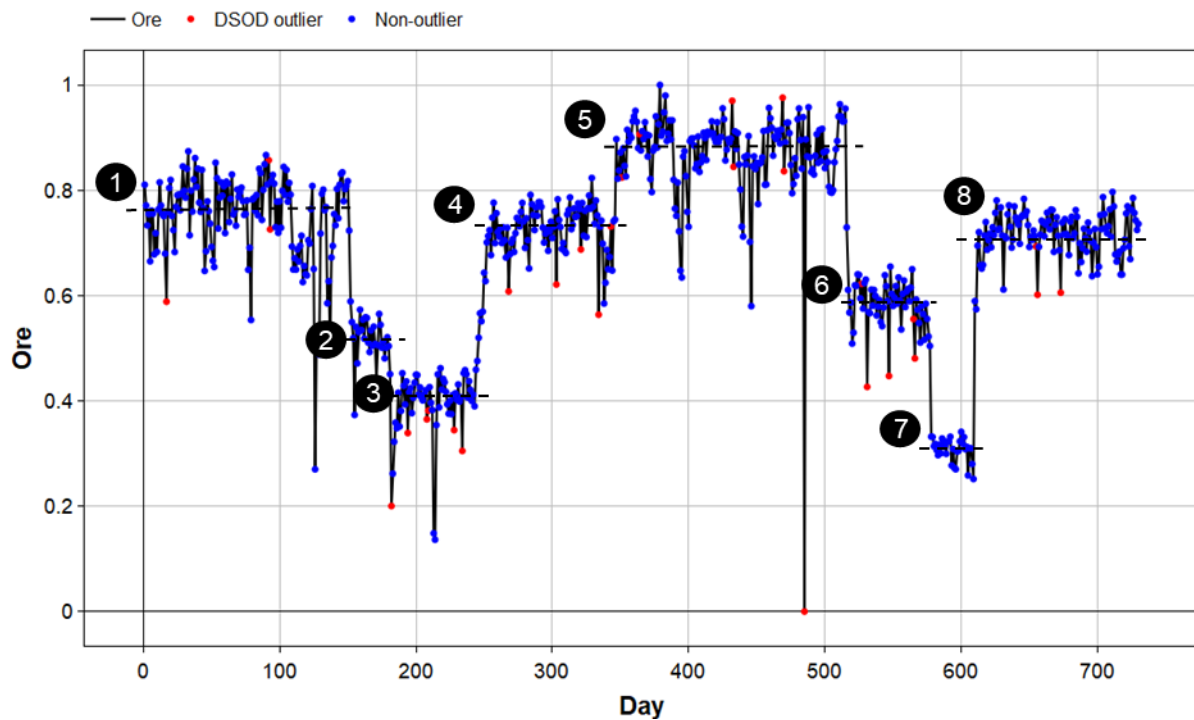


Figure 3.12: Time-series plot with DSOD outliers for the total ore

The time-series plot shows multiple step-changes that occur between different operational modes (1-8). These modes occur as different amounts of furnaces are used. The step-changes are not flagged because the process almost immediately stabilises to a new operating point. However, again it is seen that a sudden drop and recovery causes some points to be flagged as potential outliers.

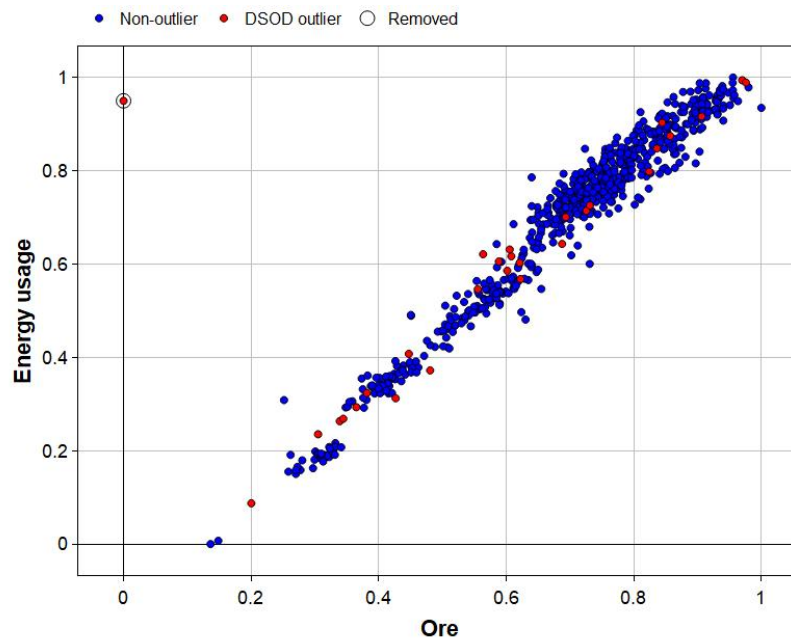


Figure 3.13: Scatter plot of the energy usage vs the ore usage showing the DSOD outliers and the point removed

The scatter plot given in Figure 3.13 shows that most of these points fall within the normal linear trend, as with the previous case study. However, there is a single point where no raw ore is supplied to the process, but the energy usage is high. This is probably due to data loss and the point is removed.

For each of the case studies, data quality was good with very few outliers removed. The plots for the remaining case studies are given in Appendix C2. After the removal of the outliers, the driver set could be done.

3.4 Selected driver sets

The techniques discussed in section 2.4 were applied to each case study to produce four different sets for each case study. These sets are referred to as driver sets A to D and the method used for each is given in Table 3.6. The four driver sets were selected in a similar manner for each case study, however, the final variables selected differed due to the differences in the processes.

Table 3.6: Selection methods for the different driver sets

Driver set	Method
A	Intuitive selection of a single driver
B	All variables included
C	Method given by Gallagher <i>et al.</i> [14]
D	PCA with EKF selection

The selection of driver sets B to D is based on to explicit methods. However, the intuitive selection of the single driver for driver set A may differ between practitioners. For case study I, the strong linear relationship between the total production and the energy usage motivates its use as a driver. The five production variables are relatively independent of one another so that fluctuations in the distribution of the products are suspected to be independent of energy usage. Using the total production variable removes this random variation.

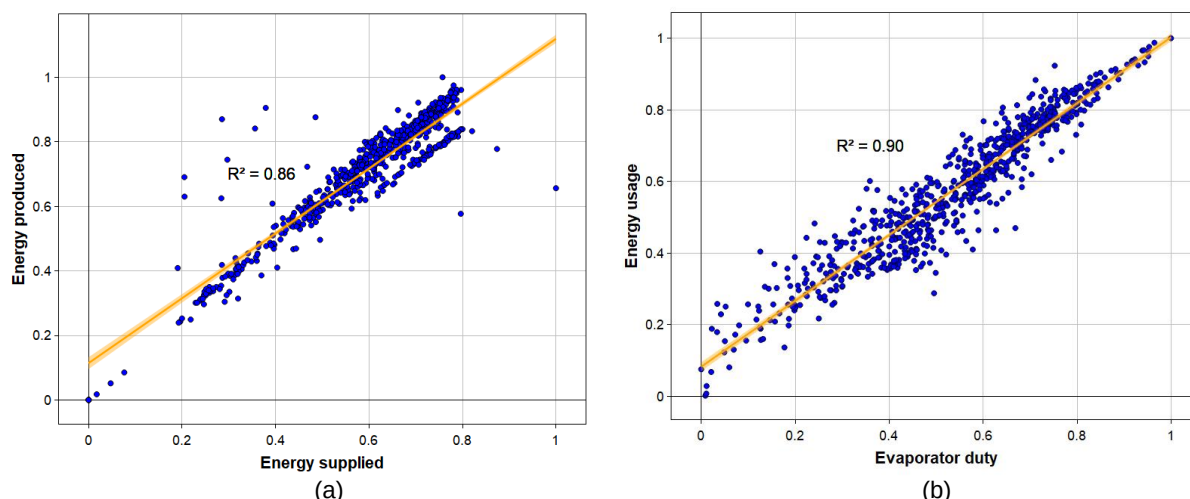


Figure 3.14: The relationship between (a) the energy produced and the energy supplied for case study II and (b) the energy usage and the evaporator duty for case study IV

For case study II, the total energy supplied to the engines (from the natural gas and the boiler feed water) and the total energy produced (in the form of electricity and steam) showed a strong linear relationship (shown in Figure 3.14(a)). Thus, the total energy into the system was used. For case study III the total production was used because it is the only variable on the output side of the process and should therefore include all the variation of the process.

For case study IV, a similar argument was used as for case study II but it is based on the strong linear relationship between the total evaporator duty (cooling delivered) and the energy supplied to the system (shown in Figure 3.14(b)). For case study V, the flow and pressure were multiplied, which relates to the kinetic energy of the compressed air. The different variables selected for each driver set of each case study are given in Table 3.7.

Table 3.7: Driver sets for each case study

Case study	Driver set	Variables
Case study I	A	One variable: Total production
	B	Five variables: Production for product A to E
	C	One variable: Production of product A
	D	One variable: 1 st PCA variable
Case study II	A	One variable: Total energy in
	B	Four variables: Natural gas normal volume flow; Natural gas heating value, BFW flow, BFW temperature
	C	One variable BFW volume flow
	D	One variable: 1 st and 2 nd PCA variables

Case study	Driver set	Variables
<i>Case study III</i>	A	One variable: Total ferrochrome production
	B	Four variables: Reductant mass; Production mass; Ore mass; Flux mass
	C	One variable: Ore mass
	D	One variable: 1 st PCA variable
<i>Case study IV</i>	A	One variable: Total evaporator duty
	B	Six variables: Evaporator volume flow; Evaporator TI; Evaporator TO; Condenser volume flow; Condenser TI; Condenser TO
	C	One variable: Evaporator volume flow
	D	Two variables: 1 st , 2 nd and 3 rd PCA variables
<i>Case study V</i>	A	One variable: Flow power ¹
	B	Three variables: Normal volume flow of air; Air pressure; Total production (tonnes hoisted)
	C	One variable: Total production (tonnes hoisted)
	D	One variable: 1 st PCA variable

¹ Flow multiplied by pressure (which relates to the kinetic energy)

3.5 Savings, errors and consistencies obtained

As explained in section 2.6, two approaches were used when training the models:

- (i) The standard M&V approach of using all the available data – the all-data approach
- (ii) The machine learning approach of determining the average when taking a random 70% sample for model construction (remaining 30% used for determination of model accuracy)

Each of these approaches was used to calculate savings. Further, two different methods were used to calculate the savings for each approach:

- (i) The traditional forward-direction calculation method
- (ii) The backward-direction calculation method

Thus, a total of four calculation methods were used to calculate the savings for each combination of the model types and selected driver sets, thus producing 24 calculation options per model type. The savings calculated for case study I using daily frequency is shown in Figure 3.15. The savings are shown on the y-axis, while the different driver sets are grouped according to model type on the x-axis. The differences are interpreted according to the guideline of Figure 2.18.

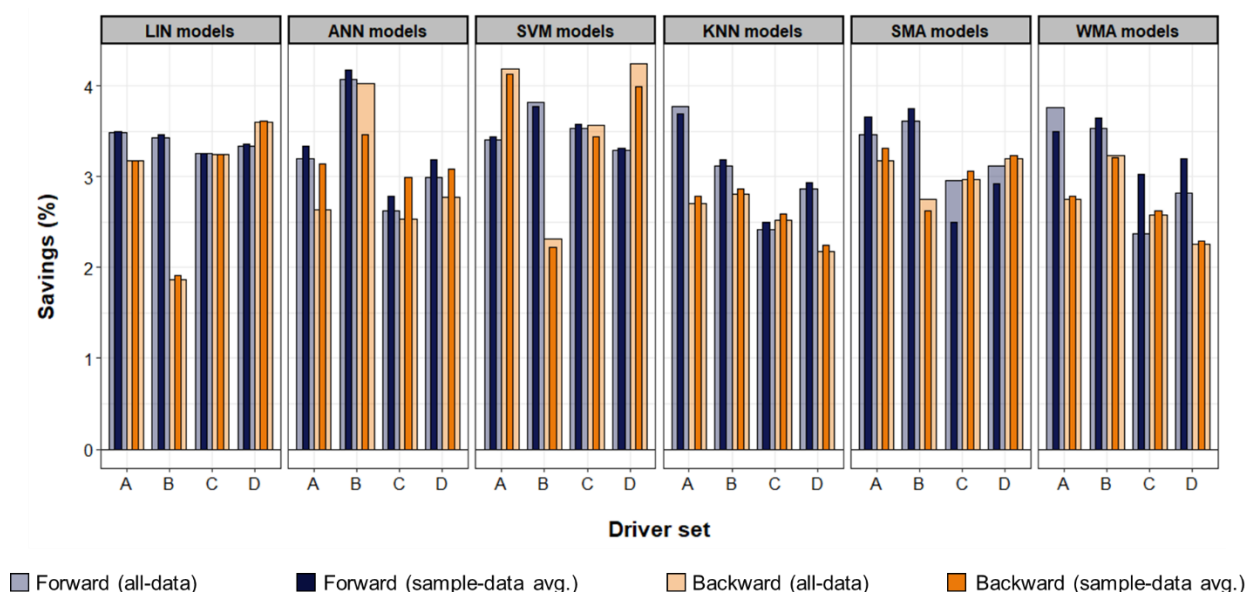


Figure 3.15: The savings calculated for case study I (daily)

Every calculation option shows a saving. However, the savings vary between 1.9% and 4.2%, with an overall average of 3.1%. Thus, the savings calculated can differ by up to 127% in comparison to the most conservative calculation, depending on the method used. For almost all the model types, the greatest savings occur when using driver set A or D (the total production or the PCA variable), either in the forward or the backward direction. The exception occurs for the ANN, where the greatest savings occurs when using the driver set B (all the measured variables) in the forward direction, which then transfers to the SMA model type. The smallest savings for each model type occurs when using driver set B or C, except for ANN model type, where it occurs when using driver set C in the backward direction which is again transferred to the SMA model type. Savings are just as much influenced by the type of driver set used than the model type chosen. The LIN model type does, however, show the most consistent savings between the various calculation options.

For most of the calculation options, the savings based on the all-data approach compare well with the sample-data savings across the different models, with differences of less than 10%. However, some calculation options did show higher differences. This includes the SMA and WMA model types that show 15.3% and 27.8% difference in the forward direction when using driver set C. The WMA model type further shows a difference of 13.3% when using driver set D in the forward direction. The ANN model type shows large differences in the backward direction for all of the driver sets (>10%).

There are thus differences between the all-data and sample-data approaches. In this case, the only non-ensemble model that stands out is the ANN model. As mentioned in Chapter 1, the ANN model can be prone to overfitting and thus it is important to use sample data as there is higher uncertainty (on a conceptual basis) when using all the data. The model-averaging techniques show larger differences because the differences between the two approaches are aggregated for each of the single model types (LIN, ANN, SVM and KNN).

The differences between the forward and backward-calculated savings vary from slight (approx. 0% for the LIN model with the all-data approach and driver set C) to great (45.6% for the LIN model with the all-data approach and driver set B). In particular, the LIN and SVM model types show differences of more than 39% for both the all-data and sample-data approaches when using the driver set B. Both these models are linear-type models and are thus greatly influenced by the randomness in the product distribution.

For the ANN, the greatest inconsistencies occur when using the all-data approach with driver set B and the sample-data approach with driver set B (17.6% and 16.9% respectively). For the KNN model type it occurs when using driver set A for both approaches (28.3% and 24.6%). This is due to the level of inter- and extrapolation being higher when using this driver set (further details on this are provided when discussing savings for the other frequencies).

Each of the model types shows some inconsistency between the forward and the backward calculations. These inconsistencies are transferred to the model-averaging techniques. The SMA shows the greatest difference when using driver set B (all the variables) because the differences for each of the single models are especially predominant for this driver set. This, in turn, is likely caused by the randomness of the product distribution. The greatest differences for the WMA on average for the different approaches occur when using driver set A because it tends to give preference to the KNN model. This is shown in Figure 3.16.

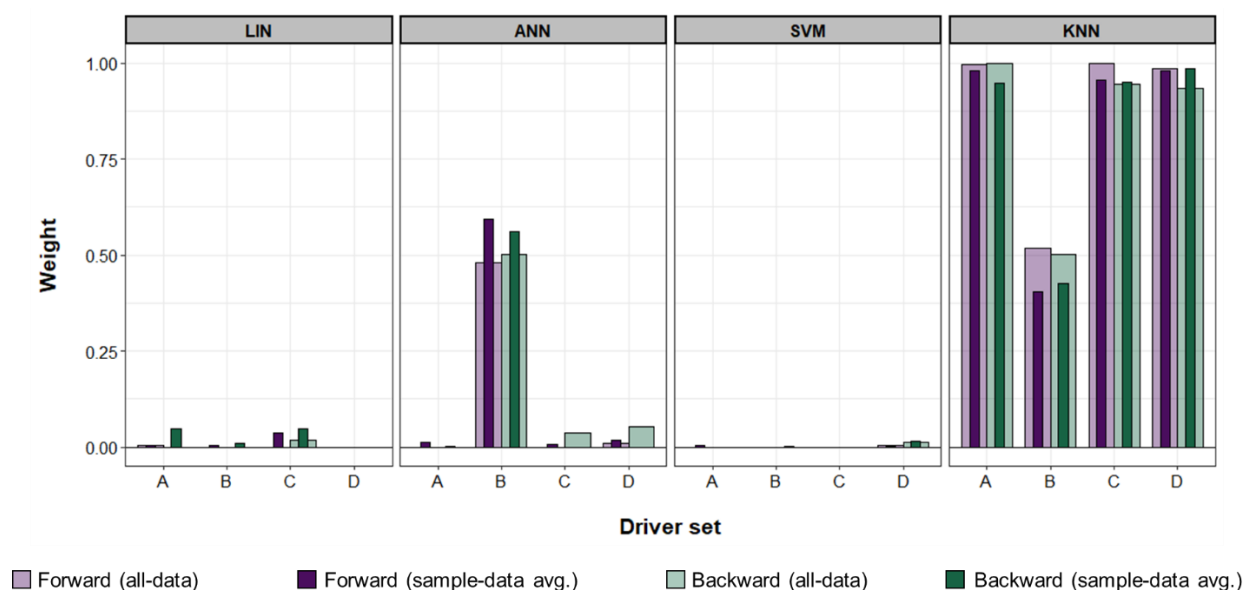


Figure 3.16: Weights for the WMA model for case study I (daily)

The figure shows the weights assigned by the WMA model to each model type for each driver set in the forward and backward direction when using the all-data approach. In each case the KNN model type has the greatest weights assigned to it and is thus favoured more. As shown in Appendix C4, the same was found for the other case studies. For the all-variable driver set the KNN model is favoured slightly less in comparison to the other driver sets while the ANN is favoured slightly more.

The savings differ when using different measurement frequencies. The bar plots for the savings for the weekly and monthly frequencies are shown in Figure 3.17 and Figure 3.18 respectively.

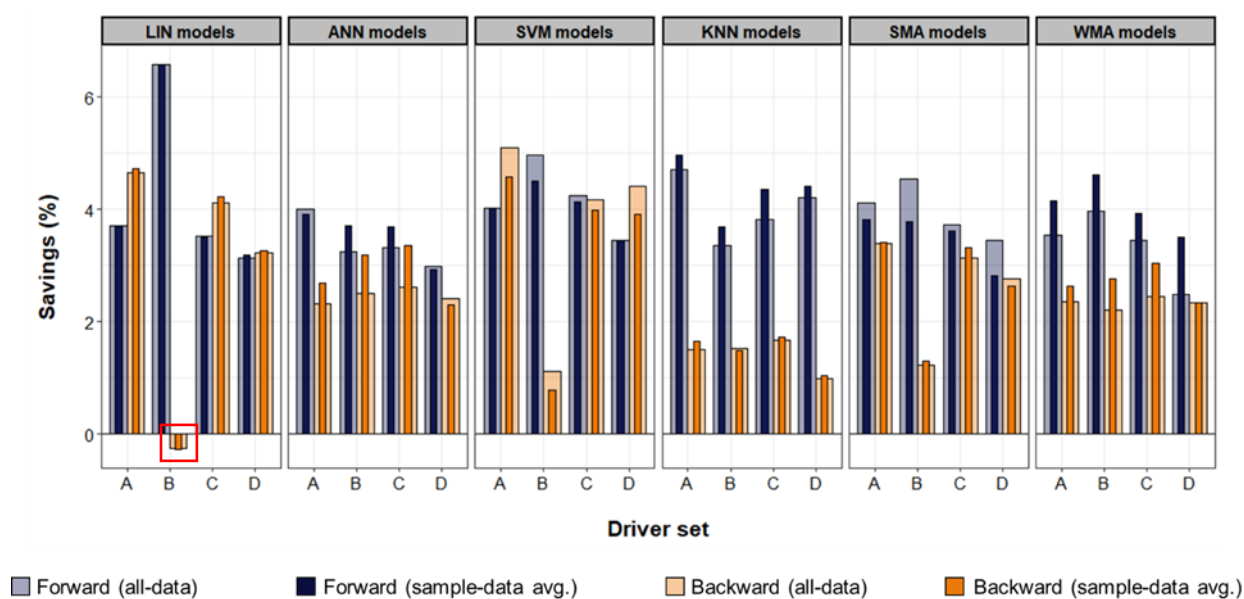


Figure 3.17: The savings calculated for case study I (weekly)

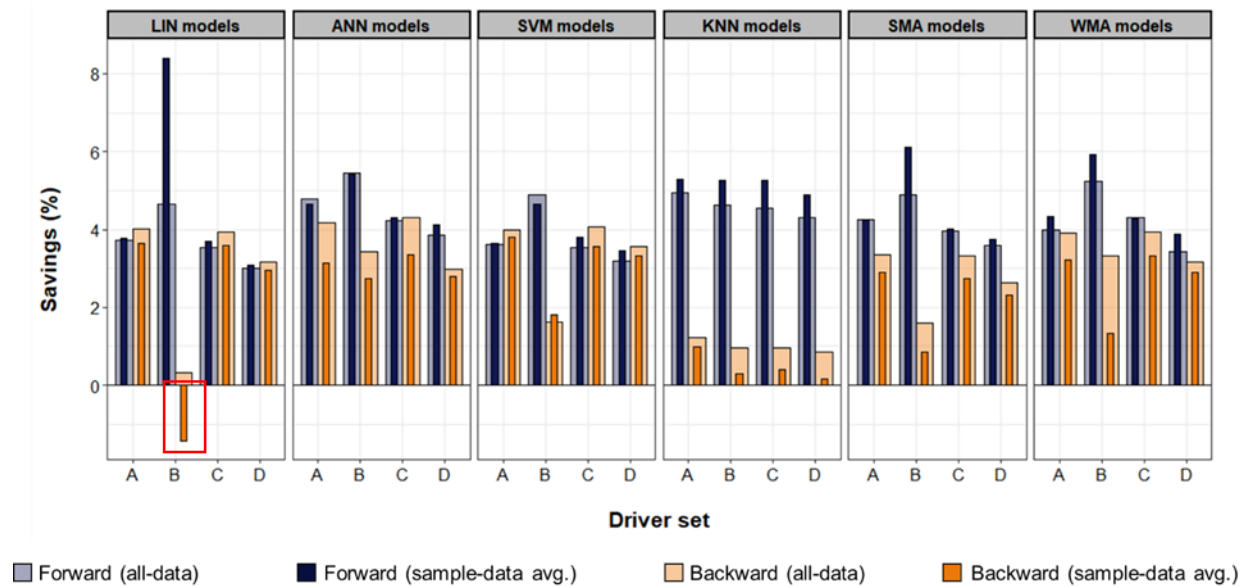


Figure 3.18: The savings calculated for case study I (monthly)

Both frequencies still show positive savings with each calculation but with three exceptions at the LIN model when using driver set B (highlighted in the figures). However, the differences between calculation options and between model types are more extreme as savings range from -0.3% to 6.6% and from -1.4% to 8.4% for the weekly and monthly frequencies respectively.

Generally, certain patterns tend to emerge across the different driver sets between the different model types at each of the measurement frequencies. For each type the greatest savings occur either when using driver set A or B. The greatest differences between the forward and backward-calculated savings also occur here. Each of the model types is more consistent between the two directions when using the other driver sets, except for the KNN model. The lowest savings tends to occur when using driver set D (the PCA variable), with similar savings for driver sets A and C. However, overall differences between different driver sets are even more prominent at higher measurement frequencies (weekly and monthly).

When comparing the daily and weekly frequencies, 36 of the 48 combinations (when accounting for the forward and backward direction separately) increased in the difference between the all-data and sample-data approach. Of these, 32 combinations showed even greater differences when using the monthly frequency. Overall the average difference between the all-data and the sample-data approach increased from 4.8% for the daily frequency, to 9.5% and 28.0% for the weekly and monthly frequencies respectively. Overall, the decrease in measurement frequency increases the difference between the two approaches.

The same observation can be made regarding the consistency between the forward and backward calculations. Of the 48 combinations, 38 increased in difference between the forward and backward direction from the daily to the weekly frequency. Overall, the average difference increased from 15.4% for the daily frequency to 35.7% and 38.3% for the weekly and monthly frequencies respectively.

Similar to the daily frequency, most of the model types show large differences (>15%) when using driver set B for both the weekly and monthly frequencies. The ANN model type is the sole exception when using the sample-data approach with this driver set at the weekly frequency and is considered a random occurrence because the ANN is likely suffering from overfitting. For both the weekly and monthly frequencies the KNN model shows poor consistency between the forward and backward directions for all the driver sets used. As mentioned before, this can be explained by considering the level of inter- and extrapolation. Figure 3.19 shows the density distributions of the total production for each measurement frequency in each year.

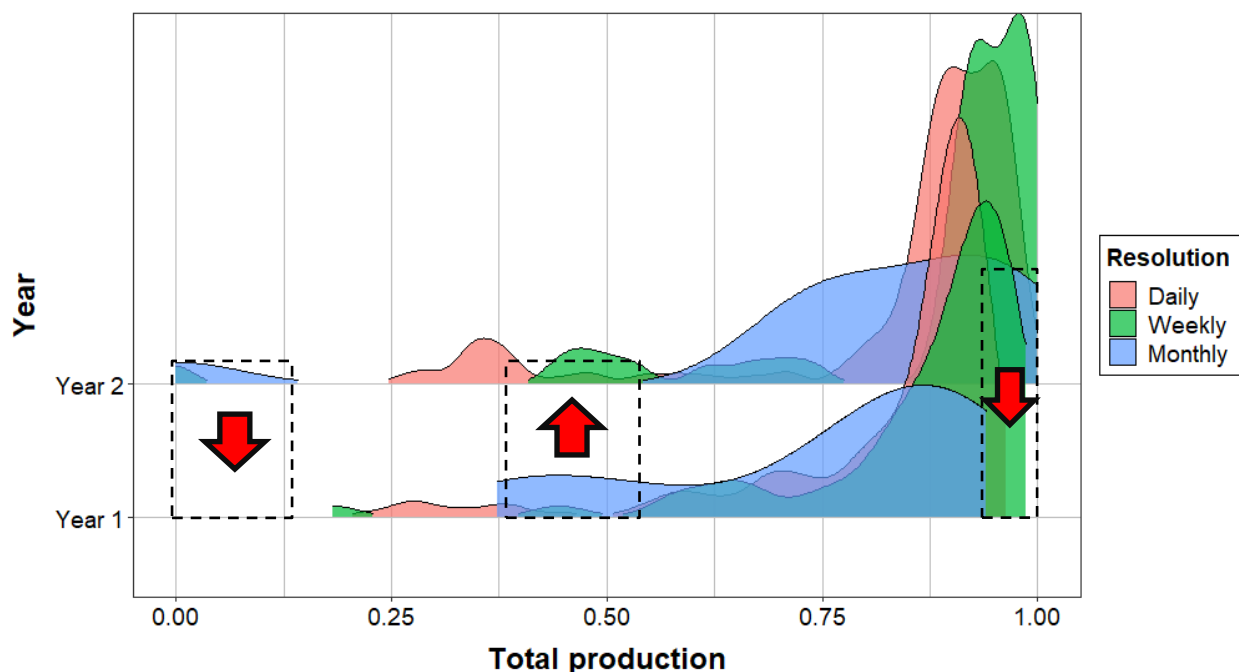


Figure 3.19: Data distribution for the total production for Year 1 and Year 2

The KNN model is greatly dependent on the distribution of data points throughout the range of use, which means it can be sensitive to inter- or extrapolation. For each frequency, there are portions where the ranges indicated by the density curves do not overlap between the two years. To illustrate, these ranges are indicated for the monthly frequency in Figure 3.19. Thus, there is a level of inter- and extrapolation in both the forward and backward directions. The extent of inter- and extrapolation is greater for the weekly and monthly frequencies than the daily frequency.

As explained in Chapter 2, the consistency of the model can also be determined by its sensitivity towards the sampling process (when using the sample-data approach). The savings density plot for the daily frequency is shown Figure 3.20 (also for case study I).

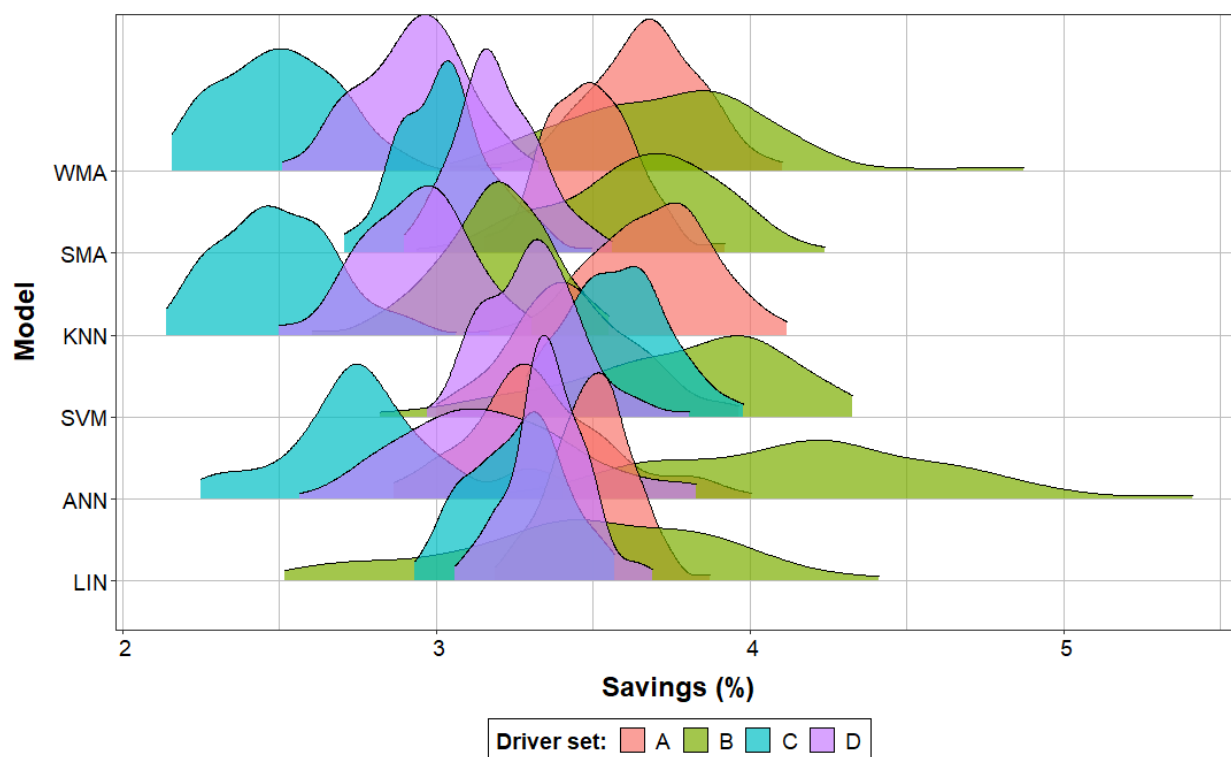


Figure 3.20: Savings density distribution for case study I (daily)

The plot shows the density curves for each driver set for each of the model types. The savings calculated range from 2.1% (minimum for KNN model type when using driver set C) to 5.4% (for the ANN model using driver set B). The savings agree well between the different model types when using the driver set A or D, while less so when using driver set B or C.

The LIN model shows the narrowest distributions for driver sets A, C and D. However, the narrowest distribution for the driver set B occurs when using the KNN model. The ANN model shows the widest distributions for each of the driver sets. Thus, while there is no model type that appears to be considerably less sensitive to the sampling process, the ANN model appears to be most sensitive to it. However, as with the average savings the widths of the distributions differ more between different driver sets. Overall, driver set B has the widest distributions. This agrees with the high level of inconsistency between the forward and backward-calculated savings when using this driver set.

The density curves of the CV(RMSE) for the daily and monthly frequencies are shown in Figure 3.21 and Figure 3.22, respectively. Note that the ranges of the x-axes are different for the two plots.

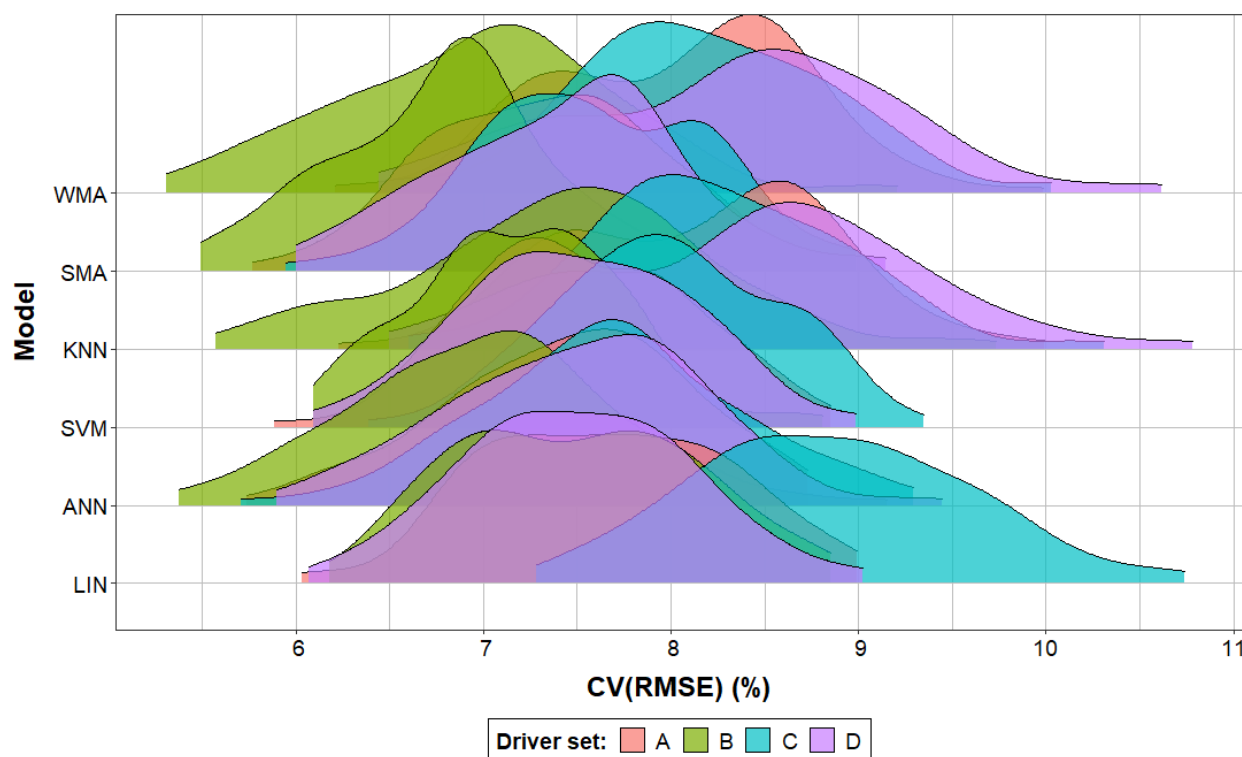


Figure 3.21: Density plot for the CV(RMSE) for the sensitivity repetitions for case study I (daily)

The CV(RMSEs) for the daily frequency range from 5.3% (for the WMA model driver set B) to 10.8%. (for the KNN model type when using driver set D). As with the savings, the average error values for the different model types agree for driver sets A and D, but also to some extent for the other two driver sets. Overall, no model type is clearly different from the others. Thus, there is an inherent average error associated with each driver set. Slight differences do occur, however, between the model types, most notably the LIN model in comparison to the others when using driver set C. Further discussion on the differences between the averages of model types follows later.

A distinction can be made between driver sets based on error. However, a more accurate driver set does not necessarily relate to greater consistency. Here, driver set B showed the lowest average CV(RMSE), but with noticeable inconsistencies in the savings. This is seen when considering the differences between the forward and backward calculations and the width of the sensitivity distributions.

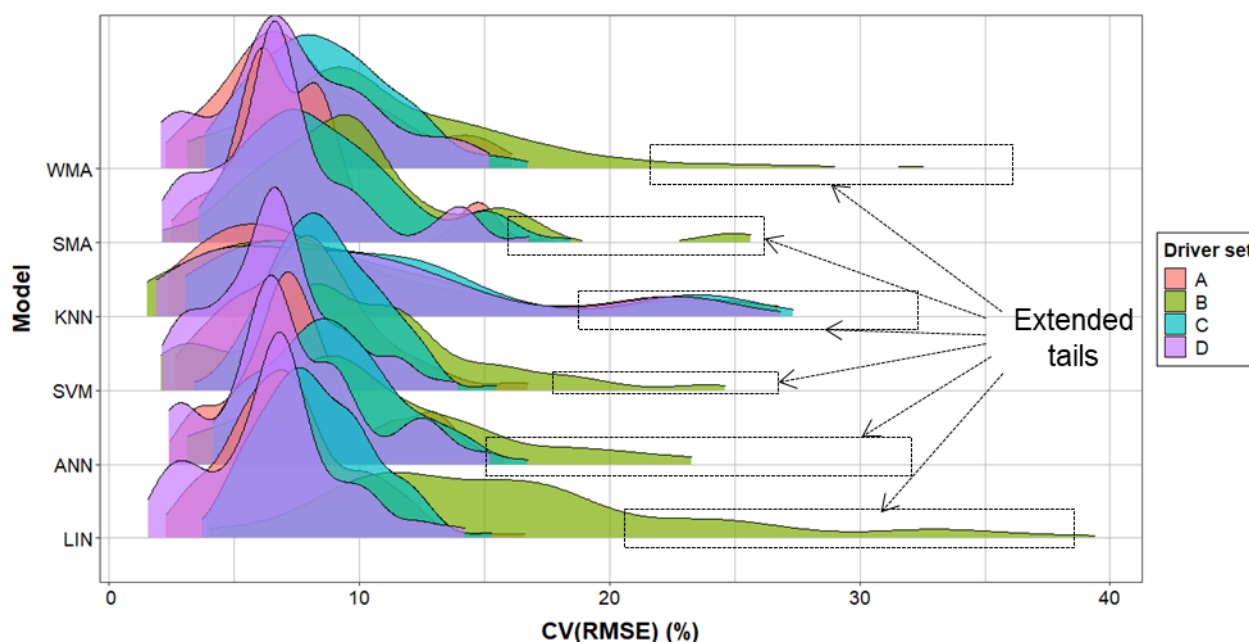


Figure 3.22: Density plot for the CV(RMSE) for the sensitivity repetitions for case study I (monthly)

The monthly frequency showed a much greater of CV(RMSEs) of 1.2% to 53.3% than the daily frequency. For most of the model types, the average errors for each driver sets increased from the daily to the monthly frequency. The exceptions are the LIN model type using driver sets A, C or D and the SVM model type when using driver set D.

The greatest increase in error for each model type occurred when using driver set B (all the measured variables). Each of the curves for driver set B shows extended tails to the right. These tails also occur when using the other driver sets for the KNN model. The predominant inconsistencies between the forward and backward calculations also occurred when using driver set B for the different model types and each driver set for the KNN model. The occurrence of these tails thus appears to be indicative of a high level of inconsistency.

The savings and CV(RMSE) for the different model types across the different frequencies show that a decrease in the measurement frequency increases the overall error and decreases the consistency for a specific model type. It also increases the differences between different model types and between different driver sets. Therefore, only the results for the daily frequency are discussed further, which was also the only measurement frequency used for the other case studies.

However, the general trend across all the measurement frequencies for case study I is that the order of savings calculated tend to agree across different models for the same driver set. The same trend was seen for other case studies, as shown in Appendix C3 and Appendix C5. Thus, in it appears that the driver set used has a greater influence than the model type on the final savings calculated.

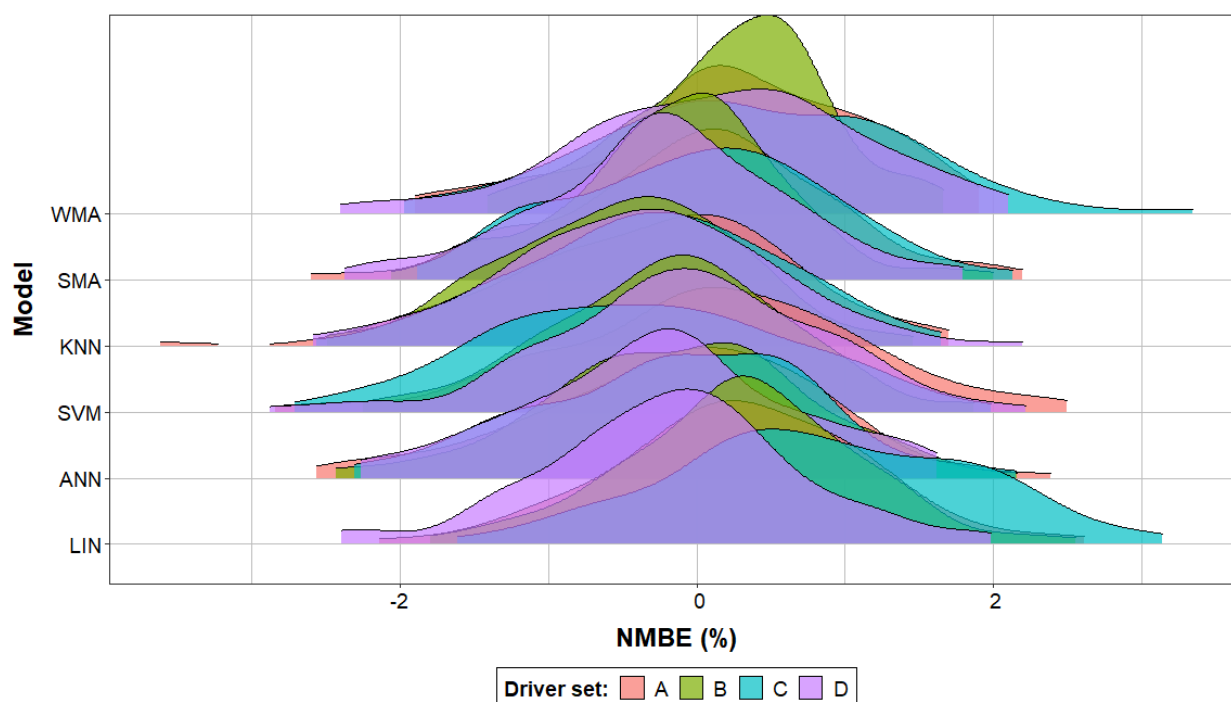


Figure 3.23: Density plot for the NMBE for the sensitivity repetitions for case study I (daily)

The NMBEs range from -3.6% (for the KNN model type when using driver set A) to 3.3% (for WMA model when using driver set C). Thus, on average the models show little bias so that the positive and negative errors tend to cancel out. The most notable exception to this, however, is the curve for the LIN model when using driver set C, which leans more to the right, thus having a positive bias. The WMA model shows the widest distributions for each of the driver sets, while curves for the other models tend to be similar besides the exception. As mentioned earlier, this is because the WMA model favours slightly different models with each iteration, thus causing greater variation.

As mentioned in Chapter 2, the NMBE and CV(RMSE) can be plotted against each other to determine whether a relationship exists between them and thus determine whether a greater CV(RMSE) leads to a greater NMBE. Consideration is first given to the all-data approach. The NMBE vs the CV(RMSE) for each model type using the different driver sets with this approach is given in Figure 3.24 (for the forward calculations).

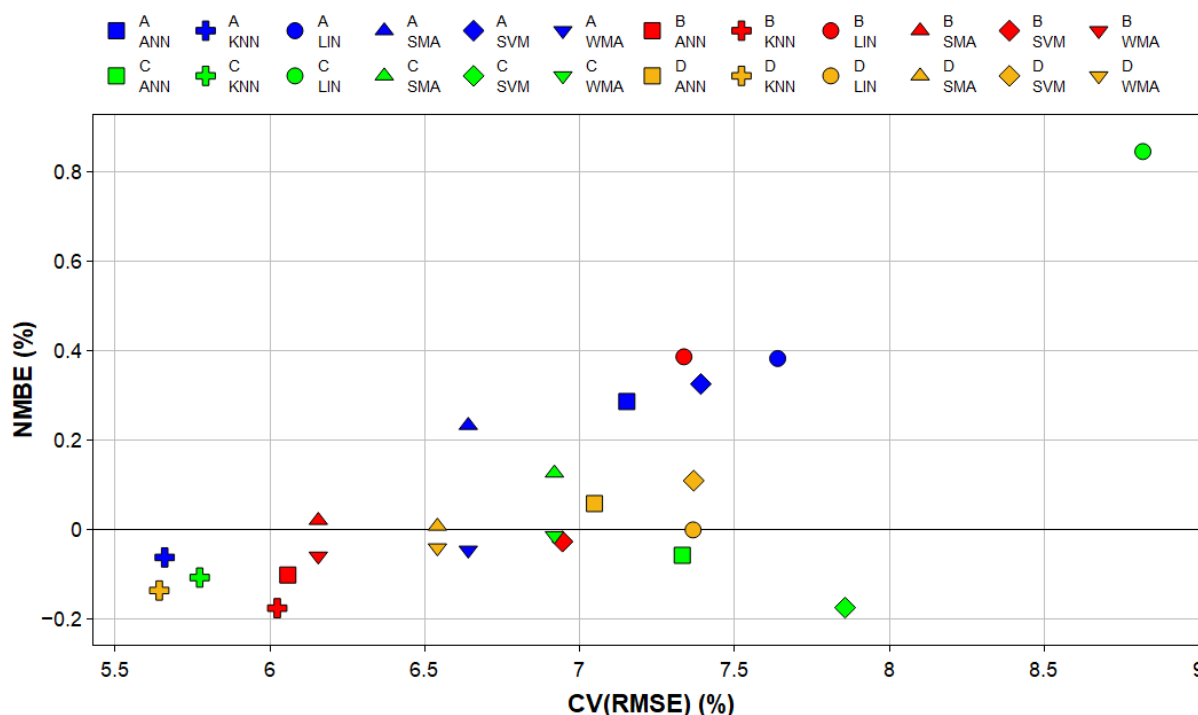


Figure 3.24: NMBE vs CV(RMSE) for the forward models when using all the training data

As the points are scattered randomly across both ranges; no definitive relationship is present. Based on CV(RMSE), the most accurate models for each driver set in order are (1) KNN (+), (2) SMA ($\blacktriangle$) and (3) WMA ($\blacktriangledown$). The least accurate model types are the ANN ($\blacksquare$), SVM ($\blacklozenge$) and LIN ($\bullet$) types in varying order. The LIN model using driver set C is the worst model based on both CV(RMSE) and NMBE. The NMBEs tend to vary for the different models, with slightly more options showing a negative bias. However, the KNN model type shows a negative bias for each driver set, which is transferred to the WMA model type as the WMA favours the KNN model more. As with the sample-data approach, grouping tends to occur according to driver set for the CV(RMSE), except for the KNN model type.

From the error values given in the plot, the KNN model type appears to be the most accurate. However, it should be noted this is for the daily frequency where a good distribution of points is available. As shown in Appendix C8, at the weekly and monthly frequencies the CV(RMSE)s for the KNN model type is amongst the worst. Further, the same is not observed when looking at the average errors for the sample-data approach, which is summarised using the same plot type in Figure 3.25.

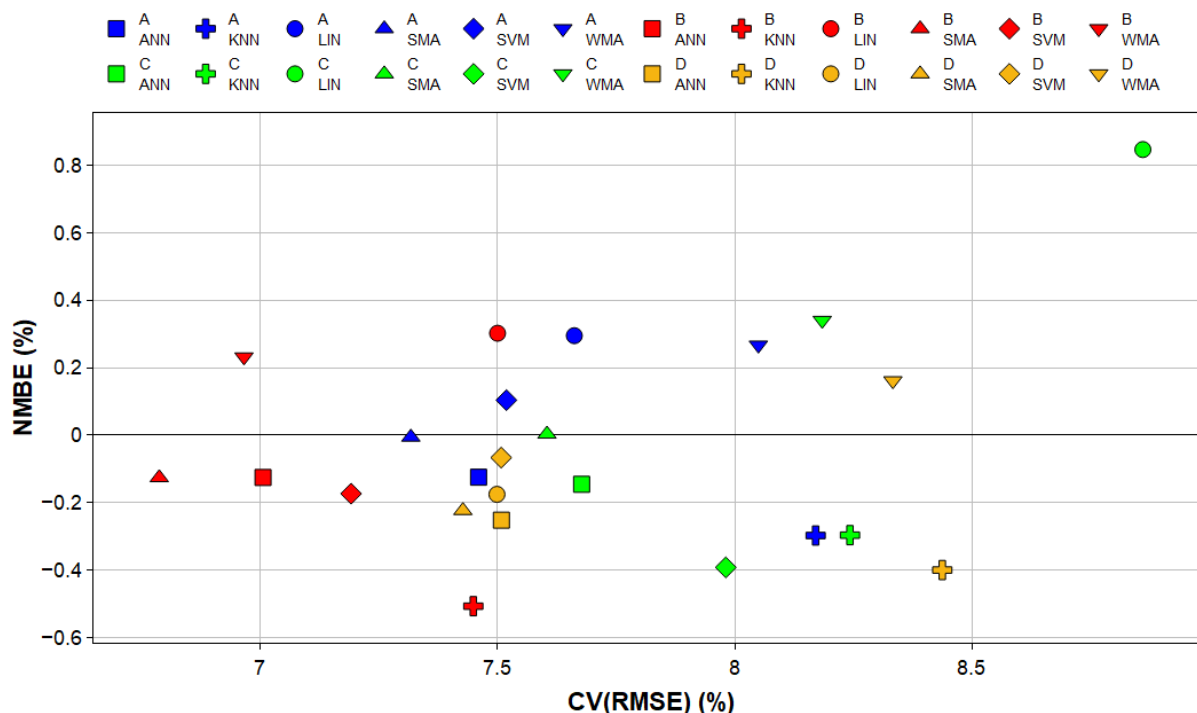


Figure 3.25: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages)

Here, the errors are even more strongly grouped according to the driver set used. The range of errors within each driver set is also reduced for the sample-data approach. For each driver set the SMA model type shows the lowest CV(RMSE) and thereafter the model types occur in varying order for the different driver sets. The LIN model using driver set C still shows the highest CV(RMSE) and NMBE. However, unlike with the all-data approach, the CV(RMSEs) for the KNN model type are amongst the highest for each driver set.

To explain this occurrence, reference is made to the working of the KNN model. As explained in Chapter 1, the basic working of the model entails taking the average of the nearest k points. Figure 3.26 contrasts the sample-data and the all-data approach for a single variable.

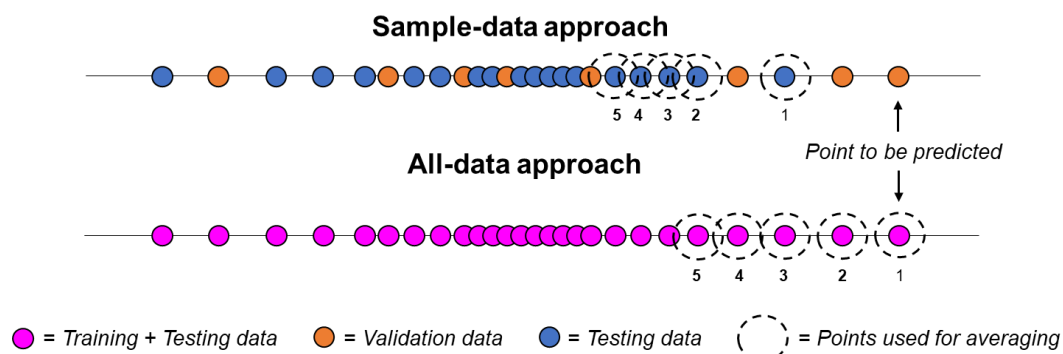


Figure 3.26: Difference between the all-data and sample-data approach for the KNN prediction

In each case the prediction is made for the last value. The five values used to calculate the average (for the prediction) are encircled. When using the sample-data approach, the distribution throughout the range of data points is not necessarily good. Further, the errors calculated are for the validation dataset (the orange points), which are those points that are not included in the training set. Therefore, in certain cases the values used to calculate the average can be far away from the points to be predicted.

For the all-data approach, the errors are calculated across all the data points so that the training dataset and the validation dataset are the same (i.e. the full baseline dataset). Thus, when making a prediction for a certain data point, one of the data points that will be included when calculating the average is the data point itself, which will help to reduce the error. Calculating the error in this way is thus inherently flawed for the KNN model type.

When using all the data for training, it can be considered as a single-sample dataset because it is one of many possibilities for data selection. As mentioned, the all-data approach results in greater variation between model errors. Thus, for a certain sample dataset, model types will differ although average differences are reduced by repetition significantly.

When applying a standard M&V process, for most cases it is valid to base the model choice on the errors calculated if a specific dataset was used to train the models. This is because each model has specific parameters that are considered as final. However, in this specific study (i.e. where model type is evaluated), basing the evaluation on the single-sample dataset is not valid because it does not provide for sampling uncertainty. This justifies the model-scoring performed based on the sample-data approach, of which the results are given in section 3.6.

The discussion of the results has thus far focused on case study I. However, for the remaining case studies the same findings were predominant – model types showed the same savings/errors for a given driver set with random deviations for certain models at certain instances. No pattern was observable for these deviations and no model type tended to deviate more times from the average than other models in most of the cases. Worth noting though is that all-variable driver set (driver set B) and the driver selected using the method of Gallagher *et al.* [14] (driver set C) tended to increase the standard deviation at either the savings or error density plots for each of the cases. All these results are given in Appendix C. An exception was found, however, in case study II. Figure 3.27 shows the density curves for the CV(RMSE) for case study II.

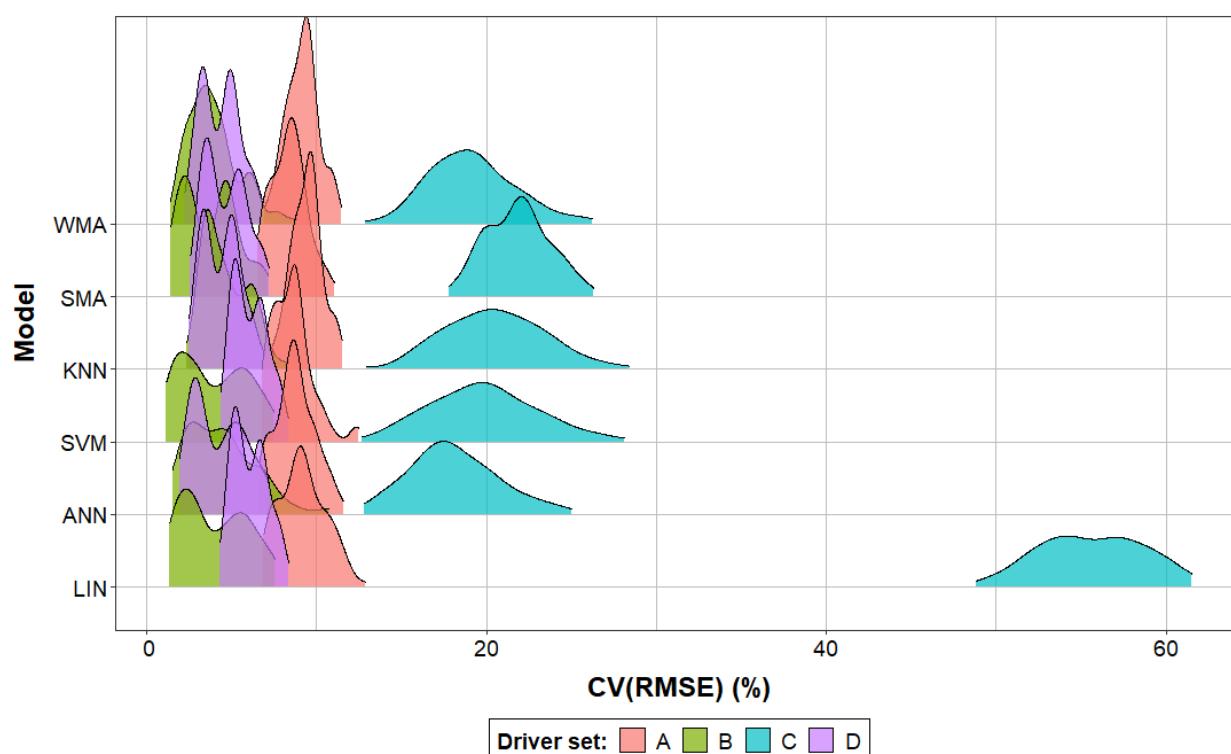


Figure 3.27: The density plot of the CV(RMSE) for case study II

Driver sets A and C have significantly larger errors than driver set B and D. In fact, on a conceptual level the BFW flow (driver set C) is a very poor predictor for the energy recovered from natural gas engines. However, the LIN model responded far worse to these driver sets than the other models. This is due to the non-linear relationship between the energy usage and these driver sets.

Most of the driver sets in the other case studies had a strong degree of linearity. It has been mentioned that SVM is also a type of linear model and it also performs worse for case study II than the other models when using driver set D. This highlights an important limitation of the linear model – the non-linear model types (ANN, KNN and to an extent SVM) can be used to model a linear process but the linear model cannot be used to model a non-linear process.

From the results presented so far, no model stood out clearly in terms of accuracy and consistency. The model-scoring technique was applied to each case study to identify potential patterns.

3.6 Results of model-scoring

The metric and category winners for each case study are given in Table 3.8 to Table 3.12, while the values which were used to do the scoring are given in Appendix C10. A summary of the category winners for each case study is presented in Table 3.13.

Table 3.8: Scoring table for case study I

Parameter	Driver set			
	A	B	C	D
Accuracy				
Error - CV(RMSE)	SMA	SMA	SMA	SMA
Error - NMBE	SMA	ANN	SMA	SVM
Overall	SMA			
Sensitivity to driver set used				
Savings std. dev. across driver sets	LIN			
Sensitivity to data sampling process				
Savings std. dev. for sensitivity repetitions	LIN	KNN	LIN	LIN
CV(RMSE) std. dev. for sensitivity repetitions	SVM	SVM	SVM	SVM
NMBE std. dev. for sensitivity repetitions	LIN	WMA	SMA	SMA
Overall	LIN			
Consistency test				
Consistency difference between forward and backward	ANN	KNN	LIN	ANN
Overall	ANN			

Table 3.9: Scoring table for case study II

Parameter	Driver set			
	A	B	C	D
Accuracy				
Error - CV(RMSE)	SMA	SMA	ANN	ANN
Error - NMBE	WMA	LIN	WMA	LIN
Overall	None			
Sensitivity to driver set used				
Savings std. dev. across driver sets	KNN			
Sensitivity to data sampling process				
Savings std. dev. for sensitivity repetitions	LIN	LIN	SMA	KNN
CV(RMSE) std. dev. for sensitivity repetitions	KNN	KNN	SMA	SVM
NMBE std. dev. for sensitivity repetitions	WMA	WMA	SVM	KNN
Overall	KNN			
Consistency test				
Consistency difference between forward and backward	LIN	SVM	SVM	WMA
Overall	SVM			

Table 3.10: Scoring table for case study III

Parameter	Driver set			
	A	B	C	D
Accuracy				
Error - CV(RMSE)	SMA	SMA	SMA	SMA
Error - NMBE	KNN	KNN	SVM	ANN
Overall	SMA			
Sensitivity to driver set used				
Savings std. dev. across driver sets	KNN			
Sensitivity to data sampling process				
Savings std. dev. for sensitivity repetitions	LIN	SMA	LIN	LIN
CV(RMSE) std. dev. for sensitivity repetitions	ANN	LIN	SMA	LIN
NMBE std. dev. for sensitivity repetitions	SMA	LIN	SMA	LIN
Overall	LIN			
Consistency test				
Consistency difference between forward and backward	SMA	SVM	SVM	SVM
Overall	SVM			

Table 3.11: Scoring table for case study IV

Parameter	Driver set			
	A	B	C	D
Accuracy				
Error - CV(RMSE)	ANN	SMA	ANN	SMA
Error - NMBE	WMA	LIN	LIN	LIN
Overall	LIN			
Sensitivity to driver set used				
Savings std. dev. across driver sets	KNN			
Sensitivity to data sampling process				
Savings std. dev. for sensitivity repetitions	LIN	SMA	LIN	SMA
CV(RMSE) std. dev. for sensitivity repetitions	LIN	SMA	LIN	ANN
NMBE std. dev. for sensitivity repetitions	SMA	SMA	LIN	SMA
Overall	SMA			
Consistency test				
Consistency difference between forward and backward	SMA	ANN	ANN	KNN
Overall	ANN			

Table 3.12: Scoring table for case study V

Parameter	Driver set			
	A	B	C	D
Accuracy				
Error - CV(RMSE)	SMA	SMA	SMA	SVM
Error - NMBE	KNN	ANN	LIN	WMA
Overall	SMA			
Sensitivity to driver set used				
Savings std. dev. across driver sets	KNN			
Sensitivity to data sampling process				
Savings std. dev. for sensitivity repetitions	SVM	SVM	LIN	SVM
CV(RMSE) std. dev. for sensitivity repetitions	LIN	WMA	LIN	KNN
NMBE std. dev. for sensitivity repetitions	SVM	SVM	WMA	WMA
Overall	SVM			
Consistency test				
Consistency difference between forward and backward	KNN	SMA	ANN	ANN
Overall	ANN			

All the model types appear at least once as a metric winner in each category in one of the case studies, except for the ANN, SMA and WMA model types in the sensitivity to driver set used category. Furthermore, where multiple metrics/driver sets are applicable, there is a good distribution of model types as metric winners in each category and with at least three model types in a category (except for the consistency test at case study II). For example, for case study I, the ANN, SMA and WMA model types appear in the accuracy category, while the LIN, SVM, KNN, SMA and WMA model types appear in category for the sensitivity to the data sampling process.

Table 3.13: Summary of category winners for each case study

Parameter	Driver set				
	CS I	CS II	CS III	CS IV	CS V
Accuracy	SMA	None	SMA	LIN	SMA
Sensitivity to driver set used	LIN	KNN	KNN	KNN	KNN
Sensitivity to the data sampling process	LIN	KNN	LIN	SMA	SVM
Consistency test	ANN	SVM	SVM	ANN	ANN

CS = Case study

In some of the case studies, a particular model type appear strong in multiple categories, but in no case study does a particular model type dominate all of the categories. Each model type appears at least once as a category winner, while SMA appeared the most times and dominated the accuracy category. This observation implies two things:

1. Model-averaging can be advantageous in the M&V process, but not necessarily WMA, which tends to favour a particular model type;
2. The fact that the simple average of the models tends to be (slightly) more accurate than the models themselves confirms that the models included in the averaging technique all compare well in terms of accuracy. However, a certain type of model may be more accurate in a particular part of the prediction range. A short discussion on this is given in Appendix C11.

The KNN model type is the winner most often in the second category. This can be understood on a conceptual level given that there often exists some relationship between different potential drivers. Case study IV is used to illustrate this argument.

Industrial chillers are typically operated in a predictable manner – a certain evaporator flow and inlet temperature will be matched with a certain condenser flow and it will produce certain outlet temperatures. In other words, specific conditions on the condenser side are matched by specific conditions on the evaporator side which relate to a certain evaporator duty. Thus, the five data points that are the closest to a data point to be predicted when using all the variables are likely to be the five data points that are the closest when using the calculated evaporator duty as predictor. Thus, the total energy usage predicted will be the same. The same argument can be applied to other processes. However, this is not the case for case study I, where the distribution between the different products are random and do not relate to the total production. Here, the LIN model type performed better, which can be expected given the strong degree of linearity between the different products and the energy usage.

The linear-type models (LIN and SVM) dominate the category of sensitivity to data sampled when looking solely at the category winners. Most of the case studies are relatively linear when considering the different driver sets (as shown in Appendix C1). These models are thus not dependent on having data throughout the applicable range. If the data points that are included during training fall close to the “true” linear line (or the “true” derived linear line for SVM), the line of best fit will be close for each iteration.

3.7 Implications of the results for M&V

The differences between models for the different values calculated (savings, errors, etc.) vary from slight to extreme. These differences are also not dependent on one another (i.e. a large difference in savings does not imply a large difference in error). The size of the differences between the models are dependent on:

1. The data sampled for the calculation
2. The degree of linearity or non-linearity of the process
3. The underlying data quality – the distribution between true variation in the data and noise
4. The driver set used

The driver set used can strongly influence data quality. From the case studies it appears that, in most cases, the driver set used influences the size of the savings more than the type of model used. It also appears that although driver sets can be distinguished based on error, a lower error does not necessarily relate to improved consistency. However, finding a driver set that allows for a good balance between accuracy and consistency can reduce the range of savings, as seen with driver set D for case study I. Distinguishing between model types appears to be more difficult than distinguishing between driver sets.

The LIN model cannot capture non-linearities and is also more sensitive to randomness in the data. The ANN and SVMs models can be slightly less accurate at times or suffer from a high level of inconsistency, while the KNN model type requires a good distribution of data points throughout the entire range of the driver set used. Thus, each model type has different drawbacks so that no one model stands out in terms of the performance metrics used.

This is verified by the findings of related studies. In section 1.6 it was mentioned that different models have been found to be superior in terms of accuracy for different cases presented in literature. From this it can be concluded that there is no clear superior model. This study shows that this finding extends to the field of M&V in terms of both accuracy and consistency.

Granderson *et al.* tested the accuracy of different models for commercial buildings by using one year for training and the subsequent year for testing [19]. They found that the different models yielded similar errors in each of the regions in the United States where they had been tested. However, not all the models used the same driver set. Thus, their study suggests that the geographical region (which influences various factors of the datasets) have a greater impact on error than the model type. The same was found in this study when using different driver sets.

They further varied the length of the training set to change the amount of data provided to the models. No one model performed better or worse across the different training periods.

Gallagher *et al.* used tested the accuracy of different model types for an industrial facility using a very similar approach to Granderson *et al.* [46]. They found different rankings for the different model types, based on CV(RMSE) for different measurement frequencies. The same was observed in this for the different case studies investigated. Gallagher *et al.* did, however, find that while for some cases a decrease in measurement frequency increased the error size, in other cases it decreased, which was also true (although less so) for this study.

In a follow-up study Gallagher *et al.* investigated a case study where an ESM was applied to the same facility [14]. They found that different models tended to give similar savings ranges when using CV(RMSE) to calculate uncertainty, which corroborates the results of this study.

Although only five case studies were considered in this study, the “randomness” that they exhibited in model type behaviour is enough to confirm the existence thereof within the industrial sector. Using the more complex models does not guarantee improved accuracy over a linear model but rather appears to be quite similar in many cases. These models also require more expertise to set up than a linear model. The question thus arises of whether it is beneficial to consider these models at all.

In the case studies presented here, if one model showed a saving for a particular driver set, the other models also showed a saving most of the time. Where a large difference did occur, these models helped to highlight non-linearity. These models can thus be used as a check for the simple linear model and be of significance not necessarily when it shows a similar saving but when it clearly shows a difference. Further, because model-averaging tends to produce lower error values, the use of these additional models can help to improve confidence in the results obtained by reducing uncertainty. Different models may also be more applicable in different cases due to the different advantages and disadvantages shown by each.

It is noted that especially the machine learning models perform better when using higher measurement frequencies, which might limit its performance in certain cases where higher frequencies are not available. However, M&V 2.0 has been enabled specifically due to higher frequencies becoming increasingly more common. Further, the LIN model type was also adversely affected in consistency by lower measurement frequencies. Thus, overall higher measurement frequencies appear to be more suitable.

Although model development from a single dataset may show large differences in errors, introducing sample iteration reduces the difference. The principle of iteration from the field of machine learning has not been adopted in M&V yet. However, it can help to remove the effect of sampling uncertainty, as a sample dataset must be used to identify possible overfitting.

The given approach requires additional expertise from the M&V practitioner, which will likely increase cost. The availability of low-cost, user-friendly software may help to mitigate this. However, such software must be developed and provide for the fundamental principles of M&V. Lastly, different groups in the M&V sector will have to decide whether it agrees with commonly accepted machine learning approaches, which do not appear to have been formally applied within M&V.

3.8 Conclusion of results

The chapter demonstrated the application of each of the steps given in the evaluation procedure developed in Chapter 2 to five industrial case studies. The main findings of the application of the procedure were discussed. Focus was placed on case study I as most of the case studies yielded similar findings.

The results show that for a particular case study and driver set, certain model types may yield similar savings while the savings calculated using other model types may differ significantly. However, there is no pattern that emerges regarding the similarity or differences between the result of different model types while the driver set used tend to influence the final results more than the model type. There is also no clear superior model type in terms of accuracy and consistency although different model types may tend to perform slightly better in different categories. This leaves a question to the M&V industry that needs to way the additional complexity of the machine learning models against the possible benefit that it can supply through verification of simpler models and the accurate modelling of non-linear systems.

CHAPTER 4 CONCLUSION

4.1 Summary of study

The application of the M&V process fulfils an important role in many industries. Through the correct and rigorous application thereof, savings achieved by ESMs can be quantified accurately. Savings quantified can influence stakeholders' decisions regarding the provision of capital for similar projects. In other cases, it can determine the remuneration of ESCOs or the monetary value of government incentives.

An important consideration during the M&V process is the construction of a model to predict the energy usage. The model influences core aspects of the M&V process, mainly accuracy and consistency. Work done towards the use of different types of SRMs for M&V applications in the industrial sector, a key contributor towards energy usage, has been limited.

This study evaluated the influence of different types of SRMs on accuracy and consistency during the calculation of energy savings within the industrial sector. This was done by developing a comprehensive evaluation procedure to evaluate the accuracy and consistency of different model types. The potential benefit of using ensemble models, specifically model-averaging, was also investigated by applying different model types and model-averaging techniques to five industrial case studies. The use of multiple case studies provided a holistic view of the performance of different model types.

Based on the findings of previous studies, four model types were selected in Chapter 1. This was the simple linear model (LIN) type and three machine learning models – ANNs, SVMs and KNNs. Based on these models, SMA and WMA could also be applied.

In order to evaluate these models, an evaluation procedure was developed in Chapter 2. The procedure included several steps to prepare for model development and evaluation. These steps were based on similar procedures found in literature and included common machine learning practices. The model development or training step made provision for the standard M&V approach of using all available data for model training and for the machine learning approach of using sample data.

Model accuracy was determined through the calculation of the CV(RMSE) and NMBE metrics. To evaluate model consistency, the sensitivity of the models towards the data sampling process and the drivers used for predication were determined. A third consistency test was also introduced in

this study based on the comparison between savings calculated in the reporting period and the de-savings calculated in the baseline period. Model-scoring was done within four categories, namely model accuracy and the three aspects of model consistency.

The results given in Chapter 3 for case study I showed that all the models behaved and performed similarly for a specific driver set. The models predicted similar savings for a driver set when looking at the daily measurement frequency. Generally, each model type yielded high consistency between the all-data and the sample-data approach for the savings calculated. In many cases, good consistency was also found between the prediction of the savings and the de-savings. Although certain types were prone to exceptions in a certain category, the exceptions did not all occur for the same model types. Overall, reducing the measurement frequency increased the number of exceptions and error sizes while decreasing the model consistencies. Therefore, the daily frequency, which was the highest measurement frequency available, was the only frequency used for the other case studies.

Concerning the model error, there were more significant differences between model types when using the all-data approach in comparison to the sample-data approach. It also gave a more consistent ranking of models between different driver sets. In contrast, the error values grouped more according to different driver sets when using the sample-data approach. For none of the approaches was there a definite relationship between the CV(RMSE) and the NMBE.

The various case studies highlighted different drawbacks for some of the model types. This includes the sensitivity of the KNN model to the level inter- and extrapolation, the sensitivity of the LIN model type to the degree of linearity of the data and the susceptibility of the ANN model type to overfitting.

Overall, differences in accuracy and consistency were not so much between different model types. The different driver sets, however, have a much bigger impact on savings and accuracy. A distinction could be made between different driver sets based on the absolute error (CV(RMSE)) but not based on the bias (NMBE). However, a lower absolute error did not necessarily yield improved consistency.

The results showed that a more accurate model is not necessarily more consistent and vice versa. A certain model type is not always more accurate or more consistent than its counterparts, although, model-scoring did highlight that certain model types may tend to perform slightly better within certain categories. This can be explained through the consideration of the different attributes of the respective model types.

SMA tended to be more accurate. This confirms the results from previous studies that showed that model-averaging increased accuracy because different model types have better accuracy in different regions of the prediction range. Model-averaging provides a way to capture the improved accuracy from different models. The KNN model type tends to be less sensitive to the driver set used as industrial facilities are often operated in a predictable manner. This forms localised relationships between different sets of variables. Further, the linear-type models, LIN and SVM, tend to be less sensitive to the data sampling process because industrial processes are often linear and any data points sampled along the standard linear trend will capture that trend.

Different model types may thus be beneficial for different cases, as demonstrated by the case studies used for this research. However, overall the results suggest that the driver set selected tends to influence the final uncertainty and the final savings calculated more so than the type of model selected given that most processes in industry have a strong degree of linearity.

Overall, this study indicated that different SRMs can be used to calculate the savings for M&V applications in the industrial sector by providing several practical examples. It demonstrated a comprehensive evaluation technique by considering applicable fundamental principles of M&V. The study also highlighted potential practical implications of employing the more complex machine learning models within the M&V industry as far as data selection and the possible need for repetitions are concerned.

The study showed that different energy drivers play a more prominent role than the model type used. There is therefore not a single, superior model type, but the different types tend to predict similar savings and yield similar accuracies for a certain driver set. However, certain types may differ significantly in certain cases because of drawbacks of each model type and due to the general imprecise nature of M&V. Nevertheless, different model types can be used to confirm the savings predicted by others and provide improved accuracy through model-averaging. Together with the machine learning approach of repetition based on sample data, the use of multiple model types can also help to reduce uncertainty and supply assurance to stakeholders.

4.2 Limitations and recommendations

This study aimed to evaluate the most common and promising model types although many more model types and ensemble techniques exist. Also, only a limited set of different hyperparameter options were considered. Therefore, further research can focus on expanding the list of model types investigated and hyperparameters considered to determine whether they yield additional findings.

While different driver set selection methods were included in this study, focus was not placed on the best selection method but rather on how a variation in driver sets influences different model types. However, the results obtained warrant further research towards the selection of driver sets. This can include exploration of other selection methods not included in this study and the impact of different methods on uncertainty. Consideration should also be given to the practical implications of different selection methods.

The repetitive approach given in this study has been used to provide for the newer machine learning models. However, the approach introduces potential additional complexities into the M&V process, such as the exact quantification of uncertainty and the implementation thereof within the standard M&V process.

Although this study evaluated different model types for application of M&V in the industrial sector, it has not considered the exact quantification of the uncertainties according to statistical principles. Rather, consideration was given to accuracy, which is only a relative indication of uncertainty. The exact quantification of uncertainty may be required by some stakeholders in certain applications of the M&V process. If uncertainty can be reduced through the inclusion of different model types and the process of repetition, stakeholders might be more easily convinced of the benefits of a project.

It is recommended that future studies look towards the exact quantification of uncertainty, based on this new approach. To this end it might be necessary to consider the Bayesian approach. Further, it may be useful to make provision for the inclusion of multiple models so that the uncertainty is based on the saving ranges given by these models and the density of repetitions around the final savings quantified. However, even if the approach given does reduce uncertainty, its inclusion into the standard M&V process adds complexity.

This study has also not considered potential techno-economical barriers prohibiting the implementation of the given methodology within the accepted M&V process. Examples of these may be cost, technical expertise and the need for a paradigm shift. Determining what these barriers are and how they can be overcome, will contribute towards a methodology for its practical implementation within the M&V industry and improve the evaluation and implementation of ESMs.

REFERENCES

- [1] Y. Iftikhar, W. He, and Z. Wang, "Energy and CO₂ emissions efficiency of major economies: a non-parametric analysis," *J. Clean. Prod.*, vol. 139, pp. 779–787, 2016.
- [2] M. S. Almozaini, "The causality relationship between economic growth and energy consumption in the world's top energy consumers," *Int. J. Energy Econ. Policy*, vol. 9, no. 4, pp. 40–53, 2019.
- [3] L. Chiari and A. Zecca, "Constraints of fossil fuels depletion on global warming projections," *Energy Policy*, vol. 39, no. 9, pp. 5026–5034, 2011.
- [4] W. P. Nel and C. J. Cooper, "Implications of fossil fuel constraints on economic growth and global warming," *Energy Policy*, vol. 37, no. 1, pp. 166–180, 2009.
- [5] "BP Statistical Review of World Energy Statistical Review of World," *Ed. BP Stat. Rev. World Energy*, 2019.
- [6] F. van der Ploeg, "The safe carbon budget," *Clim. Change*, vol. 147, no. 1–2, pp. 47–59, 2018.
- [7] Intergovernmental Panel on Climate Change, *Climate Change 2014: Impacts, Adaptations, and Vulnerability. Part A: Global and Sectoral Aspects*. UK: Cambridge University Press, 2014.
- [8] Department of Environmental Affairs (South Africa), "Climate Change Bill, 2018," *Government Gazette*, 40428:4–39, 2018. 8 Jun.
- [9] A. Trianni, E. Cagno, and A. De Donatis, "A framework to characterize energy efficiency measures," *Appl. Energy*, vol. 118, pp. 207–220, 2014.
- [10] J. W. Bleyl *et al.*, "Office building deep energy retrofit: life cycle cost benefit analyses using cash flow analysis and multiple benefits on project level," *Energy Effic.*, vol. 12, no. 1, pp. 261–279, 2019.
- [11] G. Zhang, P. Zhang, J. Xiu, and J. Chai, "Are energy conservation and emission reduction policy measures effective for industrial structure restructuring and upgrading?," *Chinese J. Popul. Resour. Environ.*, vol. 16, no. 1, pp. 12–27, 2018.

- [12] W. Hamer, “A practical approach to quantify RSA Section 12L EE tax incentives for large industry,” NWU (North-West University), 2016.
- [13] Y. Zhang, Z. O’Neill, B. Dong, and G. Augenbroe, “Comparisons of inverse modeling approaches for predicting building energy performance,” *Build. Environ.*, vol. 86, pp. 177–190, 2015.
- [14] C. V. Gallagher, K. Leahy, P. O’Donovan, K. Bruton, and D. T. J. O’Sullivan, “Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0,” *Energy Build.*, vol. 167, pp. 8–22, 2018.
- [15] Y. He, N. Liao, J. Bi, and L. Guo, “Investment decision-making optimization of energy efficiency retrofit measures in multiple buildings under financing budgetary restraint,” *J. Clean. Prod.*, vol. 215, pp. 1078–1094, 2019.
- [16] O. Gamaunova and T. Musorina, “Return on investment in the implementation of energy-efficient solutions for facade insulation,” *MATEC Web Conf.*, vol. 251, 2018.
- [17] J. Grandersona, S. Touzani, S. Fernandes, and C. Taylor, “Application of automated measurement and verification to utility energy efficiency program data,” *Energy Build.*, vol. 142, pp. 191–199, 2017.
- [18] M. Pye and A. McKane, “Making a stronger case for industrial energy efficiency by quantifying non-energy benefits,” *Resour. Conserv. Recycl.*, vol. 28, no. 3–4, pp. 171–183, 2000.
- [19] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes, “Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings,” *Appl. Energy*, vol. 173, pp. 296–308, 2016.
- [20] W. Booyesen, “Measurement and verification of industrial DSM projects,” North-West University, 2014. (Thesis – Phd).
- [21] A. Gonzalez-Caceres, M. Rabani, and P. A. Wegertseder Martínez, “A systematic review of retrofitting tools for residential buildings,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 294, p. 012035, 2019.

- [22] Y. Heo and V. M. Zavala, "Gaussian process modeling for measurement and verification of building energy savings," *Energy Build.*, vol. 53, pp. 7–18, 2012.
- [23] S. Meyers and S. Kromer, "Measurement and verification strategies for energy savings certificates: Meeting the challenges of an uncertain world," *Energy Effic.*, vol. 1, no. 4, pp. 313–321, 2008.
- [24] ISO (International Organisation for Standardization), "ISO 50001: Energy management systems - requirements with guidance for use," Chemin de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland, 2011.
- [25] J. Granderson and S. Fernandes, "The state of advanced measurement and verification technology and industry application," *Electr. J.*, vol. 30, no. 8, pp. 8–16, 2017.
- [26] K. Bhattacharjee, "Measurement and verification of industrial energy conservation projects," *Strateg. Plan. Energy Environ.*, vol. 33, no. 1, pp. 66–76, 2013.
- [27] Efficiency Valuation Organization, "International performance measurement and verification protocol - core concepts," 1629 K Street NW, Suite 300, Washington, DC 20006, USA, June, 2014.
- [28] ASHRAE, "Guideline 14-2014, Measurement of Energy, Demand and Water Savings." American Society of Heating, Refrigeration and Air-Conditioning Engineers, 1791 Tullie Cir NE, Atlanta, GA 30329, USA, 2014.
- [29] ISO (International Organization for Standardization), "ISO 50015: Energy management systems - measurement and verification of energy performance of organisations - general principles and guidance," Chemin de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland, 2014.
- [30] SABS, "SANS 50010: Measurement and verification of energy savings," 1 Dr Lategan Rd, Groenkloof, Pretoria 0027, 2018.
- [31] W. Hamer, W. Booysen, and E. H. Mathews, "A practical approach to manage uncertainty in the measurement and verification of energy savings," *South African J. Ind. Eng.*, pp. 1–3.
- [32] SABS, "SANS 50010: Measurement and verification of energy savings," 1 Dr Lategan Rd, Groenkloof, Pretoria 0027, 2017.

- [33] A. Boyd *et al.*, “South African approaches to measuring, reporting and verifying: A scoping report.” p. 38, 2011.
- [34] US Department of Energy Federal Energy Management Program, “Measurement and verification for contracts,” 2015.
- [35] T. A. Reddy and D. E. Claridge, “Uncertainty of ‘measured’ energy savings from statistical baseline models,” *HVAC&R Res.*, vol. 6, no. 1, pp. 3–20, 2000.
- [36] K. Campbell, W. Booyesen, and J. C. Vosloo, “Evaluating the feasibility of the 12L tax incentive for energy intensive industries,” *South African J. Ind. Eng.*, vol. 28, no. 3, pp. 15–28, 2017.
- [37] H. Carstens, X. Xia, and S. Yadavalli, “Bayesian energy measurement and verification analysis,” *Energies*, vol. 11, no. 2, p. 380, 2018.
- [38] H. Carstens, X. Xia, and S. Yadavalli, “Measurement uncertainty in energy monitoring: Present state of the art,” *Renew. Sustain. Energy Rev.*, vol. 82, pp. 2791–2805, 2018.
- [39] SABS, “SANS 50010: Measurement and verification of energy savings,” 1 Dr Lategan Rd, Groenkloof, Pretoria 0027, 2012.
- [40] H. Janse van Rensburg, W. Booyesen, and S. W. Van Heerden, “Structuring data for RSA section 12L energy efficiency tax incentive application,” *SAIIE27 Proc.*, no. October, pp. 1–14, 2016.
- [41] P. A. De Villiers and L. J. Grobler, “Measurement and verification methodology for an energy efficiency improvement on the cooling auxiliary equipment in the mining industry,” *Proc. 9th Conf. Ind. Commer. Use Energy, ICUE 2012*, pp. 197–202, 2012.
- [42] S. Mirianhosseinabadi, “A framework for real-time performance measurement and verification and commissioning using building automation systems in existing buildings,” North Carolina A&T State University, 2016. (Thesis – Phd).
- [43] J. Lalk, “Addressing energy problems in South Africa,” *Featur. Innov.*, vol. 9, no. 2014, pp. 2–5, 2014.
- [44] M. Culka, “Applying Bayesian model averaging for uncertainty estimation of input data in energy modelling,” *Energy. Sustain. Soc.*, vol. 4, no. 1, pp. 1–17, 2014.

- [45] A. Srivastav, A. Tewari, and B. Dong, "Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models," *Energy Build.*, vol. 65, pp. 438–447, 2013.
- [46] C. V. Gallagher, K. Bruton, K. Leahy, and D. T. J. O'Sullivan, "The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings," *Energy Build.*, vol. 158, pp. 647–655, 2018.
- [47] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Build.*, vol. 49, pp. 560–567, 2012.
- [48] 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.
- [49] S. Robert, L. Ste, and A. Jay, "State of the art in building modelling and energy performances prediction : A review," vol. 23, pp. 272–288, 2013.
- [50] Z. Wang and R. S. Srinivasan, "A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models," *Renew. Sustain. Energy Rev.*, vol. 75, no. November 2016, pp. 796–808, 2017.
- [51] A. S. Jihad and M. Tahiri, "Forecasting the heating and cooling load of residential buildings by using a learning algorithm 'gradient descent', Morocco," *Case Stud. Therm. Eng.*, vol. 12, no. September 2017, pp. 85–93, 2018.
- [52] Y. Wei *et al.*, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 82, no. September 2017, pp. 1027–1047, 2018.
- [53] N. Fumo, "A review on the basics of building energy estimation," *Renew. Sustain. Energy Rev.*, vol. 31, pp. 53–60, 2014.
- [54] E. J. Ju, J. H. Lee, S. H. Park, C. S. Park, and M. S. Yeo, "Comparison of grey-box model and artificial neural network - Prediction of surface condensation in residential space," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 609, no. 3, 2019.

- [55] V. Ismet Ugursal, A. S. Fung, and M. Aydinalp, "Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks," *Appl. Energy*, vol. 71, no. 2, pp. 87–110, 2002.
- [56] M. Aydinalp, V. I. Ugursal, and A. S. Fung, "Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks," *Appl. Energy*, vol. 79, no. 2, pp. 159–178, 2004.
- [57] S. Fiorelli and A. H. Neto, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," vol. 40, pp. 2169–2176, 2008.
- [58] H. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586–3592, 2012.
- [59] P. Therkelsen, P. Rao, D. Sholes, A. Mckane, B. Meffert, and R. Green, "The value of regression models in determining industrial energy savings," pp. 389–399.
- [60] W. Hamer, "A practical approach to quantify RSA Section 12L EE tax incentives for large industry," North-West University, 2016. (Thesis – Phd).
- [61] C. Turhan, T. Kazanasmaz, I. E. Uygun, K. E. Ekmen, and G. G. Akkurt, "Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation," *Energy Build.*, vol. 85, pp. 115–125, 2014.
- [62] L. A. Botes, "Objective evaluation of industrial energy efficiency models for the RSA Section 12L tax incentive," North-West University, 2017. (Masters – MSc).
- [63] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed., Springer, 1990. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf> Date of access: 05 Feb. 2019.
- [64] R. D. Cook and S. Weisberg, *Applied regression including computing and graphics*, 1 Wiley Dr, Somerset, NJ 08873, USA: John Wiley & Sons, 1999.
- [65] E. Burman, D. Mumovic, and J. Kimpian, "Towards measurement and verification of energy performance under the framework of the European directive for energy performance of buildings," *Energy*, vol. 77, pp. 153–163, 2014.

- [66] T. Guiterman and M. Krarti, "Analysis of measurement and verification methods for energy retrofits applied to residential buildings," *ASHRAE Trans.*, vol. 117, no. PART 2, pp. 382–394, 2011.
- [67] L. E. Singer and D. Peterson, *International energy outlook 2010*, vol. 0484, no. May. 2011.
- [68] Z. Li and G. Huang, "Re-evaluation of building cooling load prediction models for use in humid subtropical area," *Energy Build.*, vol. 62, pp. 442–449, 2013.
- [69] G. Zhang, X. Liu, S. Lu, J. Zhang, and W. Wang, "Occurrence of typical antibiotics in Nansi Lake's inflowing rivers and antibiotic source contribution to Nansi Lake based on principal component analysis-multiple linear regression model," *Chemosphere*, vol. 242, p. 125269, 2020.
- [70] X. Gao, Q. Sun, H. Xu, and J. Gao, "Sparse and collaborative representation based kernel pairwise linear regression for image set classification," *Expert Syst. Appl.*, vol. 140, p. 112886, 2020.
- [71] Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," in *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, pp. 147–154.
- [72] A. Piasecki, J. Jurasz, and R. Skowron, "Forecasting surface water level fluctuations of lake Serwy (Northeastern Poland) by artificial neural networks and multiple linear regression," *J. Environ. Eng. Landsc. Manag.*, vol. 25, no. 4, pp. 379–388, 2017.
- [73] X. Yan and X. G. Su, *Linear regression analysis*. World Scientific Publishing Co., 2009.
- [74] S. Chatterjee and A. S. Hadi, *Regression analysis by example*, 5th ed. John Wiley & Sons, 2012. Available from ProQuest Ebook Central, <https://ebookcentral-proquest-com.nwulib.nwu.ac.za/lib/northwu-ebooks/detail.action?docID=918623> Date of access: 15 Sep. 2019.
- [75] Z. Wang and R. S. Srinivasan, "A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models," *Renew. Sustain. Energy Rev.*, vol. 75, no. November 2016, pp. 796–808, 2017.

- [76] Z. Wang and R. S. Srinivasan, "A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models," *Proc. - Winter Simul. Conf.*, vol. 2016-Febru, pp. 3438–3448, 2016.
- [77] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renew. Sustain. Energy Rev.*, vol. 73, no. March 2016, pp. 1104–1122, 2017.
- [78] A. S. Ahmad *et al.*, "A review on applications of ANN and SVM for building electrical energy consumption forecasting," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 102–109, 2014.
- [79] A. Reuter, K. R. Kuhner, and N. Wohlgemuth, "Energy models: methods and trends," *Proc. Fifth Forum Energy day Croat.*, pp. 41–56, 1996.
- [80] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning of statistical and machine-learning models using spatial data," *Ecol. Modell.*, vol. 406, no. April 2018, pp. 109–120, 2018.
- [81] R. Platon, V. R. Dehkordi, and J. Martel, "Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis," *Energy Build.*, vol. 92, pp. 10–18, 2015.
- [82] T. Amundson, S. Brooks, J. Eskil, and B. P. Administration, "Elements of defensible regression-based energy models for monitoring and reporting energy savings in industrial energy efficiency operation and maintenance projects," in *ACEEE Summer Study on Energy Efficiency in Industry*, 2013, pp. 1–14.
- [83] Y. Tae, R. Horesh, Y. Hwang, and Y. M. Lee, "Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings," *Energy Build.*, vol. 111, pp. 184–194, 2016.
- [84] D. Chakraborty and H. Elzarka, "Performance testing of energy models: are we using the right statistical metrics?," *J. Build. Perform. Simul.*, vol. 1493, pp. 1–16, 2017.
- [85] Y. Bai and W.-L. Jin, *Marine structural design*, 2nd ed. 225 Wyman Street, Waltham, MA 02451, USA: Butterworth-Heinemann, 2015.

- [86] H. Carstens, X. Xia, and S. Yadavalli, "Measurement uncertainty in energy monitoring : Present state of the art," *Renew. Sustain. Energy Rev.*, vol. 82, no. August 2017, pp. 2791–2805, 2018.
- [87] Q. Dong, K. Xing, and H. Zhang, "Artificial Neural Network for Assessment of Energy Consumption and Cost for Cross Laminated Timber Office Building in Severe Cold Regions," *Sustainability*, vol. 10, no. 1, p. 84, 2017.
- [88] M. Mittal, B. Bora, S. Saxena, and A. M. Gaur, "Performance prediction of PV module using electrical equivalent model and artificial neural network," *Sol. Energy*, vol. 176, no. July, pp. 104–117, 2018.
- [89] S. A. Kalogirou and M. Bojic, "Artificial neural networks for the prediction of the energy consumption of a passive solar building," *Energy*, vol. 25, no. 5, pp. 479–491, 2000.
- [90] L. Wu, G. Kaiser, D. Solomon, R. Winter, A. Boulanger, and R. Anderson, "Improving efficiency and reliability of building systems using machine learning and automated online evaluation," in *2012 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, 2012, Farmingdale, NY, pp. 1–6.
- [91] M. Mohandes, "Support vector machines for short-term electrical load forecasting," *Int. J. Energy Res.*, vol. 26, no. 4, pp. 335–345, 2002.
- [92] Y. Heo and V. M. Zavala, "Gaussian process modeling for measurement and verification of building energy savings," *Energy Build.*, vol. 53, pp. 7–18, 2012.
- [93] M. C. Burkhart, Y. Heo, and V. M. Zavala, "Measurement and verification of building systems under uncertain data : A Gaussian process modeling approach," *Energy Build.*, vol. 75, pp. 189–198, 2014.
- [94] B. Manobel, F. Sehnke, J. A. Lazzús, I. Salfate, M. Felder, and S. Montecinos, "Wind turbine power curve modeling based on Gaussian processes and artificial neural networks," *Renew. Energy*, vol. 125, pp. 1015–1020, 2018.
- [95] 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.

- [96] M. Zamarren and P. A. Gonza, "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network," *Energy Build.*, vol. 37, pp. 595–601, 2005.
- [97] R. Yokoyama, T. Wakui, and R. Satake, "Prediction of energy demands using neural network with model identification by global optimization," *Energy Convers. Manag.*, vol. 50, no. 2, pp. 319–327, 2009.
- [98] R. Mena, F. Rodríguez, M. Castilla, and M. R. Arahall, "A prediction model based on neural networks for the energy consumption of a bioclimatic building," *Energy Build.*, vol. 82, pp. 142–155, 2014.
- [99] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Applying support vector machine to predict hourly cooling load in the building," *Appl. Energy*, vol. 86, no. 10, pp. 2249–2256, 2009.
- [100] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks," *Energy Convers. Manag.*, vol. 50, no. 1, pp. 90–96, 2009.
- [101] F. Rossi, D. Velázquez, I. Monedero, and F. Biscarri, "Artificial neural networks and physical modeling for determination of baseline consumption of CHP plants," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4658–4669, 2014.
- [102] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, "Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors," *Energy Convers. Manag.*, vol. 49, no. 8, pp. 2272–2278, 2008.
- [103] S. Mesroghli, E. Jorjani, and S. Chehreh Chelgani, "Estimation of gross calorific value based on coal analysis using regression and artificial neural networks," *Int. J. Coal Geol.*, vol. 79, no. 1–2, pp. 49–54, 2009.
- [104] 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, pp. 1825–1832, 2008.
- [105] W. Li *et al.*, "Modeling urban building energy use: A review of modeling approaches and procedures," *Energy*, vol. 141, no. December, pp. 2445–2457, 2017.

- [106] H. Huang, L. Chen, and E. Hu, "A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings," *Energy Build.*, vol. 97, pp. 86–97, 2015.
- [107] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, 5th ed. John Wiley & Sons, 2012. Available from ProQuest Ebook Central: <https://ebookcentral-proquest-com.nwulib.nwu.ac.za/lib/northwu-ebooks/detail.action?docID=1211887>. Date of access: 15 Sep. 2019.
- [108] A. Kialashaki and J. R. Reisel, "Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks," *Appl. Energy*, vol. 108, pp. 271–280, 2013.
- [109] S. Weisberg, *Applied linear regression*, 4th ed. John Wiley & Sons, 2014. Available from ProQuest Ebook Central, <https://ebookcentral-proquest-com.nwulib.nwu.ac.za/lib/northwu-ebooks/detail.action?docID=1574352> Date of access: 15 Sep. 2019.
- [110] H. Ling, C. Qian, W. Kang, C. Liang, and H. Chen, "Combination of Support Vector Machine and K-Fold cross validation to predict compressive strength of concrete in marine environment," *Constr. Build. Mater.*, vol. 206, pp. 355–363, 2019.
- [111] B. Wang and Z. Mao, "Outlier detection based on a dynamic ensemble model: Applied to process monitoring," *Inf. Fusion*, vol. 51, no. May 2018, pp. 244–258, 2019.
- [112] M. Stockmann, R. Haber, and U. Schmitz, "Comparison of data and model based sensor data pre-treatment in a hydro-cracker," in *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 2009, vol. 2, no. PART 1, pp. 379–384.
- [113] "A review of nearest neighbor support vector machines hybrid classification models," *J. Appl. Sci.*, vol. 10, no. 17, pp. 1841–1858, 2010.
- [114] K. Hechenbichler and K. Schliep, "Weighted k-nearest-neighbor techniques and ordinal classification," 2004. https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf Date of access: 12 Oct. 2019.
- [115] J. M. Cruz, "Context-based identification of energy consumption in industrial plants," *Adv. Smart Syst. Res.*, vol. 3, no. 1, p. 31, 2012.

- [116] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Appl. Energy*, vol. 127, pp. 1–10, 2014.
- [117] R. Jovanović, A. A. Sretenović, and B. D. Živković, "Ensemble of various neural networks for prediction of heating energy consumption," *Energy Build.*, vol. 94, pp. 189–199, 2015.
- [118] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy*, vol. 123, pp. 168–178, 2014.
- [119] J. Granderson, P. N. Price, D. Jump, N. Addy, and M. D. Sohn, "Automated measurement and verification: Performance of public domain whole-building electric baseline models," *Appl. Energy*, vol. 144, pp. 106–113, 2015.
- [120] J. Granderson and P. N. Price, "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models," *Energy*, vol. 66, pp. 981–990, 2014.
- [121] K. Campbell, "Evaluating the feasibility of the 12L tax incentive for energy-intensive industries," 2017. (Masters – MSc).
- [122] A. Guisasola *et al.*, "The influence of experimental data quality and quantity on parameter estimation accuracy: Andrews inhibition model as a case study," *Educ. Chem. Eng.*, vol. 2006, pp. 139–145, 2006.
- [123] W. J. Egan and S. L. Morgan, "Outlier detection in multivariate analytical chemical data," *Anal. Chem.*, vol. 70, no. 11, pp. 2372–2379, 1998.
- [124] E. Acuña and C. Rodriguez, "On detection of outliers and their effect in supervised classification," 2005.
https://pdfs.semanticscholar.org/8a6e/59e5c29734655bd371125ca16c728b7b6daa.pdf?_ga=2.4064496.1904226464.1581972235-2097876701.1581972235 Date of access: 12 Jun. 2019.

- [125] K. M. Sunderland *et al.*, “The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project,” *BMC Med. Res. Methodol.*, vol. 19, no. 1, pp. 1–17, 2019.
- [126] S. Zhang, H. Cao, Z. Ye, Y. Zhang, and X. Hei, “An outlier detection scheme for dynamical sequential datasets,” *Commun. Stat. Simul. Comput.*, vol. 48, no. 5, pp. 1450–1502, 2019.
- [127] W. Booysen, L. A. Botes, and W. Hamer, “A practical methodology for the systematic identification of outliers,” in *Proceedings of the Conference on the Industrial and Commercial Use of Energy, ICUE*, 2017, Cape Town, South Africa, pp. 1–6.
- [128] B. Dong, C. Cao, and S. Eang, “Applying support vector machines to predict building energy consumption in tropical region,” *Energy Build.*, vol. 37, pp. 545–553, 2005.
- [129] G. C. Aye, R. Gupta, and P. Wanke, “Energy efficiency drivers in South Africa : 1965 – 2014,” pp. 1465–1482, 2018.
- [130] M. S. Garver, “Threats to the validity of logistics and supply chain management research,” *J. Bus. Logist.*, vol. 40, no. 1, pp. 30–43, 2019.
- [131] N. Lazarevic, A. G. Barnett, P. D. Sly, and L. D. Knibbs, “Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives,” *Environ. Health Perspect.*, vol. 127, no. 2, p. 26001, 2019.
- [132] J. Camacho and A. Ferrer, “Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects,” *Chemom. Intell. Lab. Syst.*, vol. 131, no. January, pp. 37–50, 2014.
- [133] R. Vitale, J. A. Westerhuis, T. Næs, A. K. Smilde, O. E. de Noord, and A. Ferrer, “Selecting the number of factors in principal component analysis by permutation testing—Numerical and practical aspects,” *J. Chemom.*, vol. 31, no. 12, pp. 1–15, 2017.
- [134] O. De Jesgs and M. T. Hagan, “Backpropagation through time for a general class of recurrent network,” in *IJCNN’01. International Joint Conference on Neural Networks*, 2001, pp. 2638–2643.

- [135] S. L. Wong, K. K. W. Wan, and T. N. T. Lam, "Artificial neural networks for energy analysis of office buildings with daylighting," *Appl. Energy*, vol. 87, no. 2, pp. 551–557, 2010.
- [136] M. Rezaeian-Zadeh, S. Zand-Parsa, H. Abghari, M. Zolghadr, and V. P. Singh, "Hourly air temperature driven using multi-layer perceptron and radial basis function networks in arid and semi-arid regions," *Theor. Appl. Climatol.*, vol. 109, no. 3–4, pp. 519–528, 2012.
- [137] S. Tippmann, "Programming tools: Adventures with R," *Nature*, vol. 517, no. 7532, pp. 109–110, 2015.
- [138] A. E. Ben-Nakhi and M. A. Mahmoud, "Cooling load prediction for buildings using general regression neural networks," *Energy Convers. Manag.*, vol. 45, no. 13–14, pp. 2127–2141, 2004.
- [139] J. Devore, N. Farnum, and J. Doi, *Applied statistics for engineers and scientists*, 3rd ed. 200 First Stamford Pl, Stamford, CT 06902, USA: Cengage Learning, 2014.
- [140] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied statistics for the behavioral sciences*, 5th ed. 125 High St, Boston, MA 02110, USA: Houghton Mifflin, 2003.
- [141] M. A. Arida, "Development and testing of building energy model using non-linear auto regression artificial neural networks," Greensboro, North Carolina, 2016.
- [142] M. A. R. Biswas, M. D. Robinson, and N. Fumo, "Prediction of residential building energy consumption: A neural network approach," *Energy*, vol. 117, pp. 84–92, 2016.
- [143] Ö. Altan Dombayci and M. Gölcü, "Daily means ambient temperature prediction using artificial neural network method: A case study of Turkey," *Renew. Energy*, vol. 34, no. 4, pp. 1158–1161, 2009.
- [144] B. B. Ekici and U. T. Aksoy, "Prediction of building energy consumption by using artificial neural networks," *Adv. Eng. Softw.*, vol. 40, no. 5, pp. 356–362, 2009.
- [145] S. Patel, r B. Kuma, Y. Badhe, B. Sharma, S. Saha, and S. Biswas, "Estimation of gross calorific value of coals using artificial neural networks," *Fuel*, vol. 86, pp. 334–344, 2007.
- [146] V. Kecman and J. P. Brooks, "Locally linear support vector machines and other local models," *Proc. Int. Jt. Conf. Neural Networks*, no. July, 2010.

- [147] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 91–91, 1997.
- [148] B. Schoikopf, "The kernel trick for distances," *NIPS'00 Proc. 13th Int. Conf. Neural Inf. Process. Syst.*, vol. 283–289, 2000.
- [149] J. Ma, "Function replacement vs. kernel trick," *Neurocomputing*, vol. 50, pp. 479–483, 2003.
- [150] S. M. Clarke, J. H. Griebisch, and T. W. Simpson, "Analysis of support vector regression for approximation of complex engineering analyses," in *Proceedings of DETC'03*, 2003, pp. 1–9.
- [151] S. A. Kalogirou, "Applications of artificial neural-networks for energy systems," *Appl. Energy*, vol. 67, no. 1–2, pp. 17–35, 2000.
- [152] K. Yu and X. Zhang, "Kernel nearest-neighbour algorithm," *Neural Process. Lett.*, vol. 15, no. 3, pp. 147–156, 2002.
- [153] T. Matcharashvili, Z. Czechowski, and N. Zhukova, "Mahalanobis distance-based recognition of changes in the dynamics of a seismic process," *Nonlinear Process. Geophys.*, vol. 26, no. 3, pp. 291–305, 2019.
- [154] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [155] E. LeDell et al., "h2o: R interface for 'H2O.'" 2018.
https://cran.rstudio.com/bin/windows/contrib/3.5/h2o_3.28.0.2.zip Date of access: 10 Oct. 2018.
- [156] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien." 2018.
- [157] L. Mouselimis, "KernelKnn: Kernel k Nearest Neighbors." 2018.
https://cran.rstudio.com/bin/windows/contrib/3.5/KernelKnn_1.1.0.zip Date of access: 10 Oct. 2018.

- [158] K. M. Mullen and I. H. M. van Stokkum, “nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS).” 2012.
- [159] G. Grolemund and H. Wickham, “Dates and Times Made Easy with {lubridate},” *J. Stat. Softw.*, vol. 40, no. 3, pp. 1–25, 2011.
- [160] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
- [161] H. Wickham, “Reshaping Data with the {reshape} Package,” *J. Stat. Softw.*, vol. 21, no. 12, pp. 1–20, 2007.
- [162] C. Brown, “formula.tools: Programmatic Utilities for Manipulating Formulas, Expressions, Calls, Assignments and Other R Objects.” 2018.
https://cran.rstudio.com/bin/windows/contrib/3.5/formula.tools_1.7.1.zip Date of access: 10 Oct. 2018.
- [163] C. O. Wilke, “ggribes: Ridgeline Plots in ‘ggplot2.’” 2018.
https://cran.rstudio.com/bin/windows/contrib/3.5/ggribes_0.5.2.zip Date of access: 10 Oct. 2018.
- [164] C. Sievert, “plotly for R.” 2018.
https://cran.rstudio.com/bin/windows/contrib/3.5/plotly_4.9.2.zip Date of access: 10 Oct. 2018.

APPENDIX A – ADDITIONAL DETAIL REGARDING MODELS

Appendix A1 – Overview of algorithms of models

In this dissertation, only the necessary details regarding the functional working of the different model types were provided. However, to understand the purpose of different hyperparameters and how it influences the different model types, a deeper understanding of each of the model types and how it is constructed (trained) is beneficial.

Linear models (LIN)

Linear regression can take single- or multivariate form (using either one of multiple independent variables). This can be expressed mathematically using Eq. A-1 and Eq. A-2 respectively.

$$y = wx + b \quad (\text{Eq. A-1})$$

w : Slope/gradient

b : Intercept/constant

$$y = w_1x_1 + w_2x_2 + \dots + w_ix_i + \dots + w_nx_n + b \quad (\text{Eq. A-2})$$

w_i : Slope (weight parameter) for independent variable i

n : Number of independent variables

b : Intercept/constant

The underlying assumption is that the residuals are normally and independently distributed with an average value of zero [82]. The most used algorithm to minimise the error by adjusting w and b is the ordinary least squares algorithm (OLS) [107], [109]. For a simple linear regression model, the parameters can be calculated as follows [73]:

$$w = \frac{\left(\sum_{j=1}^N x_j y_j - \frac{1}{N} \sum_{j=1}^N x_j \sum_{j=1}^N y_j \right)}{\sum_{j=1}^N x_j^2 - \frac{1}{N} \left(\sum_{j=1}^N x_j \right)^2} \quad (\text{Eq. A-3})$$

$$b = \left(\frac{1}{N} \sum_{j=1}^N y_j \right) - w \left(\frac{1}{N} \sum_{j=1}^N x_j \right)$$

j : The j th observation

N : The total number of observations

The calculation of the parameters is thus always explicit when using OLS. This contrasts with the ANN algorithm where the calculation of the parameters is influenced by several algorithm decisions.

Artificial neural networks (ANN)

The first ANN model was developed by McCulloch and Pitts [141]. It was developed to resemble the biological neural system [135]. For the discussion, we are going to consider a simple, but popular type of ANNs, the multi-layer perceptron (MLP) [13], [55], [138], [142], [56], [57], [68], [81], [87], [102], [135], [136]. This is the same type of ANN discussed in section 1.6. A more detailed representation than that given in section 1.6 of this type ANN is shown in Figure A.1.

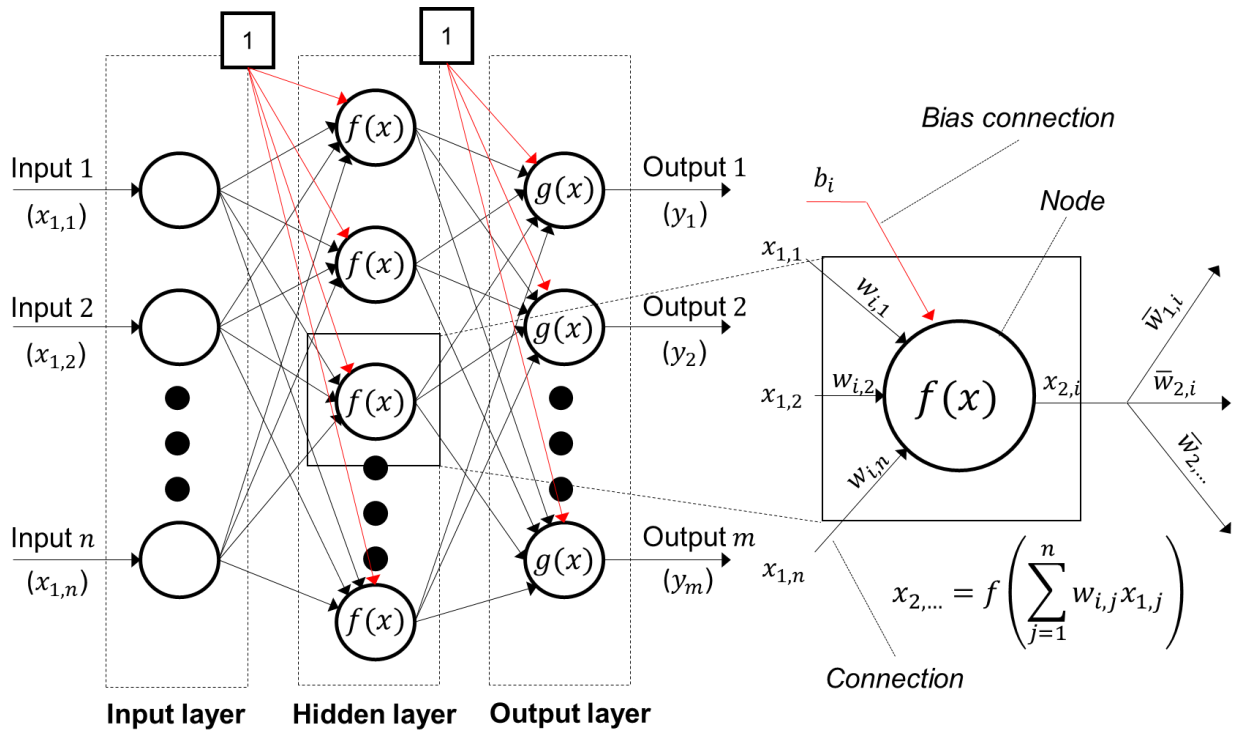


Figure A.1: A representation of a simple ANN or MLP

As mentioned in section 1.6, a MLP typically consists of three layers – an input layer, a hidden layer and an output layer [50], [102], [135], [136], [142], although multiple hidden layers are also possible [13], [51], [55], [94], [101]. Each layer consists of nodes and each node in a layer is

connected to each node in the next layer with a connection. A weight is assigned to each connection.

The input layer only feeds the inputs $x_{1,1}, \dots, x_{1,n}$ to the next layer. Each input is multiplied with the weight of each connection from that input node. Each node in the hidden layer has an activation function assigned to it, $f(x)$ [50]. Except for the input nodes, each node also has a bias assigned to it. For node i , the input to this function is the sum of the products between the relevant weights to that node and inputs plus the bias, i.e. $\sum_{j=1}^n (w_{i,j}x_{1,j} + b_i)$. The result of the function is the output of the node, $x_{2,i}$. This is then multiplied with the different weights of the relevant connections and the process repeated in the next layer.

There is thus a set of weights W and biases B that can be adjusted to minimise some objective function. This is done with some training algorithm, normally some variant of a technique called backpropagation. The algorithm uses repeated steps or epochs [143]. During each step, the algorithm cycles through the training instances provided and the weights are adjusted to decrease the average error of some error function. The weights are initially randomised to avoid symmetry [141].

Besides standard backpropagation [89], [96], [99], [135], popular training algorithms for training ANNs are the Levenberg-Marquardt Backpropagation [61], [68], [77], [87], [98], [136], [143], [144], resilient backpropagation [55], [88], [102] and Bayesian Regularisation Backpropagation [77], [83]. The Levenberg-Marquardt algorithm is especially popular as it has relatively good convergence speed without the need for computing the Hessian matrix [136]. A detailed review and comparison of these algorithms is given by De Jesús and Hagan [134].

The quality of an ANN is dependent on the quantity and the quality of the data. ANN performance can be improved by using data that is well distributed and void of instrumental noise and errors [141], [145].

Support Vector Machines (SVM)

Many authors have reviewed the working of SVM and its application to regression (SVR) [48], [99], [100], [118], [146] after its introduction by Vapnik *et al.* [147]. Therefore, only a simple review of the basic principles of SVMs is offered here.

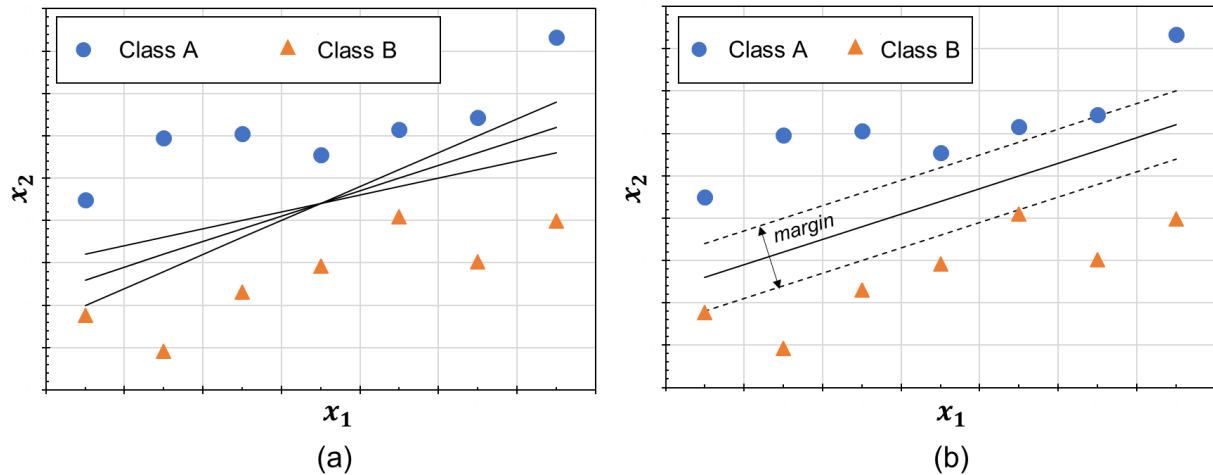


Figure A.2: Demonstration of separating two classes with a hyperplane with a certain margin (in a two-dimensional space)

The application of SVM to a classification problem is first considered. This is represented in Figure A.2. The goal of SVM is to find the separator(s) or hyperplane(s) that best separates between the classes of data [113]. For the case presented, there are only two classes of data, so there is only one hyperplane. The hyperplane is a straight line because the problem occurs in a two-dimensional space. In a three-dimensional space, the hyperplane would be a flat plane and in higher spaces it would be a multi-dimensional plane.

There are infinitely many lines that can be drawn to separate the two classes (illustrated in Figure A.2(a)). However, the goal is to find the line that best separates the two classes by maximising the *margin* between the data points (illustrated in Figure A.2(b)). The margin is the sum of the distances from the hyperplane to the closest data point of each class [113].

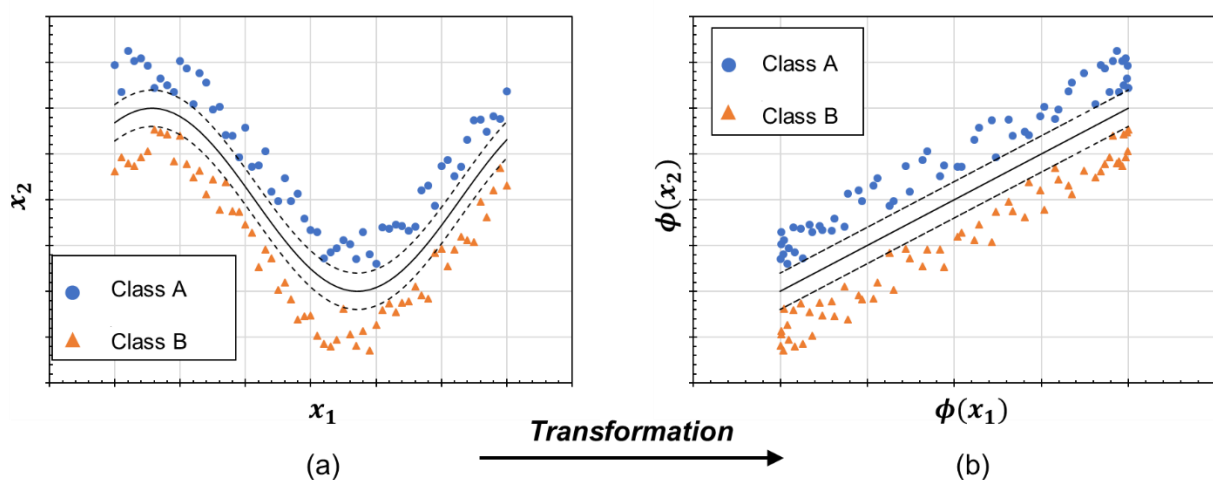


Figure A.3: Illustration of the mapping of non-linear SVM

For the example in Figure A.2, the classes are linearly separable. This is, however, not the case for the adjusted scenario presented in Figure A.3. A linear separating line can be fit to minimise the number of wrongly classified data points, but there is a better approach.

The feature space can be transformed from $\mathbf{x} = (x_1, x_2)$ to $\boldsymbol{\phi}(\mathbf{x}) = (\phi(x_1), \phi(x_2))$ so that it becomes a linear problem. This is done by applying functions f_1 and f_2 to x_1 and x_2 respectively. In this example these functions are

$$f_1(x_1) = \sin(x_1) \quad (\text{Eq. A-4})$$

$$f_2(x_2) = x_2 \quad (\text{Eq. A-5})$$

Another option might have been to apply a multivariable function f_3 that take both x_1 and x_2 into account to create a new variable y and transform the non-linear two-dimensional problem to a linear three-dimensional problem.

$$y = f_3(x_1, x_2) \quad (\text{Eq. A-6})$$

As mentioned, the principle of SVM can be transferred to regression. Consider again the scenario of Figure A.2. Suppose that we are not interested in the classes of data, but in the relationship between x_1 and x_2 .

Now, let x_1 be x and x_2 be y . The problem now becomes a simple linear equation [99], [100].

$$y = f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \quad (\text{Eq. A-7})$$

For the purpose of a generic discussion, we will consider $\mathbf{x}$ as an n -dimensional vector. However, instead of trying to minimise the total error as with linear regression, all errors that are within a certain distance, ε , of the true value is ignored and only the errors greater than ε is considered [99], [100], [118]. This is the epsilon-SVR method, the most popular in literature, as opposed to the Least Squares SVR method [118]. The solution is

$$\min \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \frac{1}{N} \sum_{i=1}^N L_{\varepsilon}(y_i, f(x_i)) \right) \quad (\text{Eq. A-8})$$

$$L_{\varepsilon}(y_i, f(x_i)) = \begin{cases} 0 & \text{for } |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon & \text{for others} \end{cases}$$

where N is the number of data points used to train the SVM.

The term for the weights vector, $\mathbf{w}$, is included to make the function as flat as possible to control the capacity of the function, called the regularised term [100], [118]. The preference that is given to the second term is controlled by the cost or penalty parameter, C . This determines how much the error terms influence the algorithm.

The function can now be rewritten as

$$\min \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \frac{1}{N} \sum_{i=1}^N (\xi_i + \xi_i^*) \right) \text{ by varying } \mathbf{w}, \xi_i, \xi_i^* \quad (\text{Eq. A-9})$$

where ξ_i and ξ_i^* is the residue of the absolute error greater than ε , i.e.

$$\text{subject to: } \begin{cases} y_i - \mathbf{w}_i \cdot \mathbf{x} - b \leq \varepsilon + \xi_i^* \\ \mathbf{w}_i \cdot \mathbf{x} + b - y_i \leq \varepsilon + \xi_i \\ \xi_i \geq 0, \xi_i^* \geq 0 \end{cases} \quad (\text{Eq. A-10})$$

This problem can be transformed to a maximisation problem based on another set of weights, $\{\alpha_i\}, \{\alpha_i^*\} \forall i = 1, \dots, N$ (referred to as the Lagrange multipliers) and the dot product between the various vectors that form the dataset.

$$\max \left(-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N ((\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\mathbf{x}_i \cdot \mathbf{x}_j)) - \varepsilon \sum_{i=1}^N (\alpha_i - \alpha_i^*) + \sum_{i=1}^N y_i (\alpha_i - \alpha_i^*) \right) \text{ by varying } \alpha_i, \alpha_i^* \quad (\text{Eq. A-11})$$

$$\text{subject to: } \begin{cases} \sum_{i=1}^N (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

However, the problem becomes more difficult when considering the non-linear problem given in Figure A.3. As before, a function can be applied to map $\mathbf{x}$ to higher dimensional space where the problem is linear, $\phi(\mathbf{x})$. Unfortunately finding a suitable function is often not simple and now the dot product that needs to be calculated is between $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$.

Here, a clever idea called the kernel trick allows us to transform the data to a higher-dimensional space without the actual function of $\phi(\mathbf{x})$ being known [148]. Instead of trying to find a function

$\phi(x)$, we assume a kernel function, $K(x_i, x_j)$, which satisfies Mercer's Theorem and calculates the dot product of $\phi(x_i)$ and $\phi(x_j)$ [149].

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (\text{Eq. A-12})$$

The final model then becomes

$$y = f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \quad (\text{Eq. A-13})$$

So instead of having a vector of weights, w , we now have two sets of weights $\{\alpha_i\}, \{\alpha_i^*\} \forall i = 1, \dots, N$.

There are many different types of kernel functions. A popular function is the Gaussian radial basis function (RBF) [90], [99], [100], [128].

SVMs can solve non-linear problems even with small quantities of data [58]. One advantage that SVMs have over ANNs is that global minima can be located instead of local minima [150]. SVMs are also referred to as SVR when used for regression [118]. The main advantage of SVMs, is that the structure risk minimisation principle is adopted which has been shown to be superior to the empirical risk minimisation principle [99], [128].

ANNs need a lot of training data, but can be set up in a generic manner with good results [49]. For SVMs the kernel function must be assumed, but it does not need the same amount of data for each variable and retains some level of interpretability.

The working of the MLP has been discussed, but different kinds of ANNs exist, such as backpropagation neural nets and recurrent neural nets [52]. There are various uses of ANNs – diagnostics, optimisation, process control and energy management [89], [151].

k-Nearest Neighbours (KNN)

As mentioned in section 1.6, KNN is a simple algorithm that predicts the output value for a new data point based on the k -nearest data points in an n -dimensional space [152].

Consider again the example given Figure 1.11 in section 1.6. It shows a two-dimensional mapping of three-dimensional dataset, $y \sim f(x_1, x_2)$. The y -value for each data point is shown next to it. As mentioned, if we want to approximate the output value for a new data point, x^* , referred to as the

query point, for $k = 5$. To do this, the five closest values to that data point in the input space (in this case a two-dimensional space) is identified and the average determined of the corresponding y -values ($\frac{1}{k} \sum_{i=1}^k y_i$).

However, with this method equal preference is given to each data point. In many cases it might be better to give less preference to data points that are further away from our query point, x^* . To do this, larger weights are assigned to points close to the query point and smaller weights to points further away from it and then normalise according to the sum of the weights.

$$y^* = \frac{\sum_{i=1}^k y_i c_i}{\sum_{i=1}^k c_i} \quad (\text{Eq. A-17})$$

To determine the weights, we turn again to kernel functions of which the output decreases as distance increases. This is further discussed in Appendix A2.

Appendix A2 – Review on hyperparameter options

As mentioned, each of the different models has different hyperparameters. The applicable hyperparameters are listed per model in Table A.1, along with a description of each.

Table A.1: Description of hyperparameters

Model	Hyperparameter	Description
LIN	Intercept	The inclusion or omission of the intercept or constant
ANN	Number of hidden layers	The number of hidden layers influence the number of passes and ultimately the number of weights
	Number of neurons	The number of neurons used in each hidden layer
	Activation functions	The function used to transform the net input to a neuron.
SVM	Kernel function	The function to transform the non-linear problem into a linear problem. Specifically, it is the function that is used to predict the dot product of the transformed variables (see Appendix A1).
	Lambda	The constant parameter of the kernel function.
	Cost	The penalty parameter that determines the combined influence of the error terms on the SVM algorithm.
KNN	k -value	The number of closest points that should be used when determining the average.
	Calculation method	The method for calculating the distances.
	Kernel functions	The function used to assign a weight to the dependent value of each data point when calculating the average.
	Lambda	The constant parameter of the kernel function.

The LIN model type only has one hyperparameter – the inclusion or omission of the intercept term. (as used by Gallagher *et al.* [46]). Many of the hyperparameters for the machine learning model types are either positive integers or positive real numbers. The number of hidden layers and the number of neurons in each hidden layer for ANN and the number of nearest-neighbours (the k -value) for the KNN are positive integers greater than zero. The lambda and cost value of the SVM model type are positive real numbers greater than zero.

Each of the machine learning models also use some function as a hyperparameter – the activation function for the ANN model type and the kernel functions for the SVM and KNN model types. For each of these there are different options. Table A.2 gives four different options for the activation function of the ANN model type and occurrences of the use thereof in literature.

Table A.2: Different activation functions used in literature for ANNs

Activation function		Used in hidden layer	Used in output layer
Name	Formula		
<i>Hyperbolic tangent</i>	$\tanh(x)$	[13], [55], [143], [144], [56], [68], [77], [87], [96]–[98], [101]	[13], [55], [56], [77], [96]
<i>Logistic sigmoid</i>	$\frac{1}{1 + e^{-x}}$	[55], [56], [61], [83], [95], [99], [100], [102], [136], [138]	[55], [56], [61], [83], [99], [100], [138]
<i>Linear</i>	cx	[143]	[55]–[57], [68], [98], [101], [102], [136], [144][87], [143]
<i>Rectified linear unit</i>	$\max\{0, x\}$	[51]	[51]

For the ANN activation function, the function must have the property that the output is between negative one and one. Often a different activation function is used in the hidden layers than in the output layer. Thus, multiple activation functions can be used for a single ANN model. This is not the case for the kernel functions of the SVM and KNN model types.

For the SVM model type, there are different examples. These include the linear function, the Gaussian RBF and the sigmoid function, given respectively in Eq. A-18 to Eq. A-20.

$$K(x_i, x_j) = x_i \cdot x_j \quad (\text{Eq. A-18})$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\lambda^2}\right) \quad (\text{Eq. A-19})$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\lambda \mathbf{x}_i^T \mathbf{x}_j + c) \quad (\text{Eq. A-20})$$

The linear kernel function is the dot product between two data points ($\mathbf{x}_i$ and $\mathbf{x}_j$) in the calculation procedure. That is, the data remains unchanged and no transformation of the domain is done. The only difference between the SVM model when using the linear kernel function and a standard LIN model, is that the SVM model fits the parameters in a different manner, as explained in Appendix A1. The remaining functions each represent a different type of transformation. The kernel functions of the SVM model type relate to the kernel functions of the KNN model type. Different examples of these are illustrated in Figure A.4, with the mathematical equations given in Table A.3.

As far as the KNN model type is concerned, the first hyperparameter after the k -value is the distance calculation method. The distance between various data points can be calculated using various techniques [114]. Examples include:

Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{s=1}^n (x_{is} - x_{js})^2 \right)^{\frac{1}{2}} \quad (\text{Eq. A-21})$$

Absolute distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^n |x_{is} - x_{js}| \quad (\text{Eq. A-22})$$

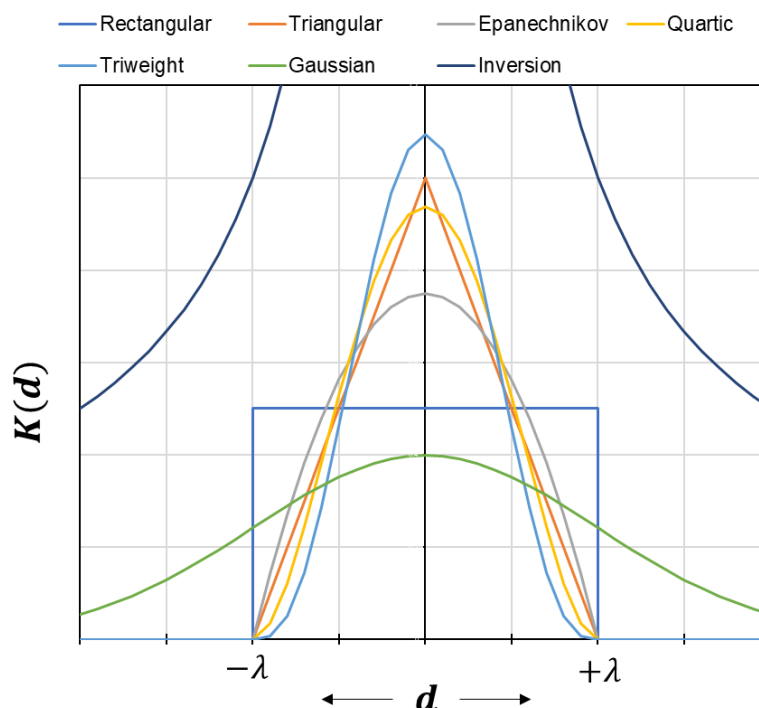


Figure A.4: Illustration of different kernel functions for KNN

Table A.3: Mathematical equations of different kernel functions for KNN [63], [114]

Name	Formula
Rectangular kernel	$K(d) = \frac{1}{2} \cdot I(d \leq \lambda)$
Triangular kernel	$K(d) = (\lambda - d) \cdot I(d \leq \lambda)$
Epanechnikov kernel	$K(d) = \left(\frac{3}{4}(\lambda - d^2)\right) \cdot I(d \leq \lambda)$
Quartic kernel	$K(d) = \left(\frac{15}{16}(\lambda - d^2)^2\right) \cdot I(d \leq \lambda)$
Triweight kernel	$K(d) = \left(\frac{35}{32}(\lambda - d^2)^3\right) \cdot I(d \leq \lambda)$
Gaussian kernel	$K(d) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d^2}{2\lambda^2}\right)$
Inversion kernel	$K(d) = \frac{1}{ d }$

The different functions have different deterioration rates from the centre. Many of the functions ignore points beyond a distance of λ , such as the rectangular- and triangular kernel functions. With the use of these types of kernel functions, it is especially important that there is a good distribution of data points, so that there is at least one data point closer than a distance of λ from any point to be predicted.

For the ANN model type, only one hidden layer was used according to common practice [77], [83], [103], [135], [136], [143]. The maximum number of neurons to be included was chosen as 11. For the SVM model type, the linear, radial and sigmoid kernel functions were chosen to include a variety of function forms. The lambda and cost values were chosen accordingly by giving consideration to that included by Gallagher *et al.* [46].

From the available list of kernel functions for the KNN model type, the uniform and triangular kernel types were chosen. This ensures that both an equal-weight option (where no preference is given between the different data points) and an option where greater preference is given to the values of points closer to the point to be predicted. The lambda-value is chosen to be the integer values greater and equal to one (1, 2 and 3) to ensure that no points are excluded in the scaled dataset. The Mahalanobis, a more complex distance calculation method [153], and the Euclidean distance calculation method were both included.

APPENDIX B – MODEL CONSTRUCTION & METRIC CALCULATION

Appendix B1 – Model construction with the R-programming language

This section provides additional details regarding the R-programming packages used for model construction and also gives additional details regarding model training.

The R-programming packages and the main functions used from it are summarised per model type in Table B.1.

Table B.1: Summary of the R-packages used per model type

Model	Package	Version	Reference	Function	Purpose
LIN	stats	3.5.3	[154]	lm	Training
				predict.lm	Prediction
ANN	h2o	3.20.0.2	[155]	h2o.deeplearning	Training
				h2o.predict	Prediction
SVM	e1071	1.7-0	[156]		Training
	stats	3.5.3	[154]	predict	Prediction
KNN	KernelKnn	1.0.8	[157]	KernelKnn	Training / Prediction
SMA	base	3.5.3	[154]	mean	Prediction ¹
WMA	nnls	1.4	[158]	nnls	Training ²

¹Model training is done separately for each of the constituting functions

²Prediction is done through matrix multiplication from the coefficients determined using the “nnls” function

Other packages that had been used include:

- lubridate (v. 1.7.4) [159]
- ggplot2 (v. 3.2.0) [160]
- reshape2 (v. 1.4.3) [161]
- formula.tools (v.1.7.1) [162]
- ggrridges (v. 0.5.1) [163]
- plotly (v. 4.9.0) [164]

For the ANN model training, a maximum number of epochs for the ANN training process was decided upon and kept constant throughout the entire procedure to remove the influence of a varying number of epochs. The “seed”-value on which the random number generation process is based when assigning the initial weights to the ANN was also fixed. This ensures that only the training data provided will influence the results produced by the ANN training process.

The WMA model training was done through non-negative least squares (NNLS) regression without the use of an intercept. Thus, the output of each model type was treated as a variable in the dataset and NNLS regression used to match outputs of the model types with the true energy usage. The final function can be expressed mathematically as,

$$\hat{y}_{WMA} = k_{LIN}\hat{y}_{LIN} + k_{ANN}\hat{y}_{ANN} + k_{SVM}\hat{y}_{SVM} + k_{KNN}\hat{y}_{KNN} \quad (\text{Eq. B-1})$$

$\hat{y}_i$: Series of output predicted by model i

k_i : Non-negative weight assigned to model i

Appendix B2 – Formulas for error metrics

The formulas for the different error metrics that are referred to in this study are given in Eq. B-2 to Eq. B-6.

$$\text{MAPE} = \frac{1}{N} \sum_{j=1}^N \left| \frac{y_j - \hat{y}_j}{y_j} \right| \times 100\% \quad (\text{Eq. B-2})$$

y_j : The j^{th} true value

$\hat{y}_j$: The j^{th} predicted output

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2} \quad (\text{Eq. B-3})$$

y_j : The j^{th} true value

$\hat{y}_j$: The j^{th} predicted output

$$\text{CV(RMSE)} = \frac{\text{RMSE}}{\bar{y}} \times 100\% \quad (\text{Eq. B-4})$$

$\bar{y}$: The average (mean) of y

$$\text{NMBE} = \frac{\frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2}{\bar{y}} \times 100\% \quad (\text{Eq. B-5})$$

$$\text{Std. dev.} = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \bar{y})^2} \quad (\text{Eq. B-6})$$

APPENDIX C – ADDITIONAL RESULTS

Appendix C1 – Initial Inspection Plots

This section provides the inspection plots for case studies II to IV, as done for case study I in section 3.1. The case studies are discussed consecutively.

Case study II

The time-series plots for the energy produced and for the various potential drivers are given in Figure C.1 and Figure C.2.

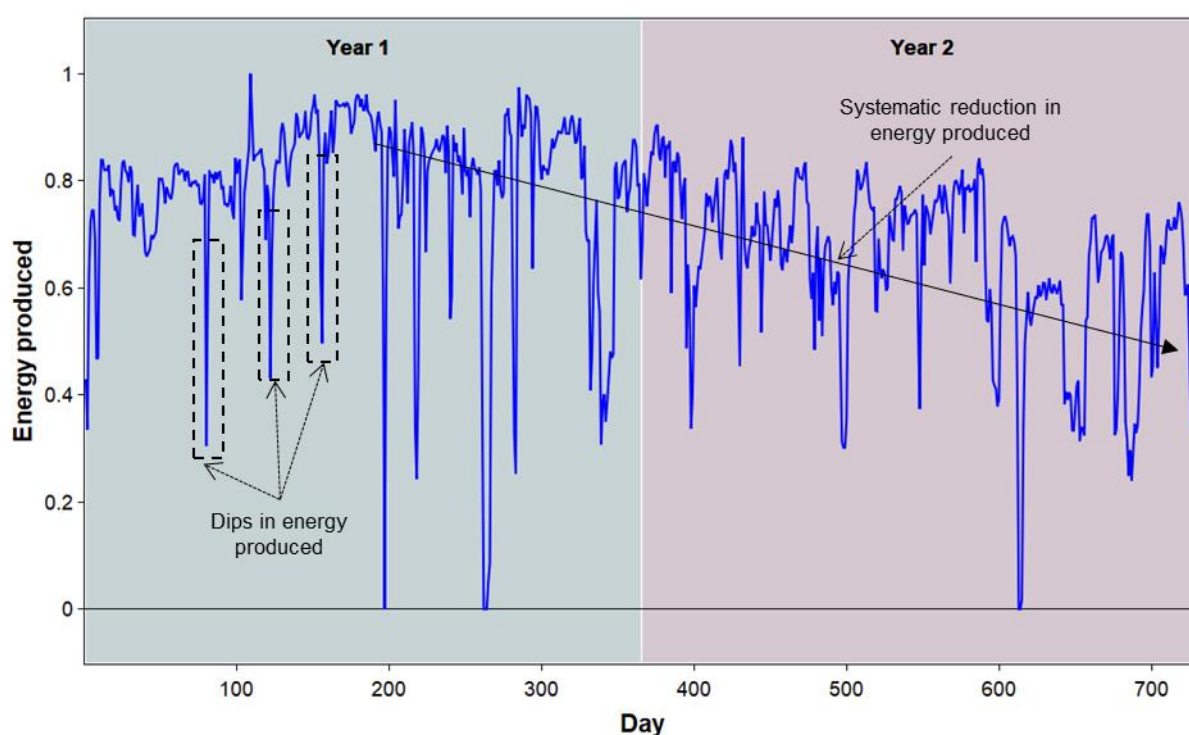


Figure C.1: Time-series plot of the energy produced for case study II

In Figure C.1, clear dips in the energy produced are seen as well as an overall, systematic reduction over time of the average energy output. The same is true when looking at the total energy supplied or the natural gas flow in Figure C.2. This is because the supply of natural gas is quite variable and reduced over time. The NG flow and the total energy supplied is closely related to one another as there is very little variation in the NG HV, which can also be seen from Figure C.2.

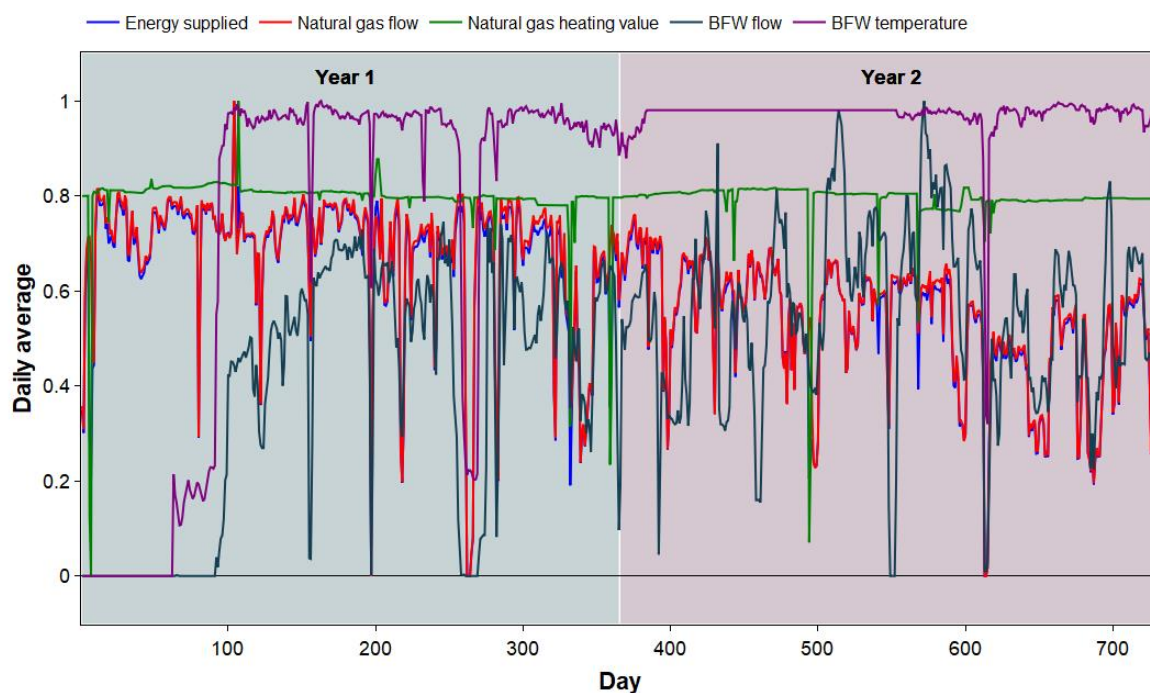


Figure C.2: Time-series plot of the potential variables for case study II

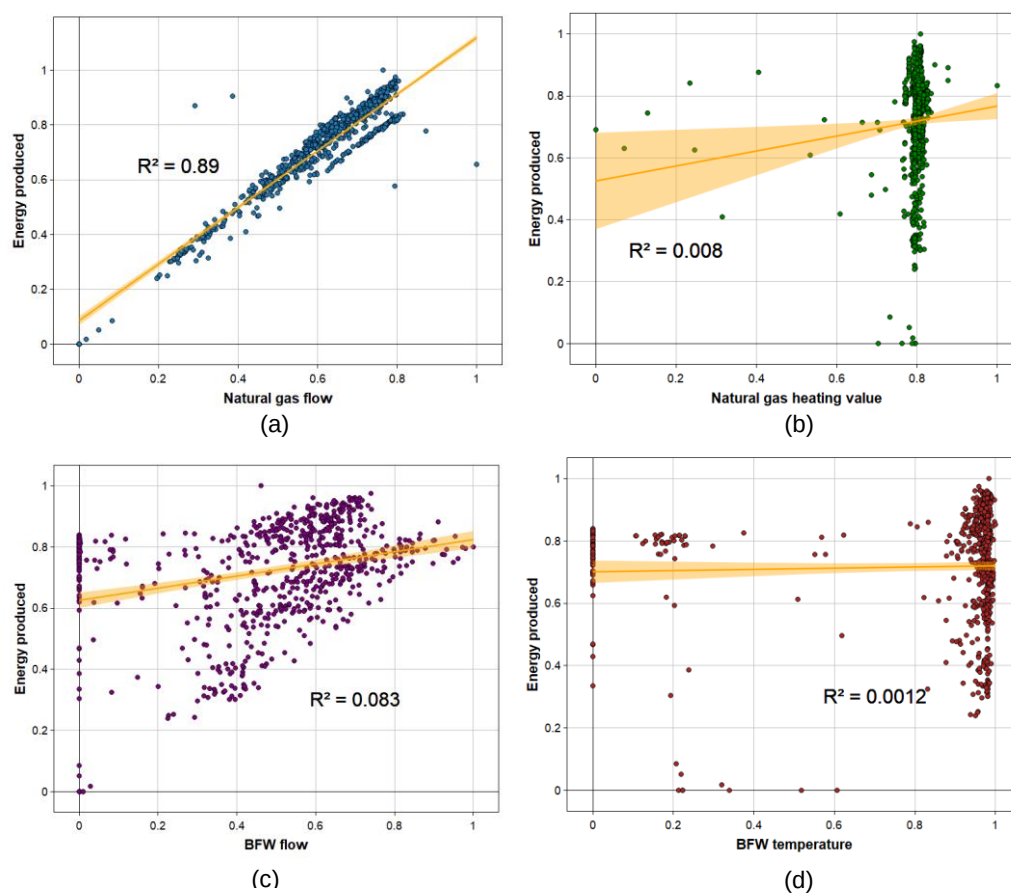


Figure C.3: Scatterplots of the energy produced vs (a) the natural gas flow; (b) the NG heating value; (c) BFW flow; and (d) BFW temperature for case study II

Figure C.3 shows the various scatter plots of the energy produced against the different potential drivers. It has already been shown in section 3.4 that there is a strong linear relationship between the energy supplied and the energy produced. This is consequently also true for the energy flow, as seen in Figure C.3(a). However, all the other potential drivers show very weak linear relationships with the energy output and do not seem to be non-linear. This explains the poor performance of the linear model when using the BFW flow as a driver, as discussed under section 3.5.

Case study III

The time-series plot of the energy usage for case study III is shown in Figure C.4 while the same is shown for the potential drivers in Figure C.5.

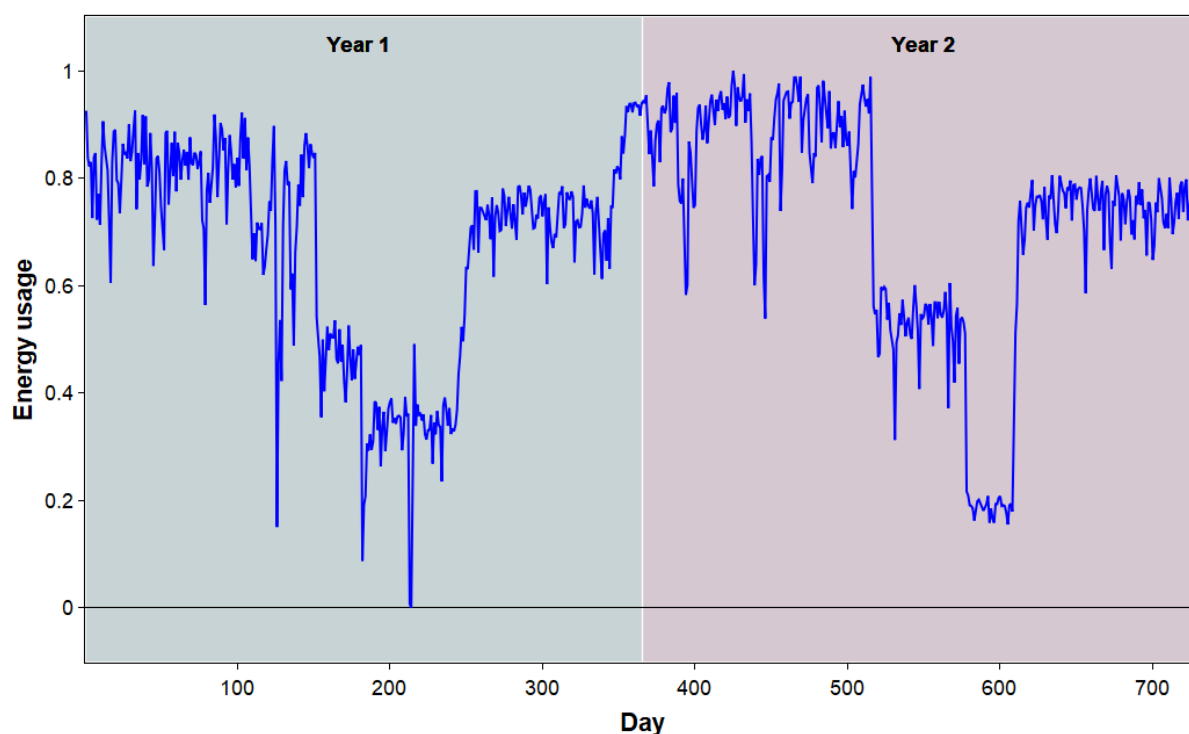


Figure C.4: Time-series plot of the energy usage for case study III

In Figure C.4, the different modes of operation that occur due to the different amounts of furnaces being used are again apparent as it was with the outlier inspection plot given in section 3.3. The same can be seen when looking at the reductants and ore used as well as the total production in Figure C.5. This is somewhat less apparent when looking at the flux used, as changes in the flux type were made.

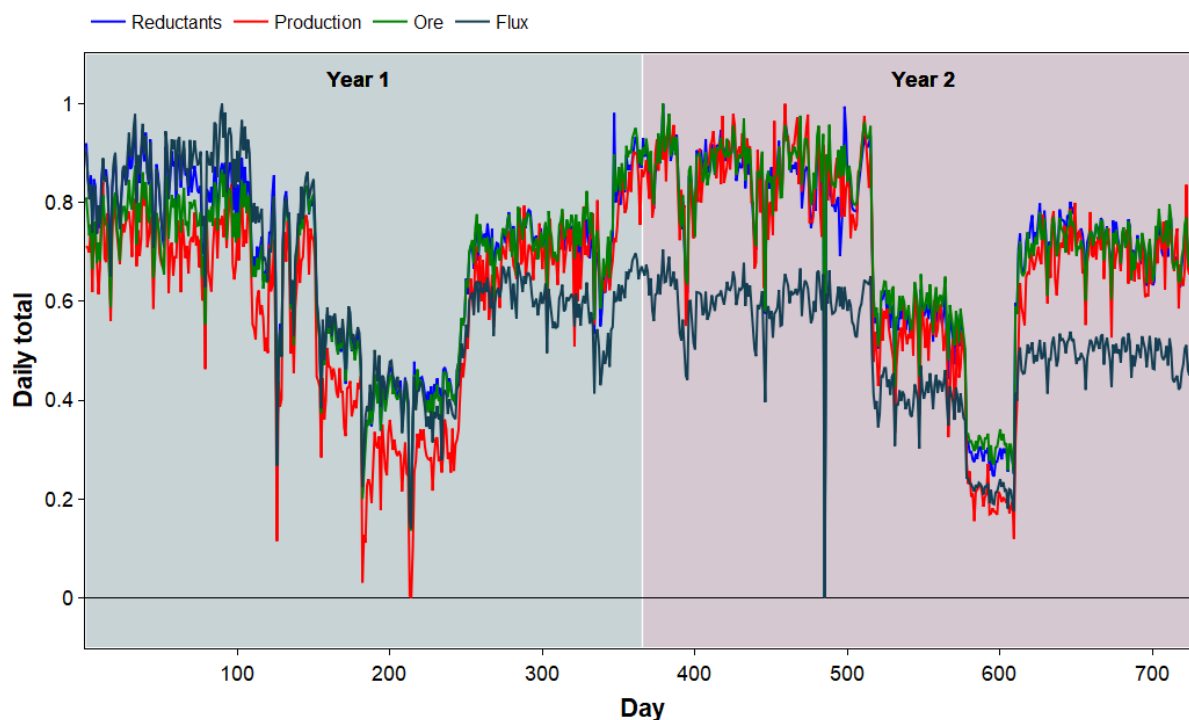


Figure C.5: Time-series plot of the potential drivers for case study III

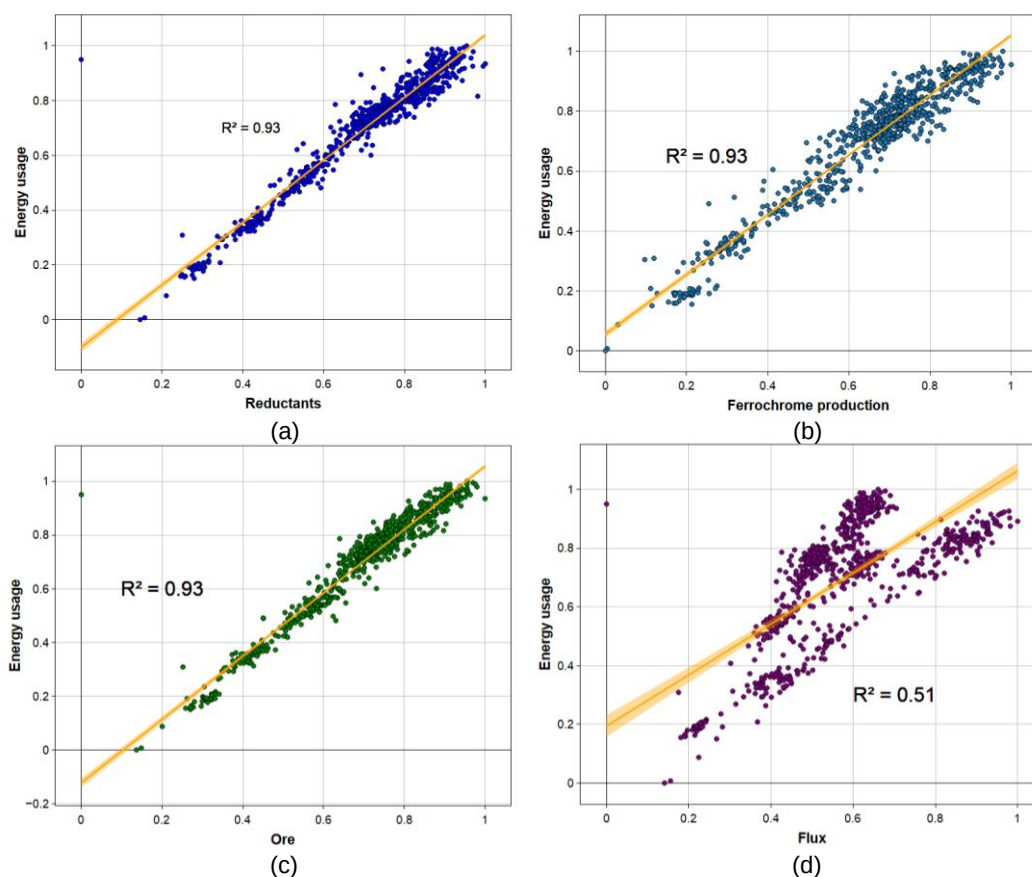


Figure C.6: Scatterplots of the energy usage vs (a) the reductant usage; (b) the ferrochrome function; (c) the ore usage; and (d) flux usage for case study III

Figure C.6 shows the scatter plots of the energy usage against the respective potential drivers. There is a strong linear relationship between the energy usage and the reductants used, the ore used and the ferrochrome production. There is a weak linear relationship between the energy usage and the flux used due to the changes that occurred.

Case study IV

The time-series plots for the energy usage and the different potential drivers are shown in Figure C.7 and Figure C.8 respectively. Seasonal effects are apparent when looking at most of the parameters, except the evaporator outlet temperature that needs to be maintained between certain limits for the different min services. Certain dips can also be seen in the different parameters, which is done purposefully by the mine to save energy during certain times of the day and over non-production weekends.

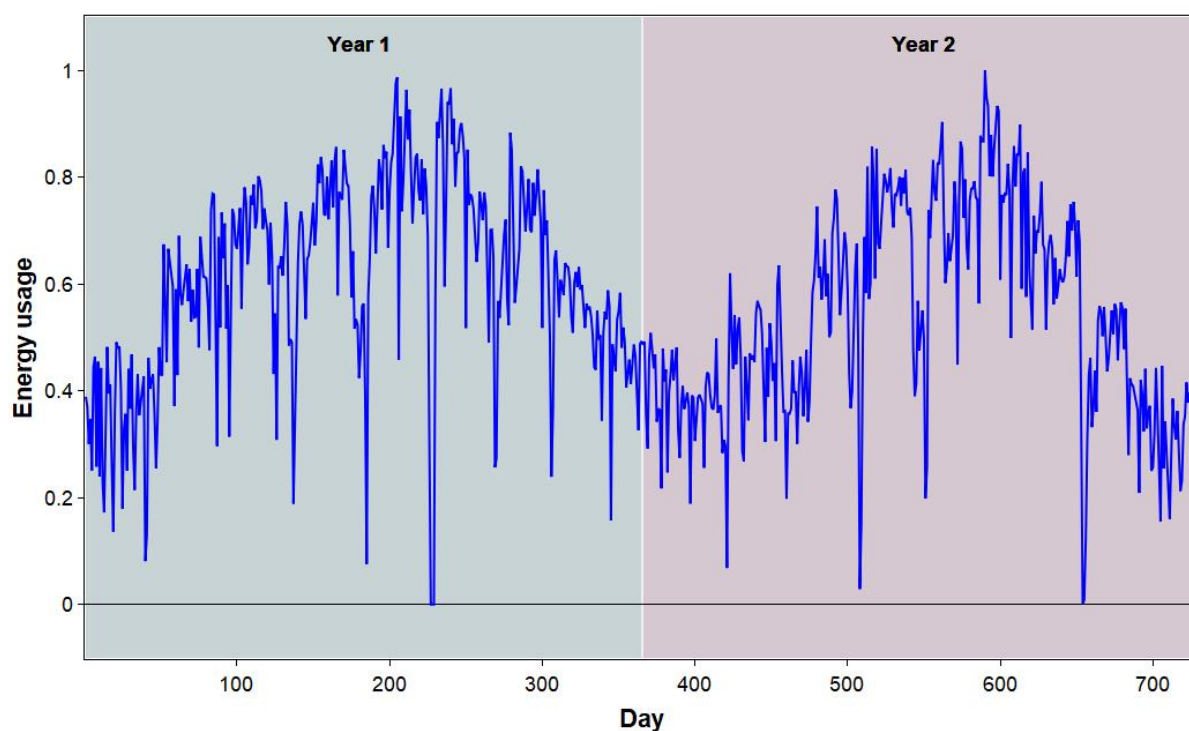


Figure C.7: Time-series plot of the energy usage for case study IV

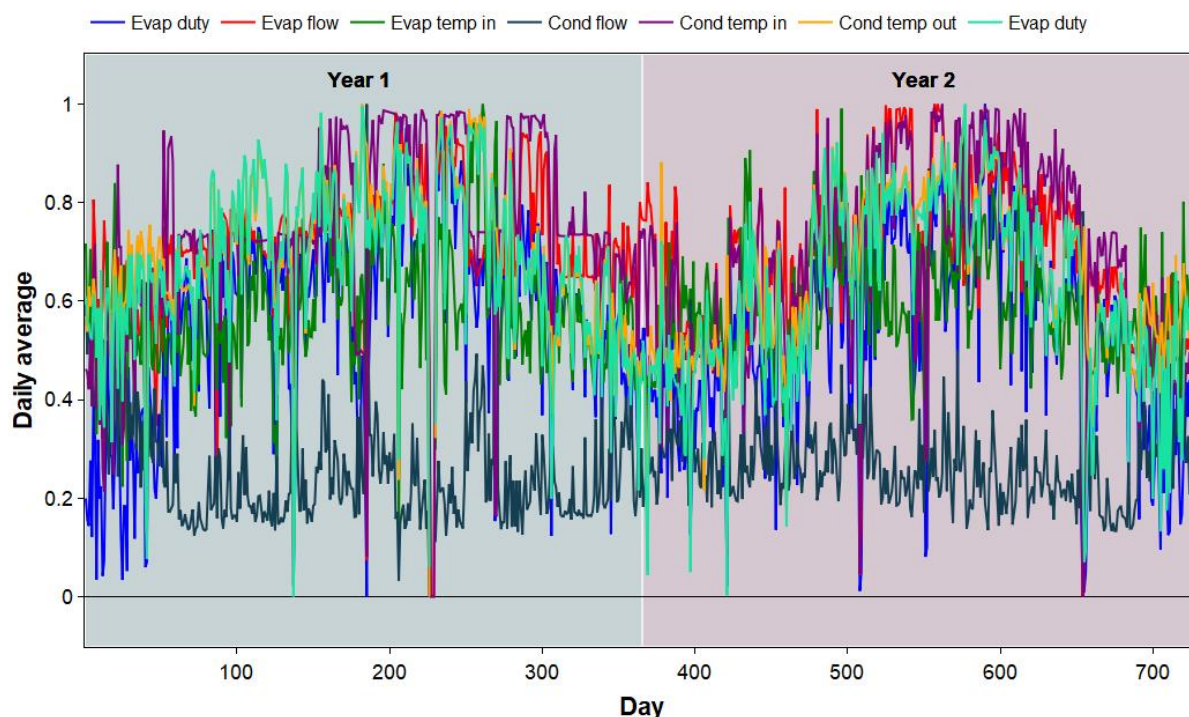


Figure C.8: Time-series plot of the potential drivers for case study IV

Figure C.9 shows the respective scatter plots of the energy usage against the potential drivers. The strongest degree of linearity appears between the energy usage and the evaporator and condenser flows. There are positive linear relationships between the energy usage and the condenser temperatures but a very weak relationship between the energy usage and the evaporator inlet temperature (TI).

There appears to be a non-linear relationship between the evaporator outlet temperature (TO) and the energy usage, specifically an inversely proportional relationship. As the TO increases, the energy usage decreases less and less. Put differently, the greater the reduction in TO, the greater the relative increase in energy usage to obtain the reduction. This explains the low R^2 value.

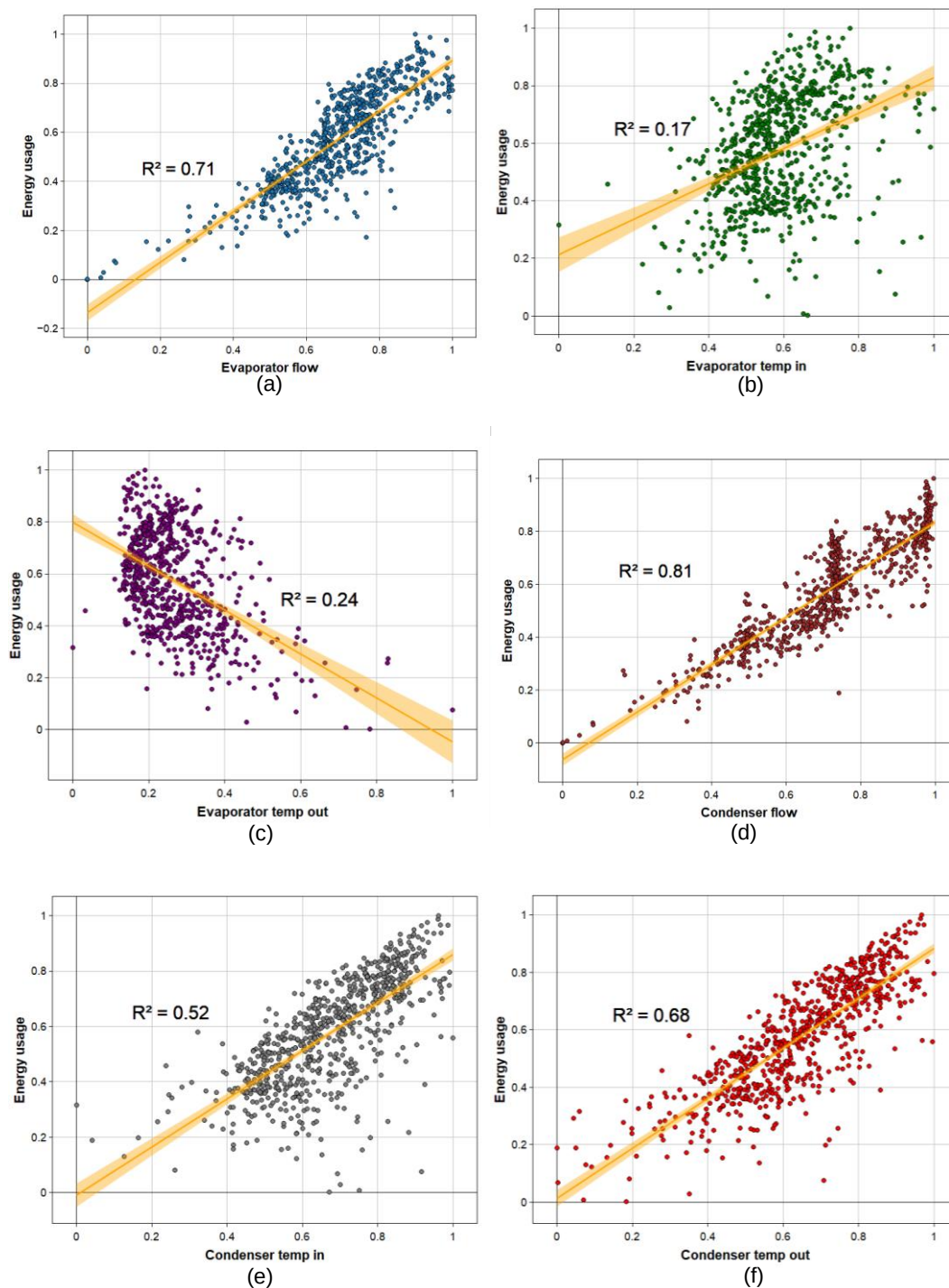


Figure C.9: Scatterplots of the energy usage vs (a) the evaporator flow; (b) the evaporator temperature in; (c) the evaporator temperature out; and (d) the condenser flow; (e) the condenser temperature in; and (f) the condenser temperature out for case study IV

Case study V

From the time-series plots for case study V shown in Figure C.10 and Figure C.11, it is apparent there are several regular dips in the energy usage and the different potential energy drivers. These dips occur over weekends when the mine is extracting much less or no ore and thus less flow and pressure are needed. There are also three noteworthy dips that occur over a longer period. These dips occur due to shutdowns on the mine when focus is placed on maintenance.

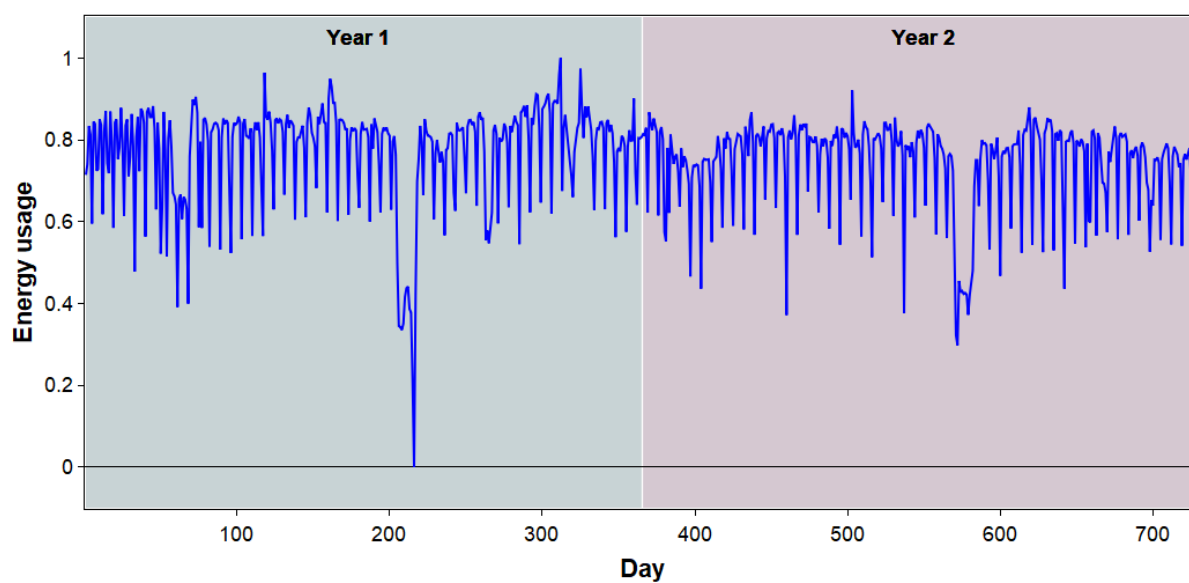


Figure C.10: Time-series plot of the energy usage for case study V

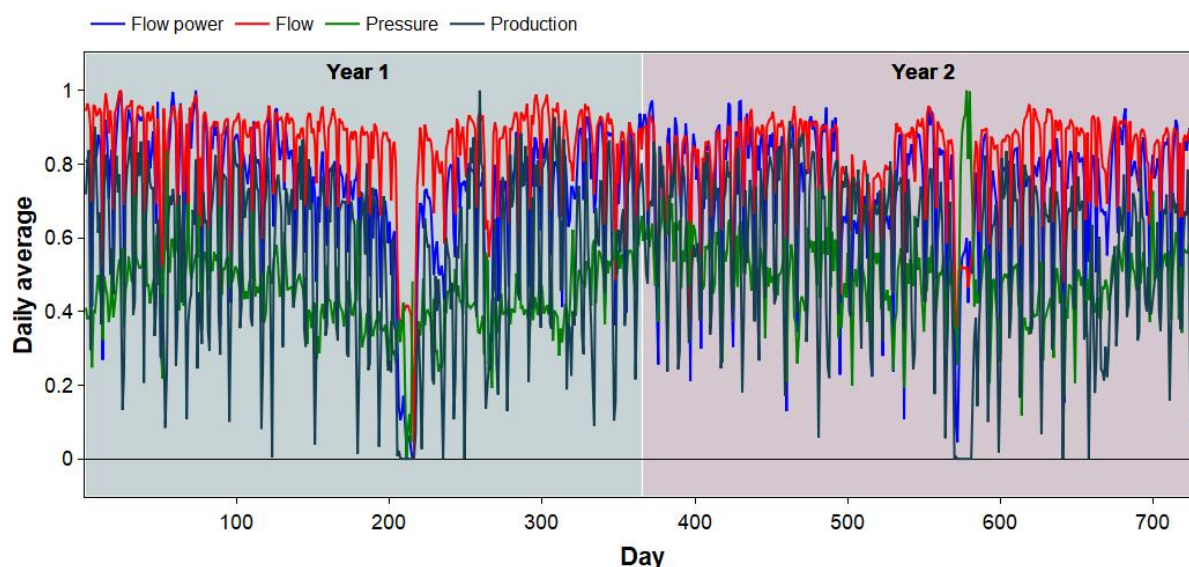


Figure C.11: Time-series plot of the potential drivers for case study V

The respective scatter plots of the energy usage vs the different potential drivers are shown in Figure C.12. There is a strong linear relationship between the energy usage and the flow. However, there are weak linear relationships between the energy usage and the flow power and the production. There is almost no relationship between the energy usage and pressure.

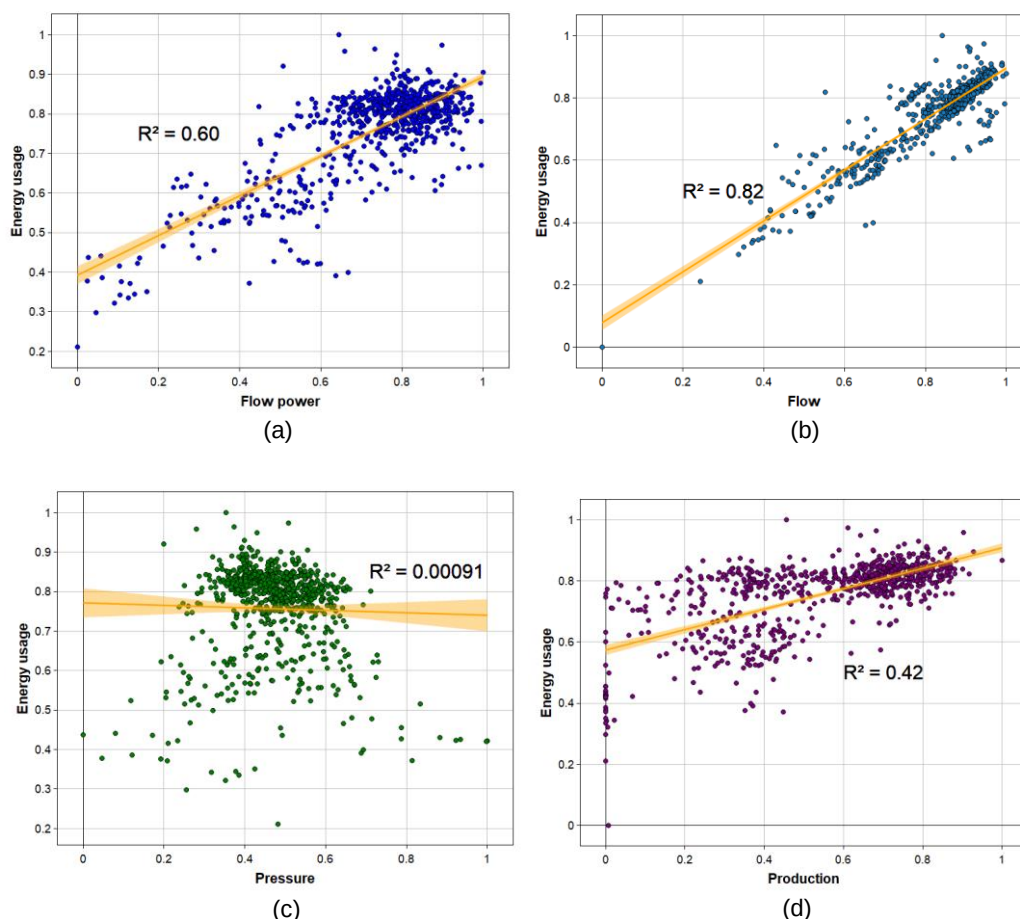


Figure C.12: Scatterplots of the energy produced vs (a) the reductant usage; (b) the ferrochrome function; (c) the ore usage; and (d) flux usage for case study III

Appendix C2 – DSOD parameters and outlier inspection plots

As mentioned in section 3.3, the DSOD method makes use of four user defined parameters. The values for these parameters were chosen based on values used by Zhang *et al* [126]. The different parameters as declared by Zhang *et al.* are given in Table C.1 along with the chosen values [126].

Table C.1: The parameters for the DSOD method with the chosen values

Parameter	Description	Chosen value
g	The search length for the close-neighbour-distance	2
k	The search length for the far-neighbour-distance	6
θ	The set threshold for the backward outlier factor and the forward outlier factor	2
ϕ	The set threshold for the dynamic sequential outlier factor	2

The process described in section 2.3 and demonstrated in section 3.3 was applied to each of the case studies. The same arguments apply in each case and the plots for each of the case studies are given hereafter in Figure C.13 to Figure C.22.

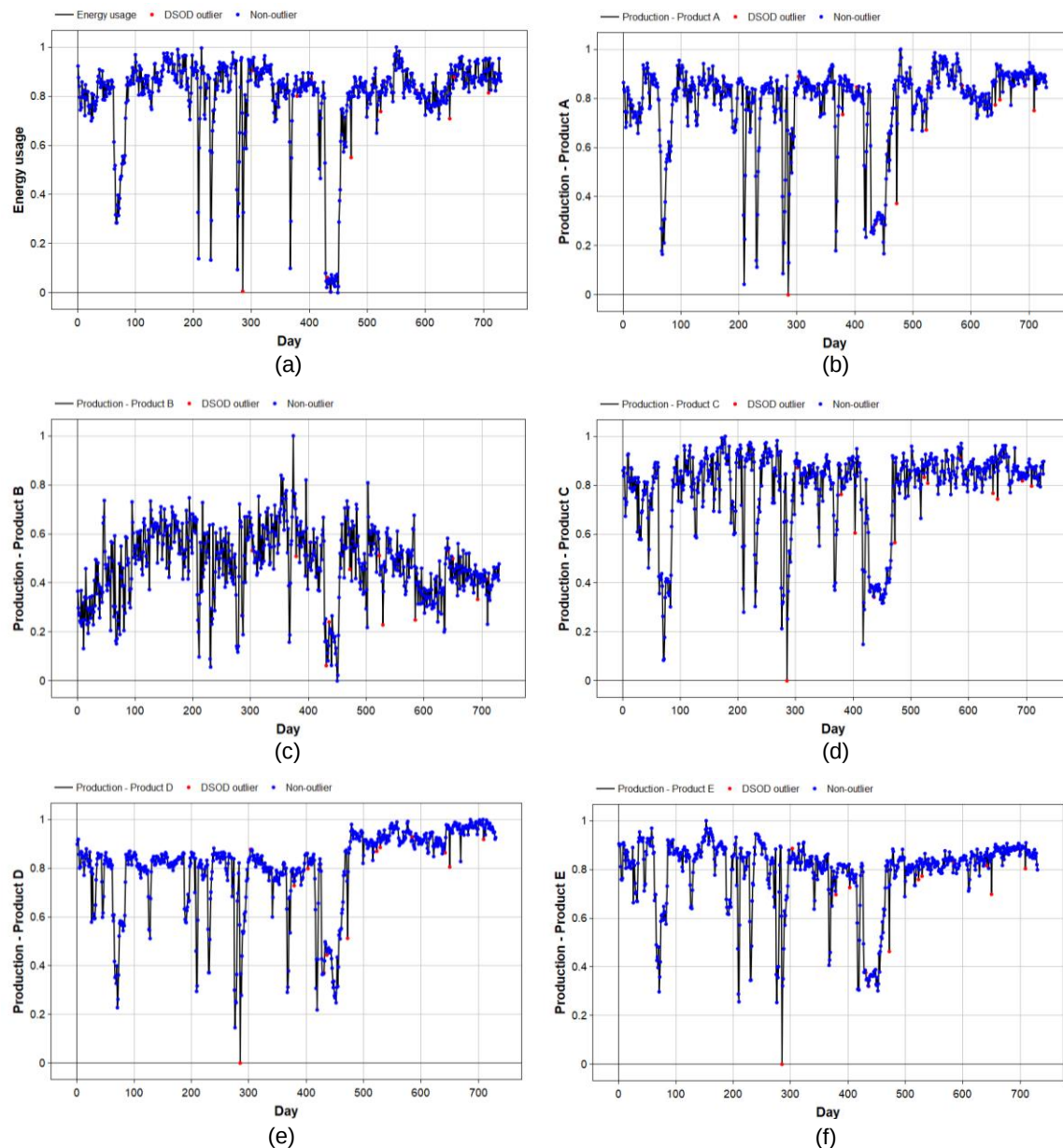


Figure C.13: Time-series plots of (a) the energy usage and (b)-(f) the products with the DSOD outliers for case study I

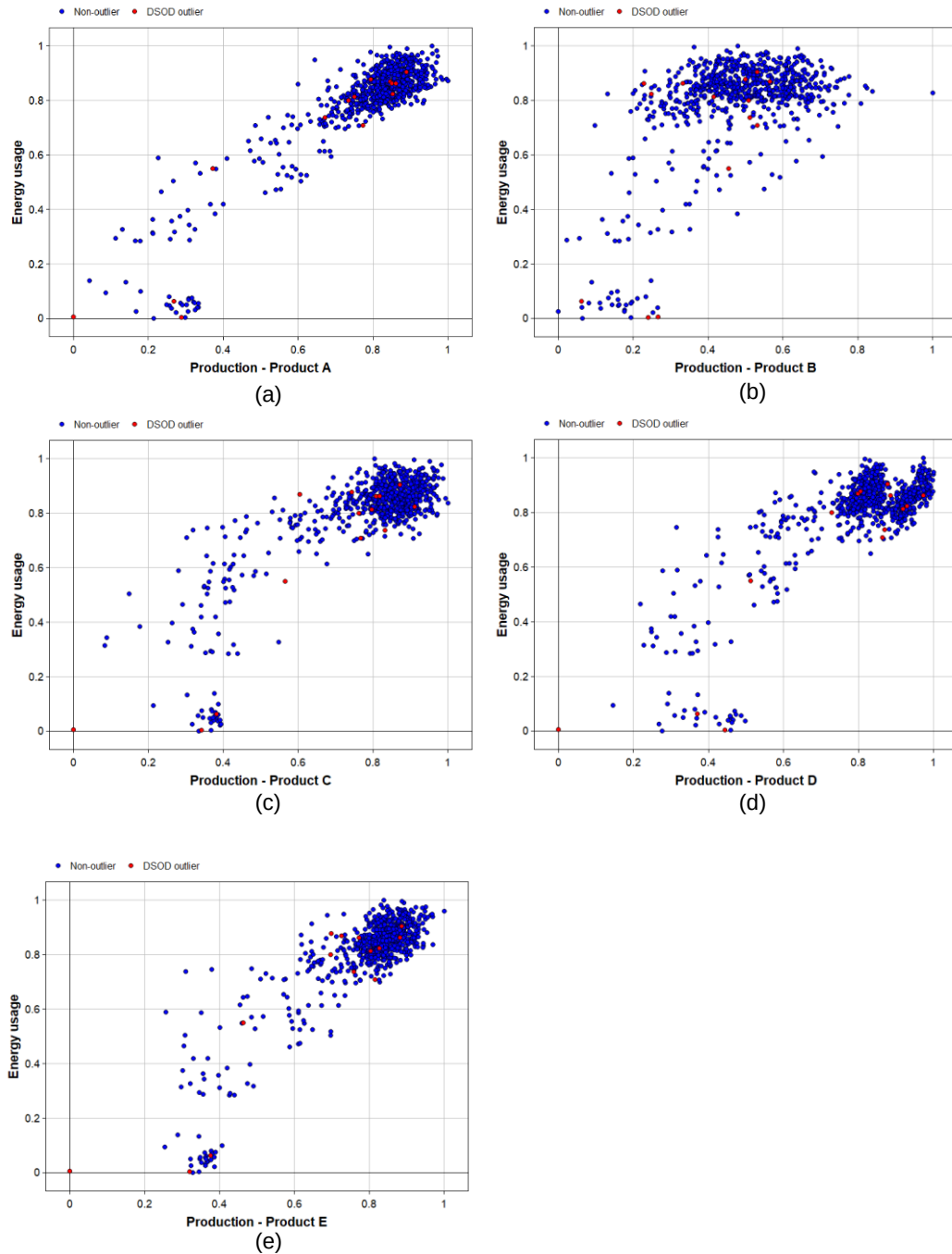


Figure C.14: Scatter plots of the energy usage against (a) product A; (b) product B, (c) product C; (d) product D; (e) product E for case study I

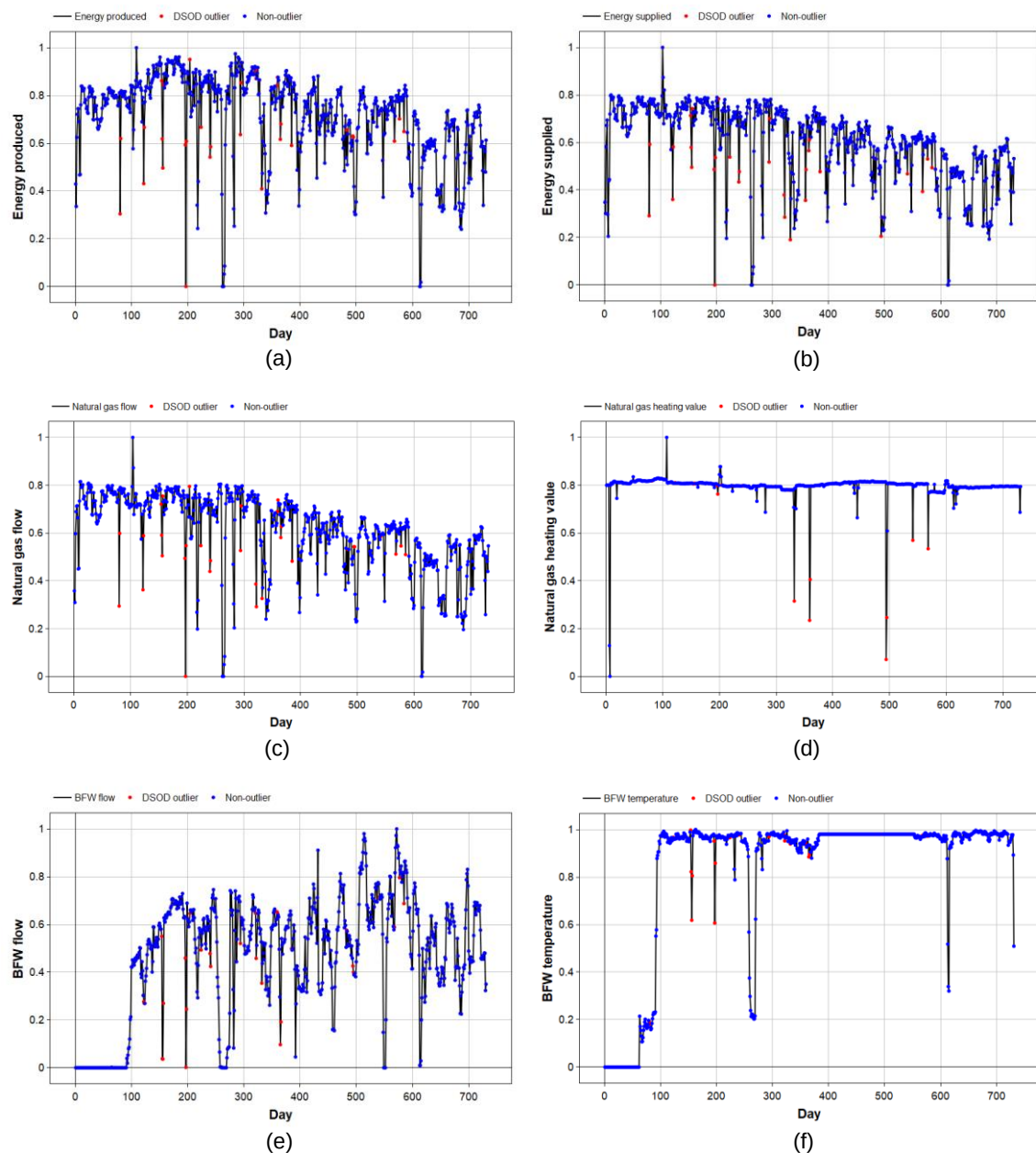


Figure C.15: Time-series plots of (a) the energy produced; (b) the energy supplied; (c) the natural gas flow; (d) the natural gas heating value; (e) the BFW flow and (f) the BFW temperature with the DSOD outliers for case study II

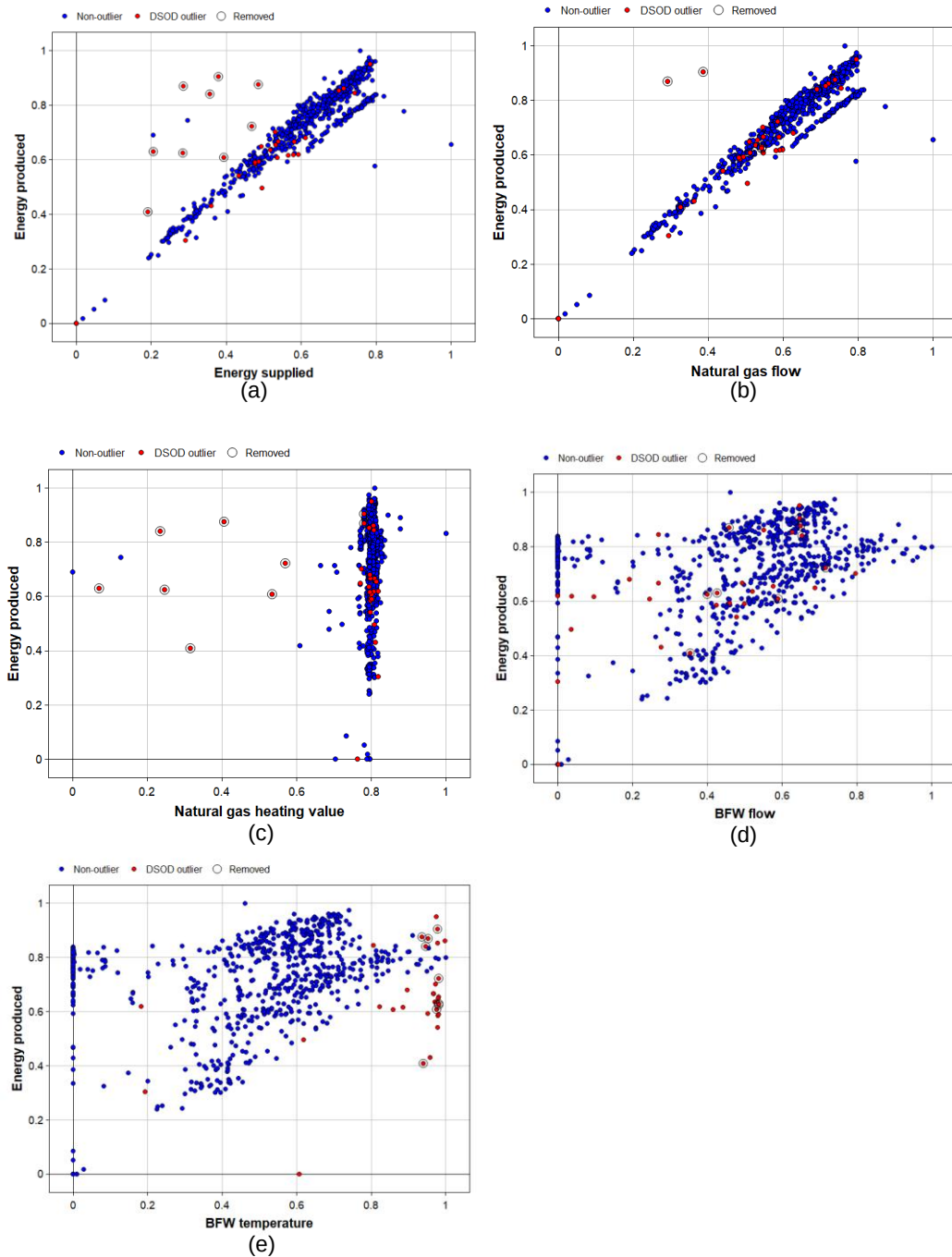


Figure C.16: Scatter plots of the energy produced against (a) the energy supplied; (b) natural gas flow; (c) the natural gas heating value; (d) BFW flow and (e) BFW temperature with the DSOD outliers and removed data points for case study II

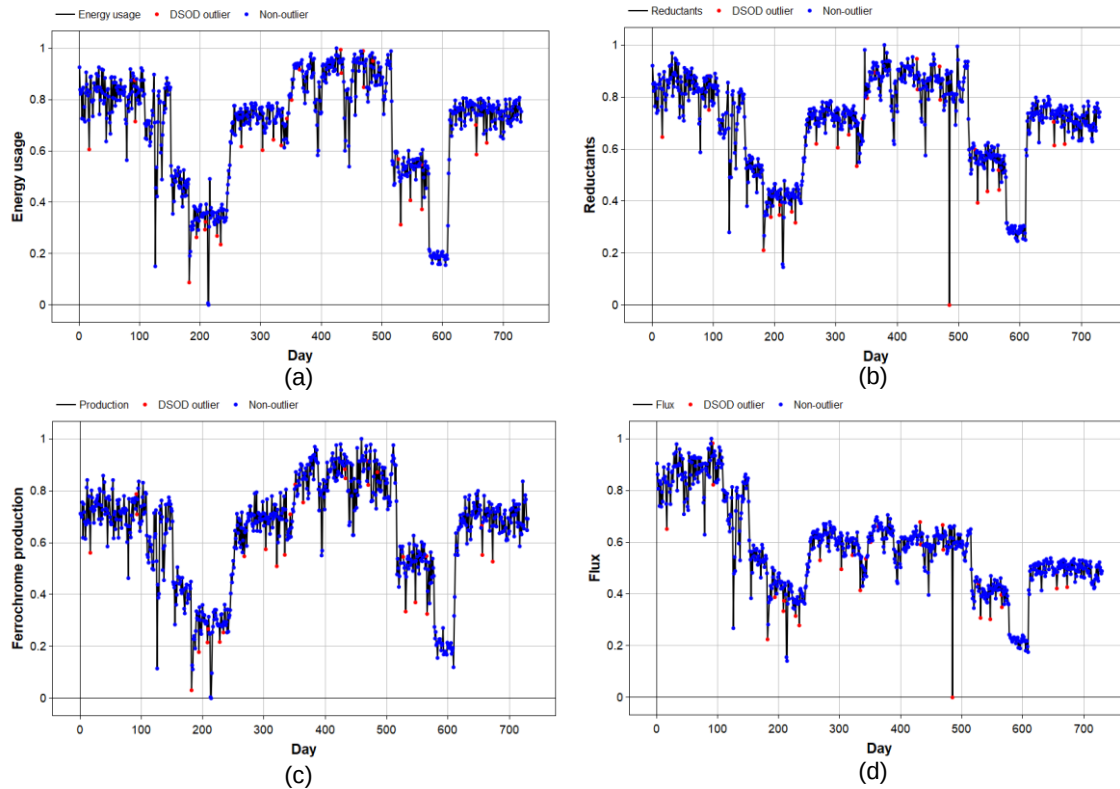


Figure C.17: Time-series plots of (a) the energy usage; (b) the reductants used; (c) the ferrochrome production; and (d) the flux used for case study III with the DSOD outliers

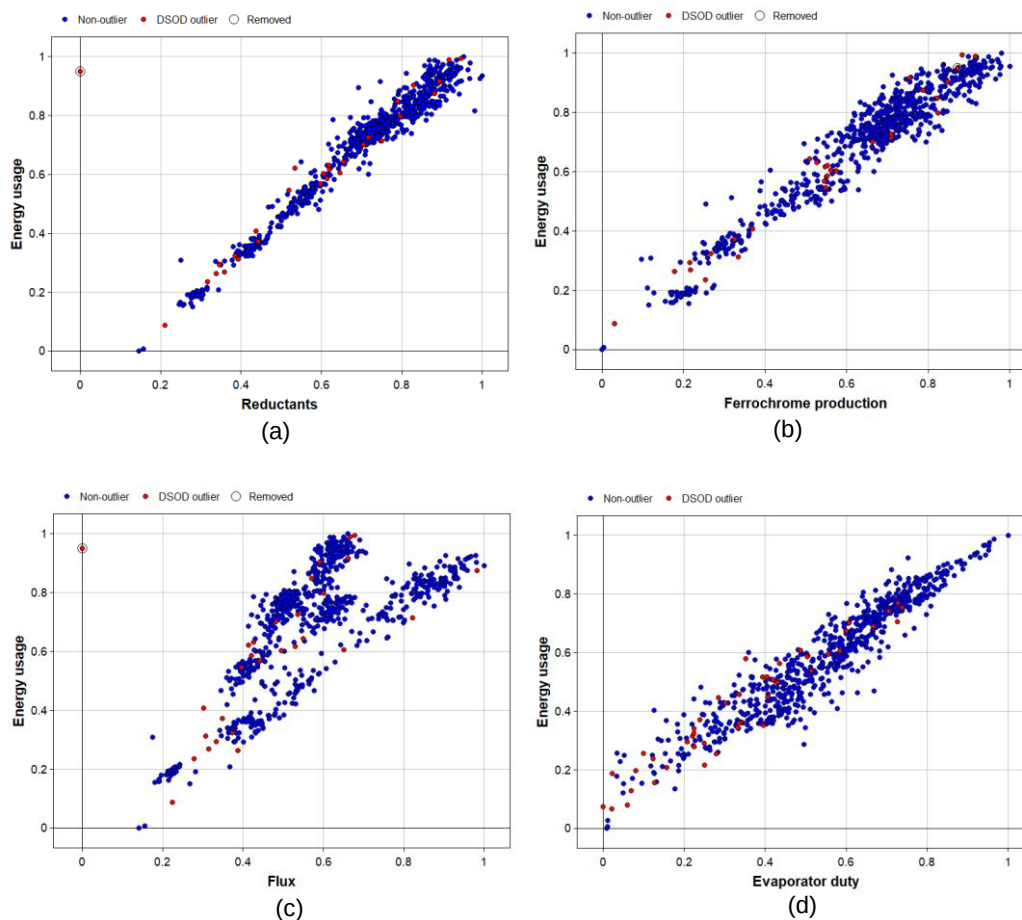


Figure C.18: Scatter plots of the energy usage against (a) the reductants used; (b) the ferrochrome production; (c) the flux used and (d) the evaporator duty with the DSOD outliers for case study III

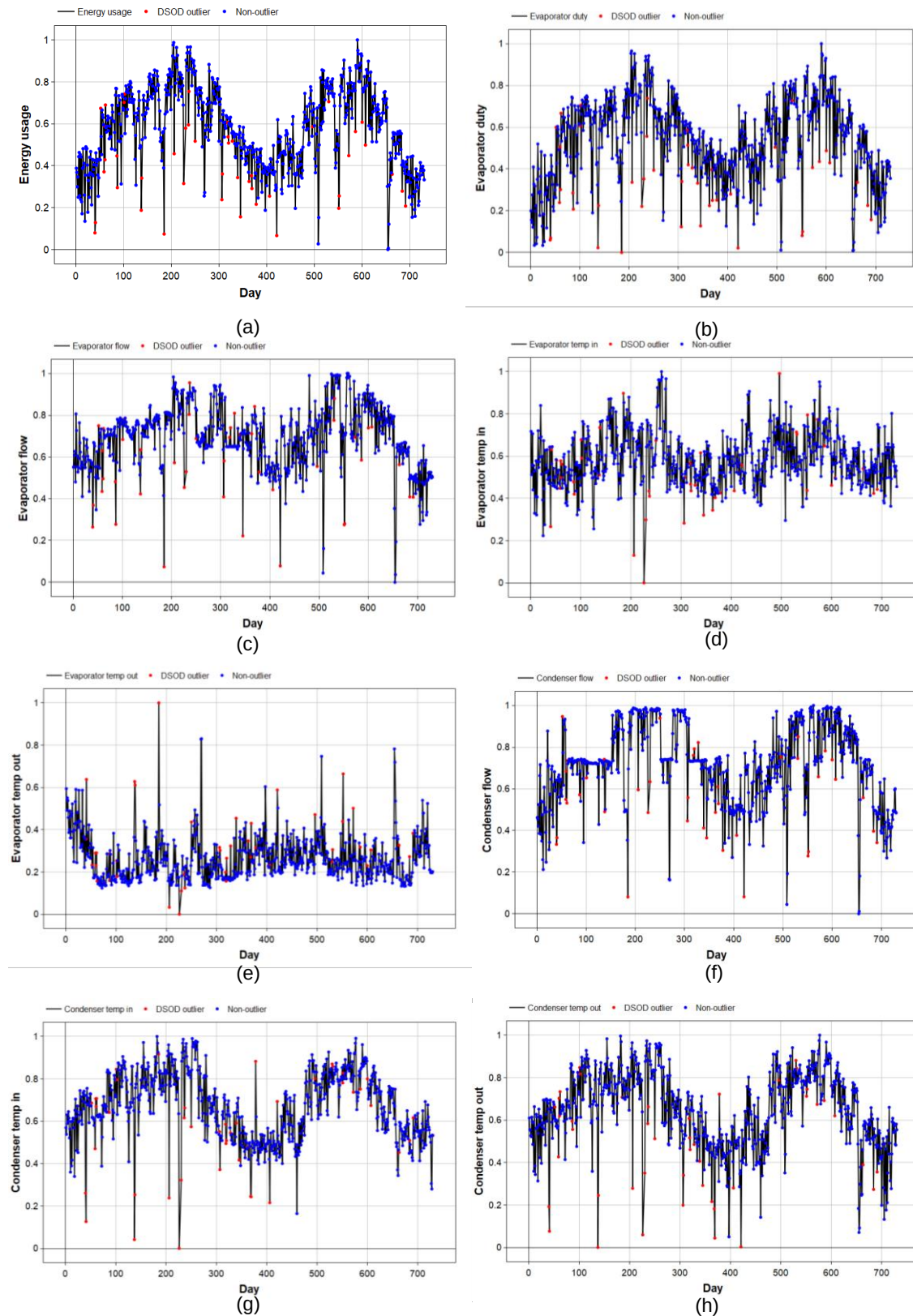


Figure C.19: Time-series plots of (a) the energy usage (b) the evaporator duty; (c) the evaporator flow; (d) the evaporator TI; (e) the evaporator TO; (f) the condenser flow; (g) the condenser TI; and (h) the condenser TO for case study IV

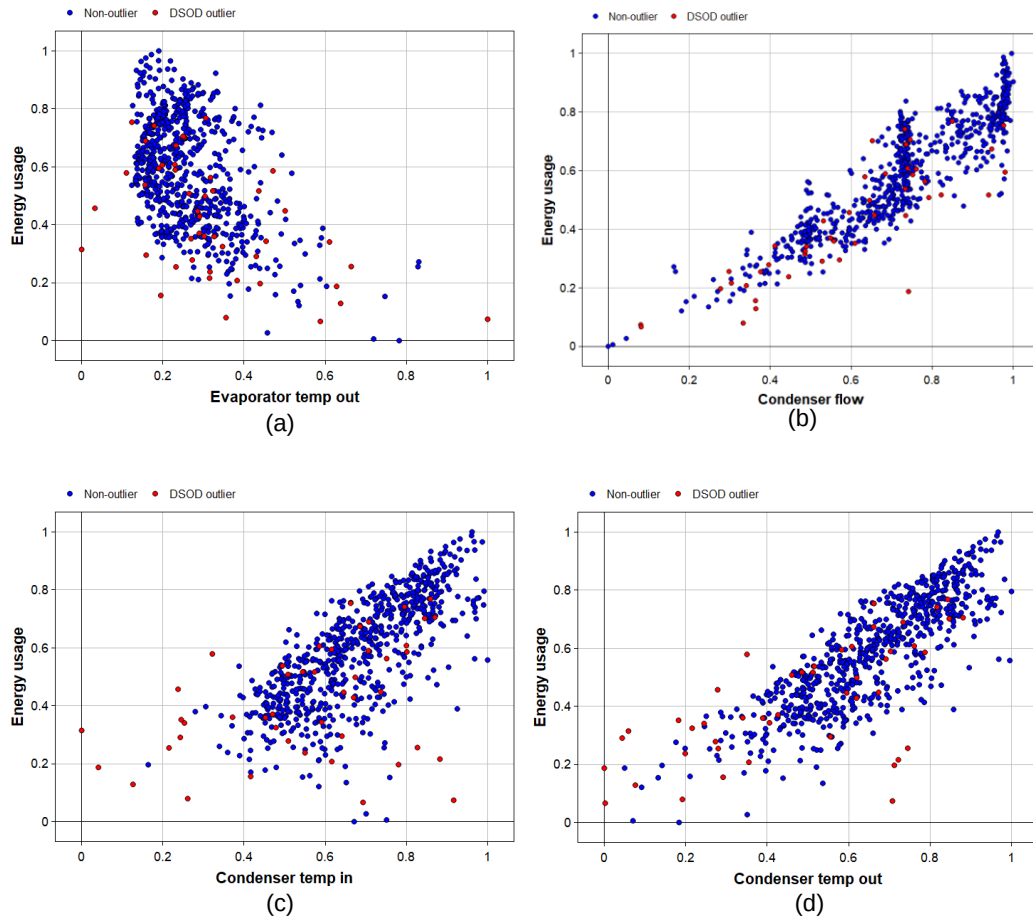


Figure C.20: Scatter plots of the energy usage against (a) the evaporator temperature out; (b) the condenser flow; (c) the condenser temperature in; and (d) the condenser temperature out for case study IV

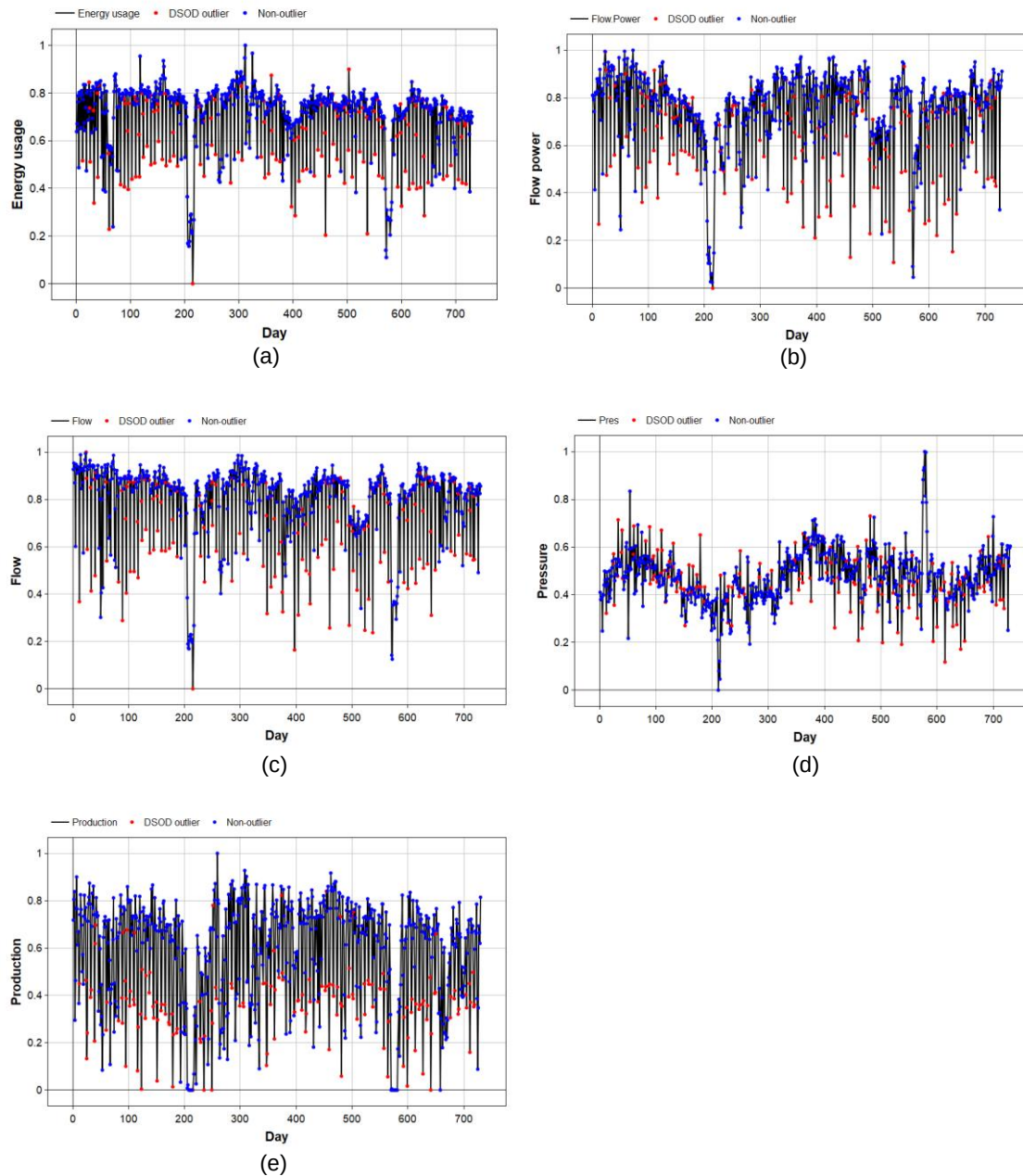


Figure C.21: Time-series plots of (a) the energy usage; (b) the flow power; (c) the compressed air flow; (d) the compressed air pressure; and (e) production for case study V

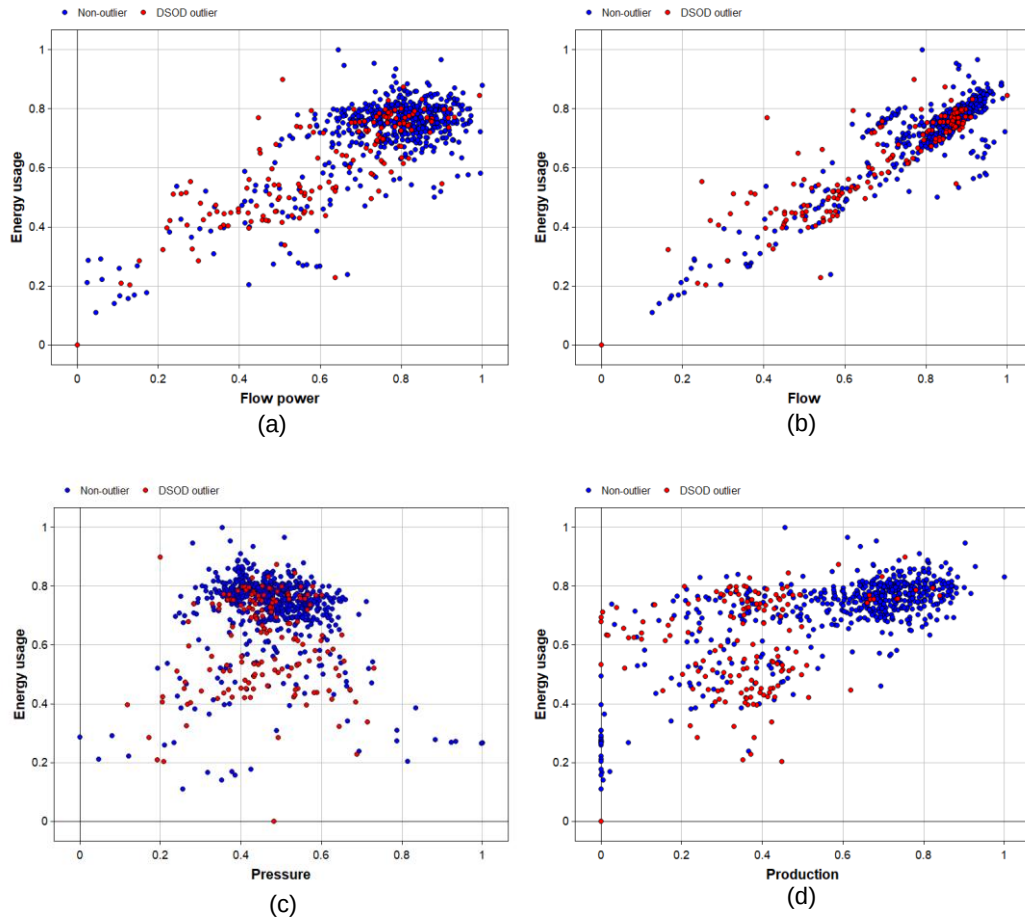


Figure C.22: Scatter plots of the energy usage against (a) flow power; (b) the flow; (c) the pressure and; (d) the production for case study V

Appendix C3 – Bar plots for the calculated savings

In this section, the bar plots of the savings calculated using the different approaches are given for each case study. A short discussion is given on each. As mentioned in Chapter 3, the same findings were generally apparent for each case study.

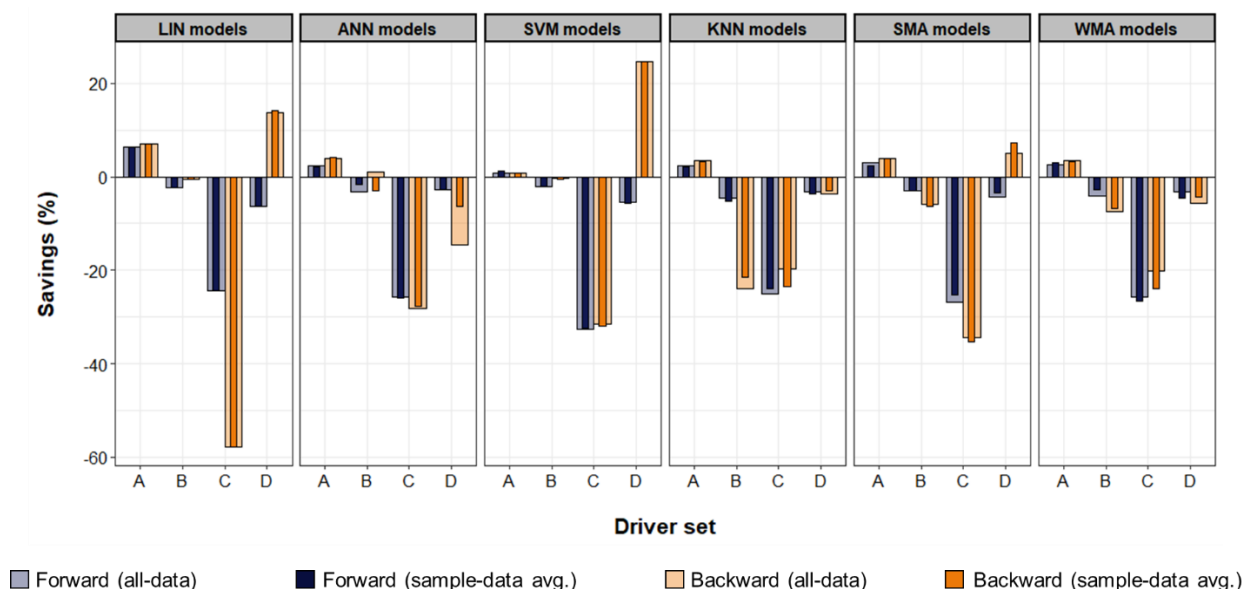


Figure C.23: The savings calculated for case study II

The bar plot of the savings calculated for case study II is given in Figure C.23. The saving ranges between -57.7% and 24.7%. The large variation in savings is due to the weak relationship with many potential drivers. These variables effectively add noise to the data which influence driver sets B to D greatly. Driver set A (the total energy supplied) is mainly uninfluenced by this as it is mostly determined by the NG flow.

In most of the cases there is a good consistency between the all-data and the sample-data approach. The greatest exception is the ANN model type for driver set B and D in the backward direction. The difference between the forward and backward-calculated savings varies from small to great and no model type is excluded from this, although the WMA model type is the least affected by it on average across the different driver sets.

The case study also demonstrates that differences are greater between model types than between different driver sets. Each of the model types shows smaller positive savings for driver set A with good consistency between the four different techniques. All the model types also show large negative savings for driver set C while showing relatively small positive or negative savings for driver set B. The greatest exceptions occur at driver set D where the linear-type models (LIN and SVM) show large positive savings for the backward calculation.

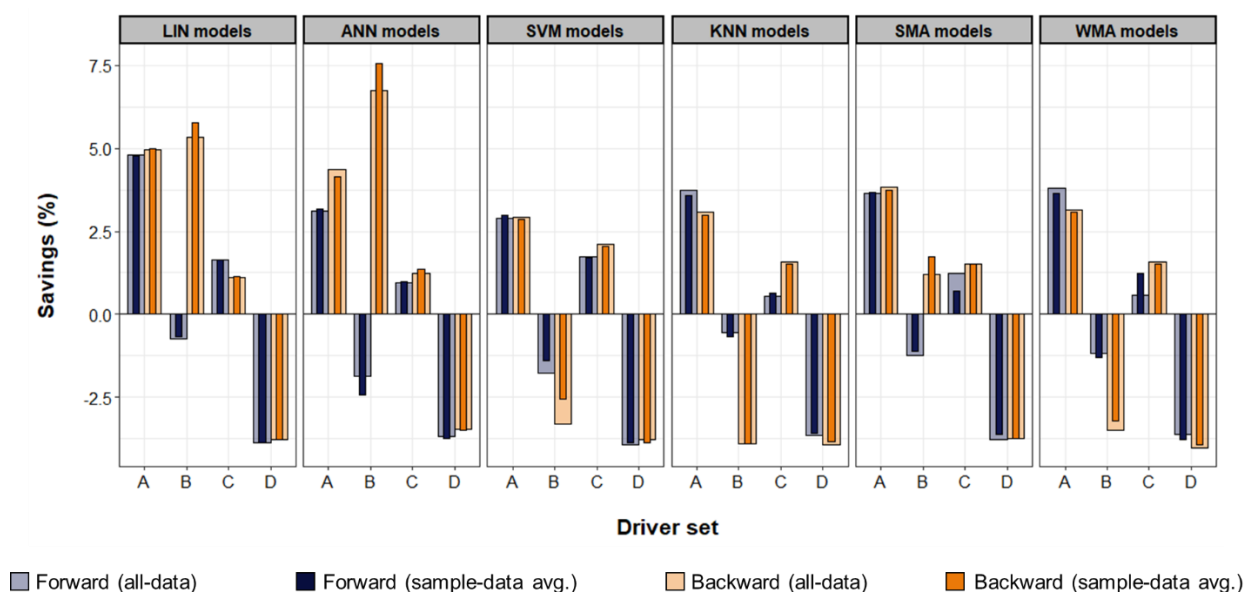


Figure C.24: The savings calculated for case study III

Figure C.24 gives the bar plot with the savings for case study III. The savings range between -4.0% and 7.6%. There is good consistency between the all-data and sample-data approach in most of the cases and the consistency between the forward and backward-calculated savings vary from small to great, although all the model types show good consistency between the forward and backward-direction when using driver set A or D. The sole exception to this is the ANN model type when using driver set A.

All the model types show positive savings of comparable size for driver set A but smaller savings for driver set C. Each of the model types show negative savings for driver set D of relatively the same size. As with case study I, the greatest exceptions occur when using driver set B. Here the LIN and ANN model types show relatively large positive savings in the backward direction while the SVM and KNN model types large negative savings in the backward direction. The differences are transferred to the model-averaging techniques.

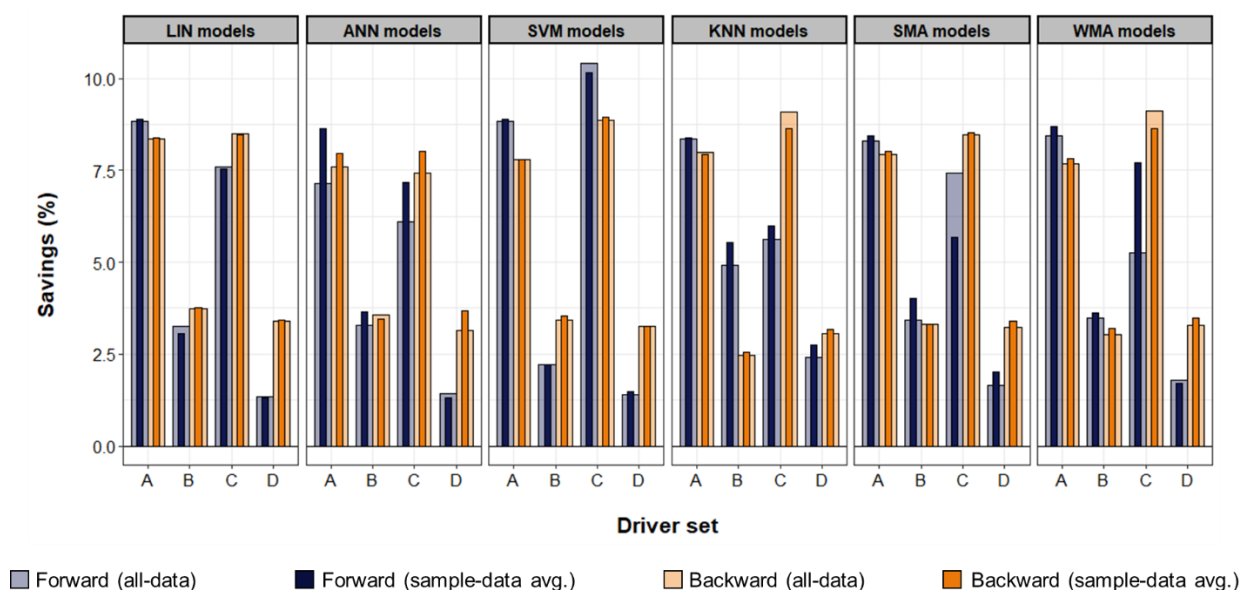


Figure C.25: The savings calculated for case study IV

The influence of the driver set is especially apparent for case study IV, as seen in Figure C.25. All the model types show positive savings for all the different calculation options. However, for each model type significantly larger savings are found when using driver set A or C in comparison to driver sets B and D. For each of the model types the saving for the backward-calculation is also greater than the forward-calculated saving when using driver set D.

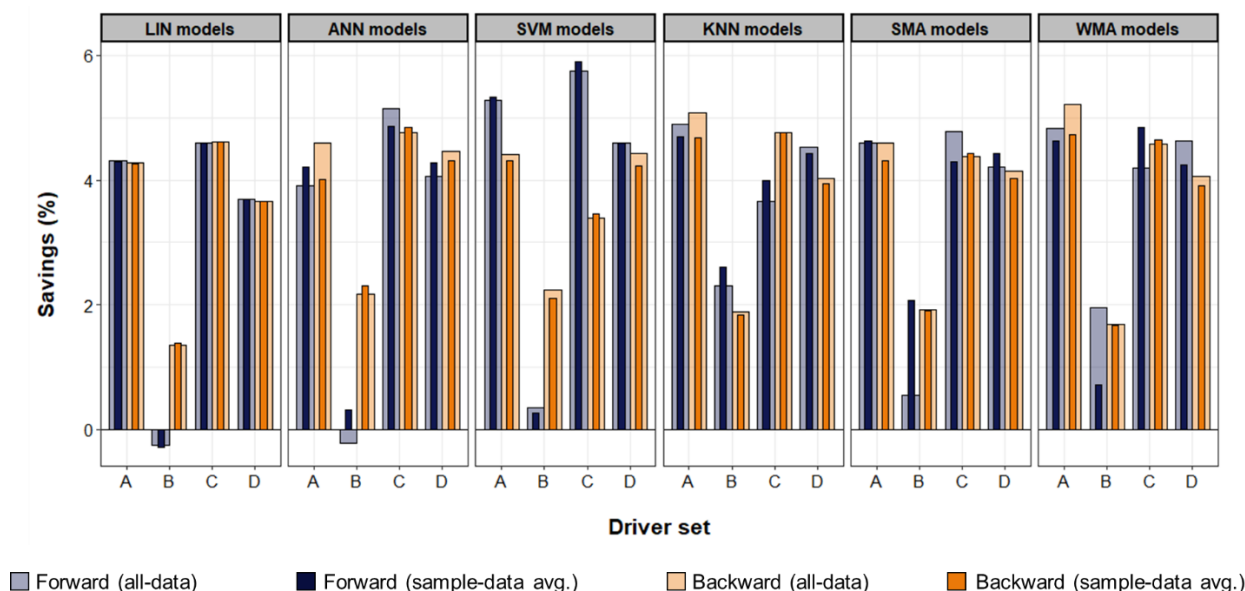


Figure C.26: The savings calculated for case study V

For case study V there is good consistency between savings calculated by the all-data and sample-data approaches, as with the other case studies (see Figure C.26). For each model type, the smallest savings occur when using driver set B. The greatest differences between the forward and backward-calculated savings occur when using driver set B for most of the model types with the backward-calculated savings being significantly larger. The sole exception to this is the KNN model type which consequently influences the model-averaging techniques. All the model types show larger savings when using driver sets A, C or D which are comparable in size.

Appendix C4 – Bar plots for the WMA weights

The bar plots with the weights to the different model types are given in Figure C.58 to Figure C.63 for the respective case studies. For case studies II to V, the greatest weights are assigned to the KNN model type for most of the options (see Figure C.27 to Figure C.32). However, for case studies II to IV there are certain exceptions that occur where greater weights are assigned to the ANN or SVM model types. For the weekly and monthly frequencies of case study I, the greatest weights were assigned to the LIN or ANN model types for most of the situations due to the poor performance of the KNN model type at these frequencies.

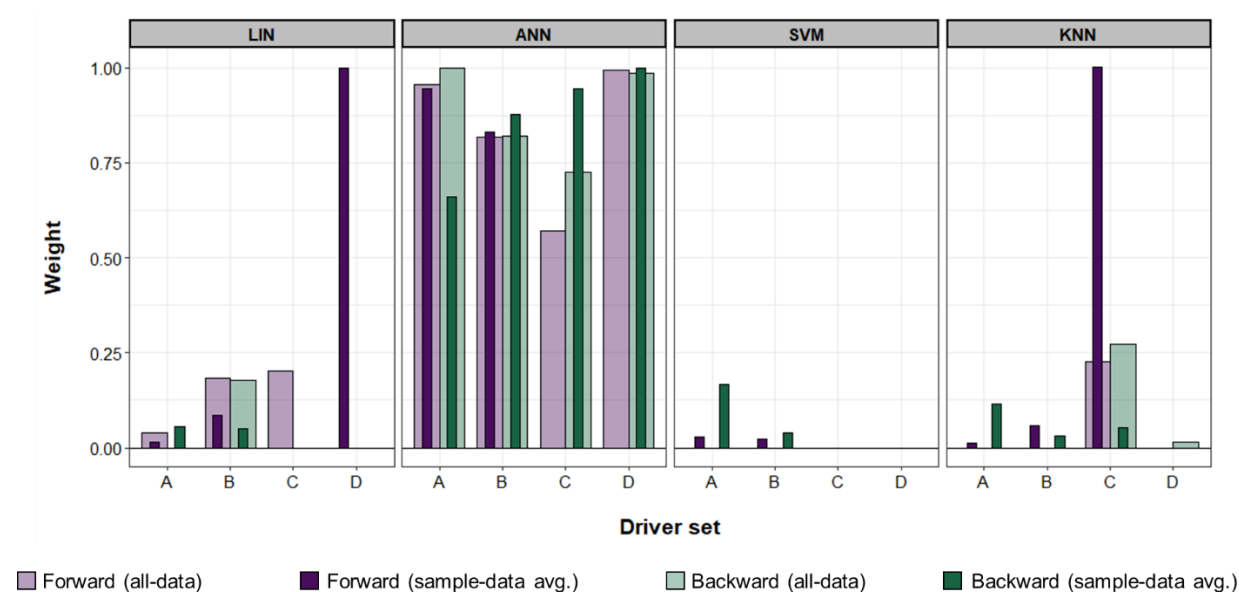


Figure C.27: The weights calculated for case study I (weekly)

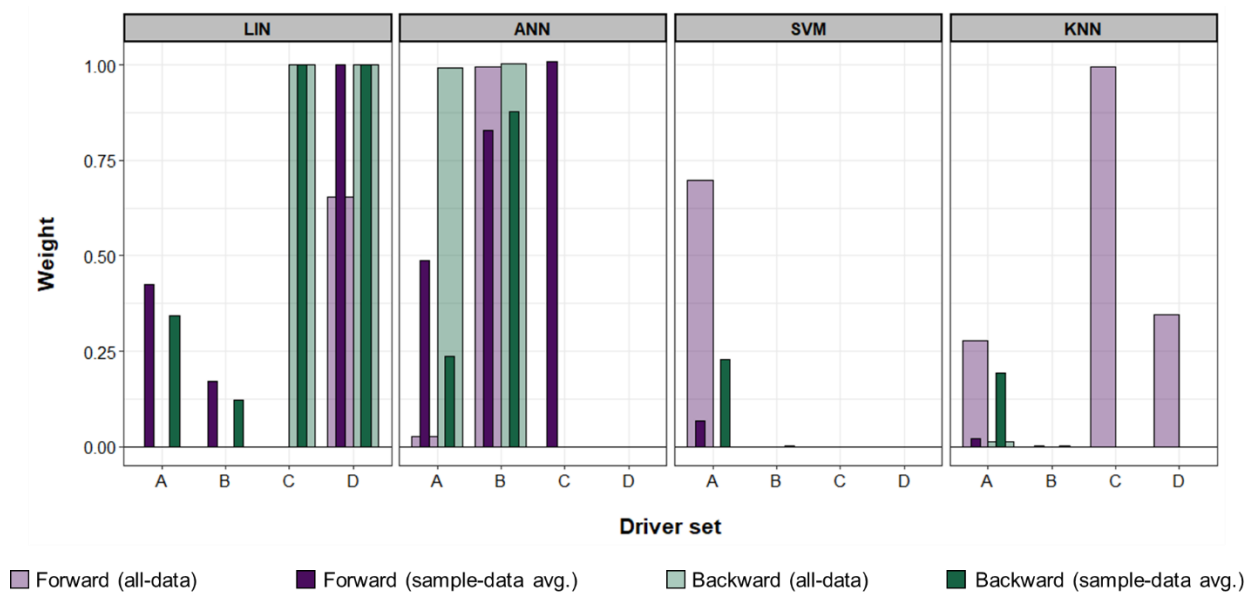


Figure C.28: The weights calculated for case study I (monthly)

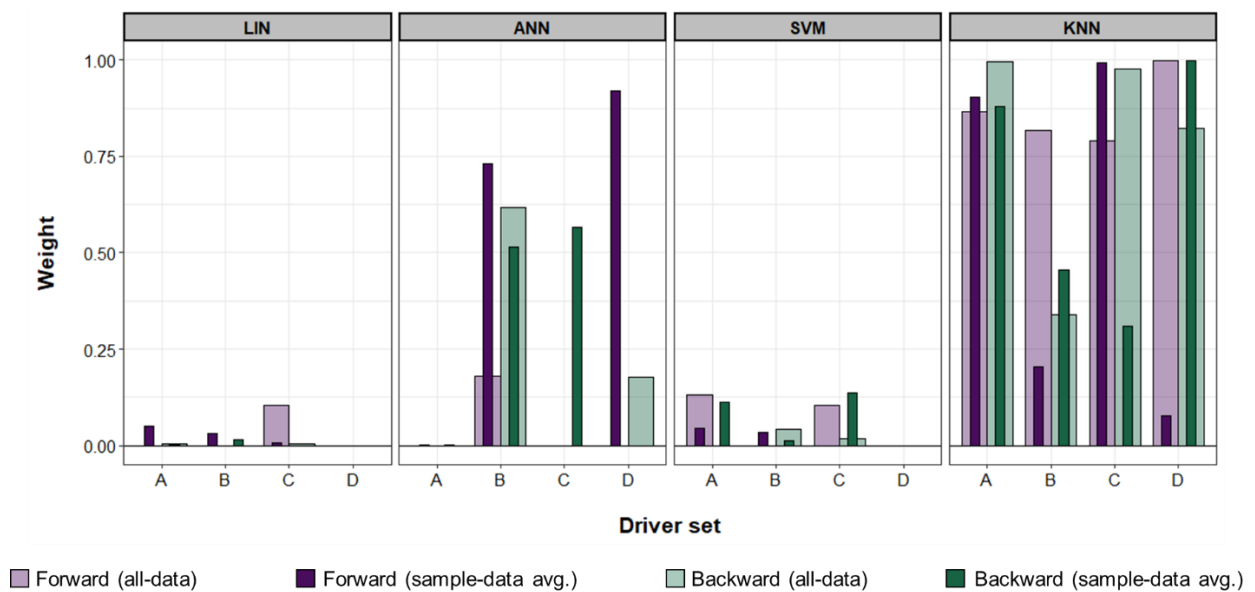


Figure C.29: The weights calculated for case study II

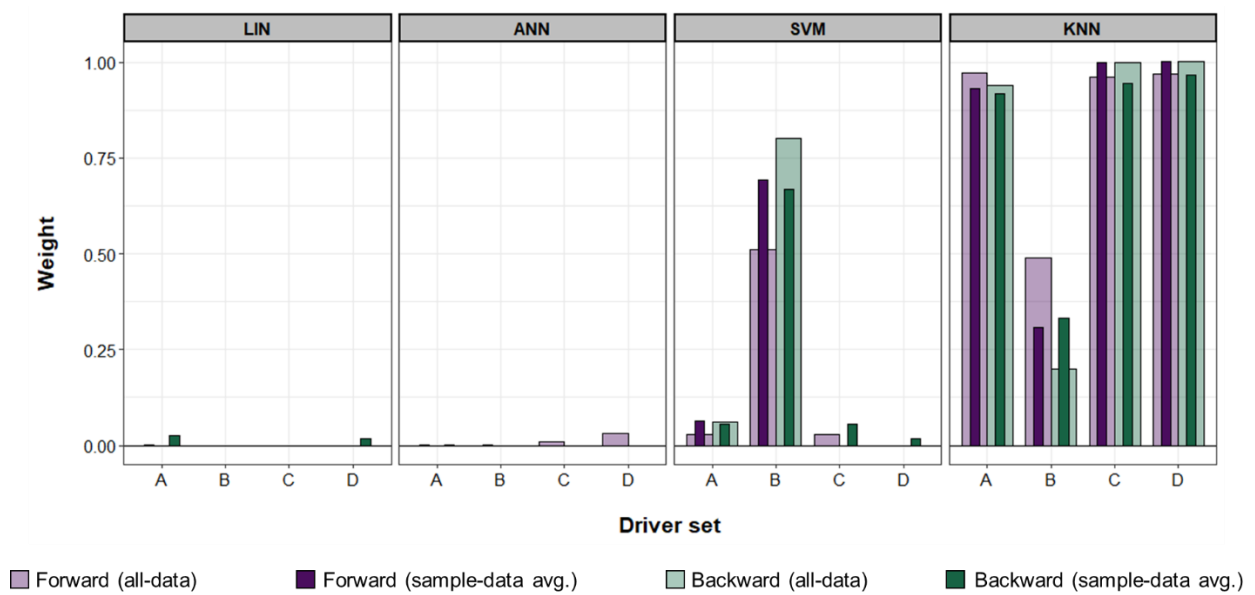


Figure C.30: The weights calculated for case study III

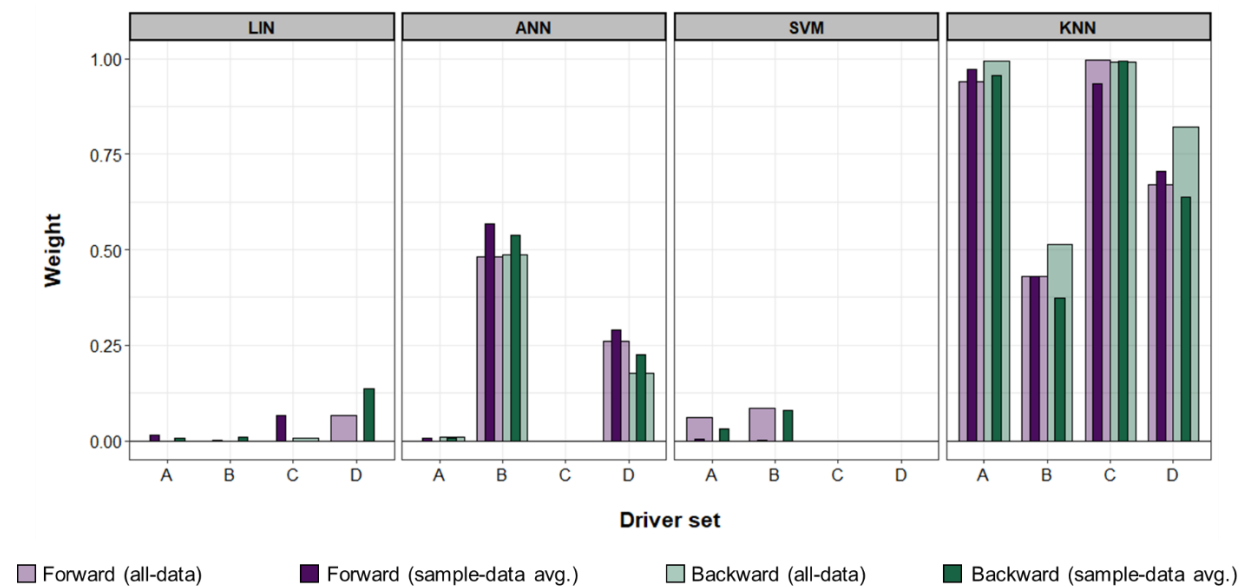


Figure C.31: The weights calculated for case study IV

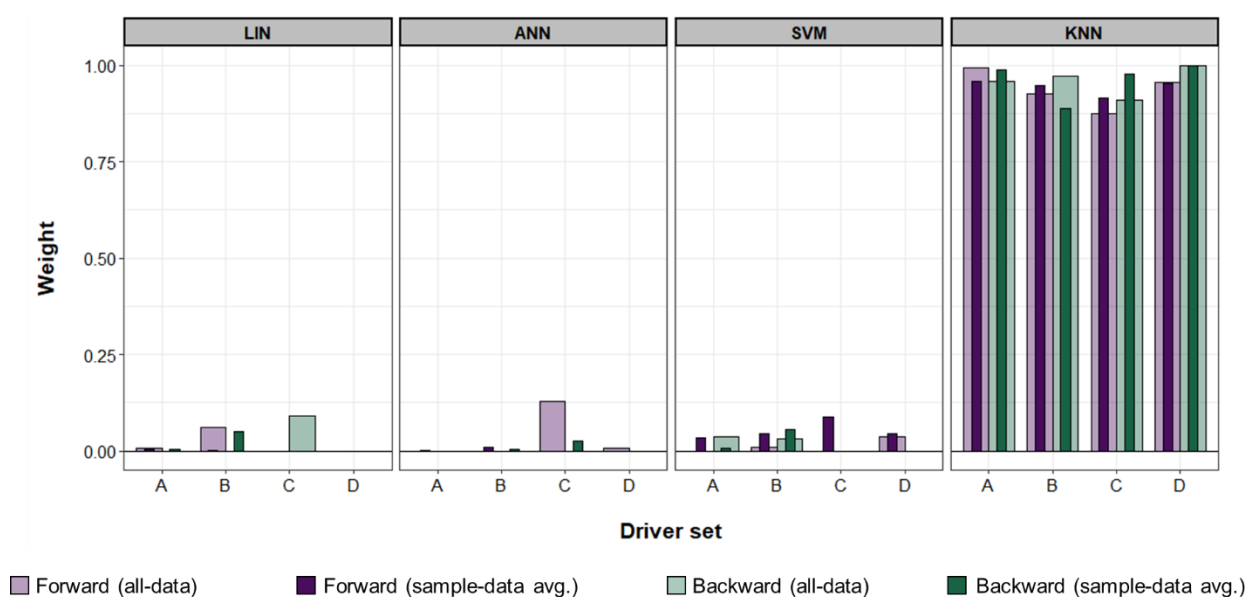


Figure C.32: The weights calculated for case study V

Appendix C5 – Savings density plots

The savings density plots for the different case studies are given in Figure C.33 to Figure C.36. The savings calculated for each case study have already been discussed in Appendix C3. Therefore, a more general discussion is presented here.

For most of the case studies, a split between the different driver sets is apparent. For case study II, driver sets C and A are split from driver sets B and D. The greatest differences occur here when using the LIN model with driver set A, the ANN model with driver sets B and D and the SVM model type with driver set C.

For case studies III and IV, the splits are relatively clear between the different driver sets with some variation between the different model types for a certain driver set. The greatest exception is the SVM model type at case study IV when using driver set C. The widest distributions occur when using the SVM model type with driver set B at case study III and when using the ANN model type with driver set C.

For case study V, the savings for driver set B are significantly lower for all the model types. However, the savings split between driver sets A, C and D is weak. Yet ranking the average savings size in decreasing order, for most of the model types, is as follows: (1) C, (2) A, (3) D and (4) B.



Figure C.33: The savings density plot for case study II

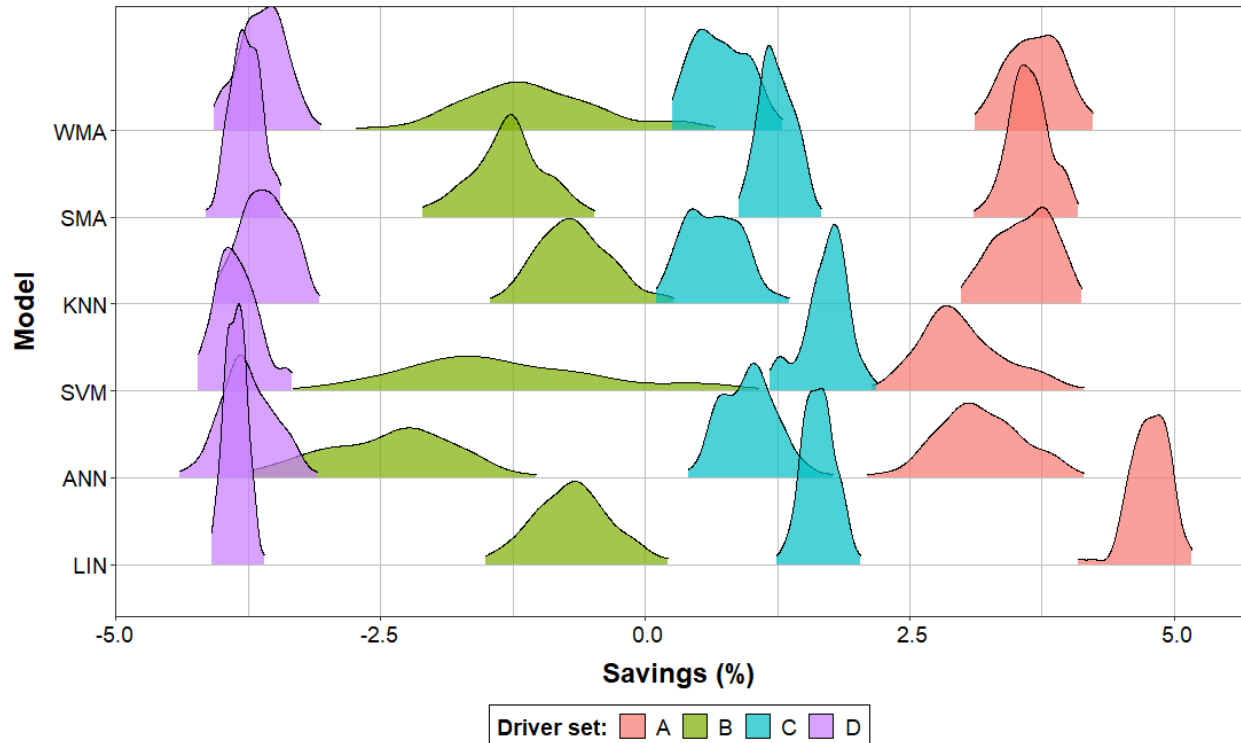


Figure C.34: The savings density plot for case study III

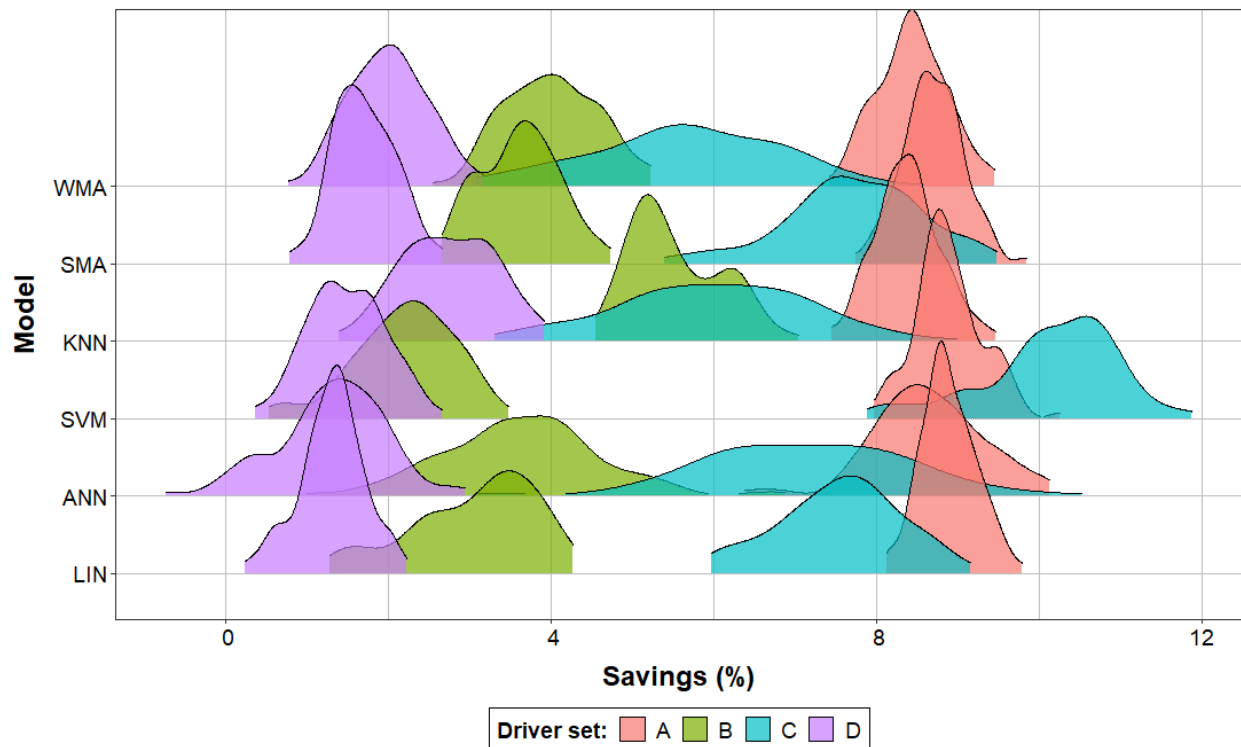


Figure C.35: The savings density plot for case study IV

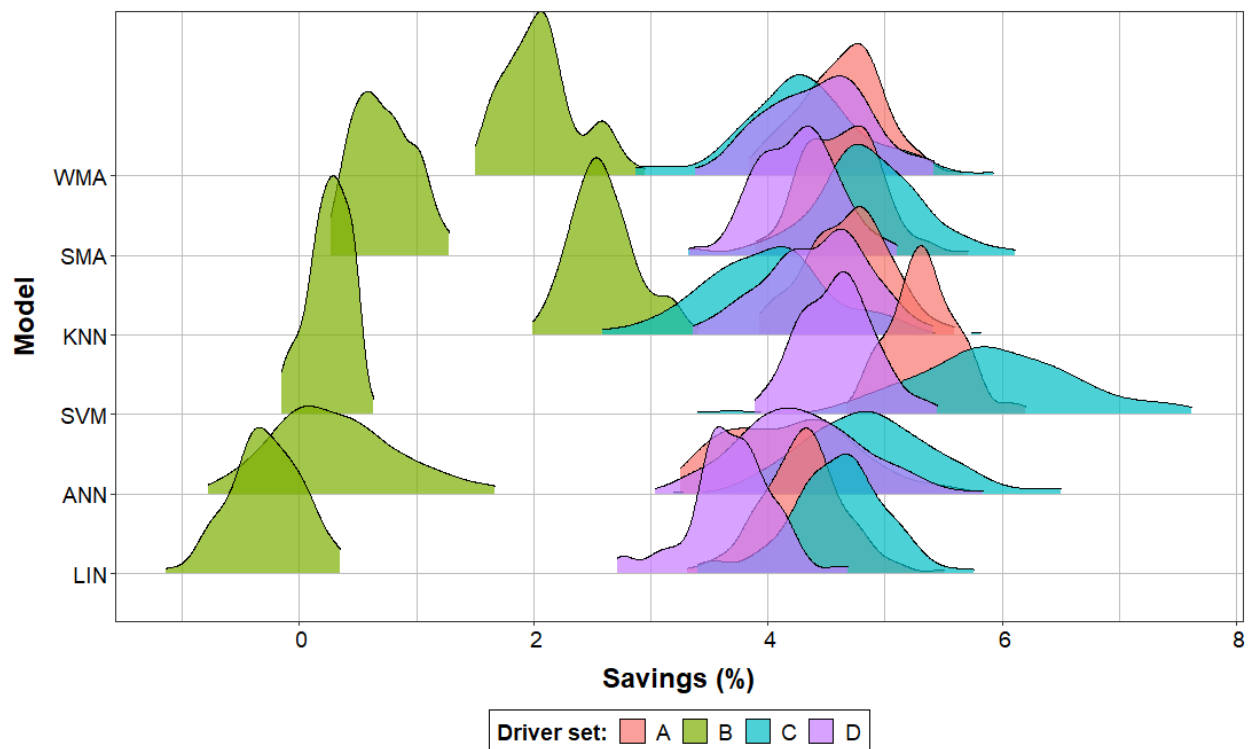


Figure C.36: The savings density plot for case study V

Appendix C6 – CV(RMSE) density plots

The density plots for the CV(RMSE) for the respective case studies that were not shown in Chapter 3 are given in Figure C.37 to Figure C.39. The splits between driver sets are less clear for driver set for case studies III and IV although driver sets A and C show significantly higher error values at the respective case studies. In each of the case studies, the ranking of average errors for the different driver sets is consistent between each of the model types. This confirms that there is an inherent absolute error that can be associated with each driver set.

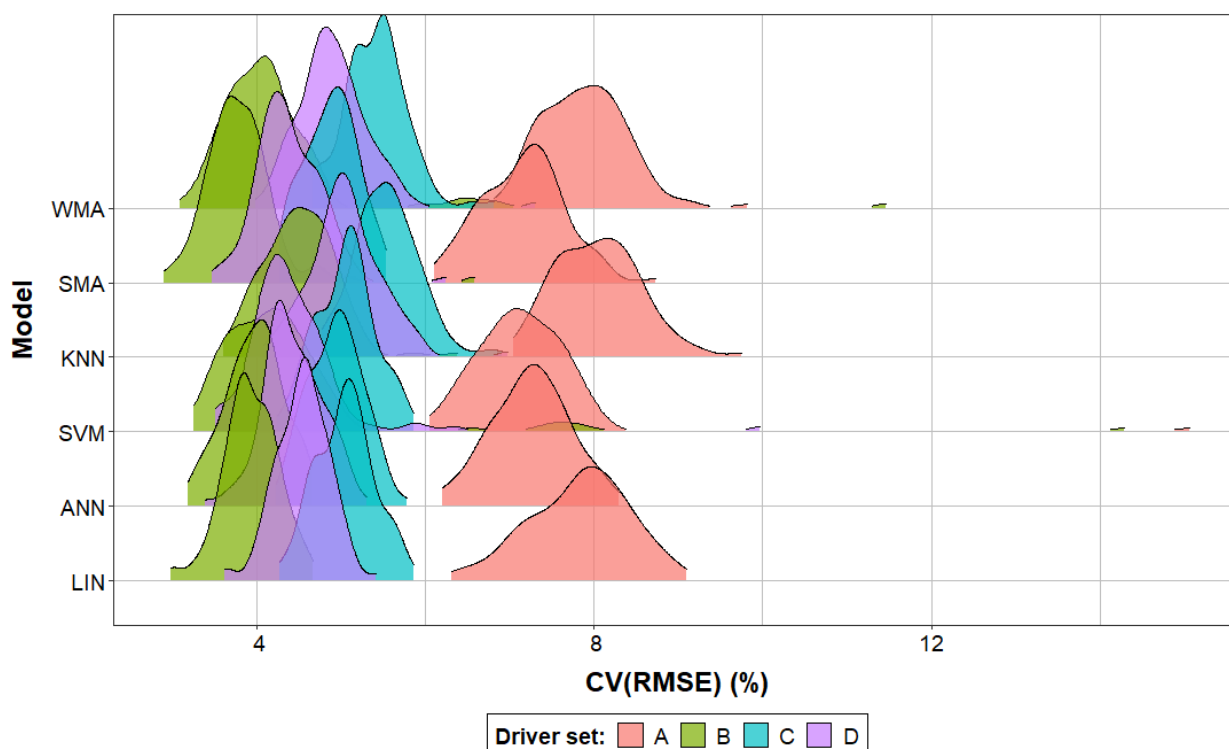


Figure C.37: The savings density plot of the CV(RMSE) for case study III

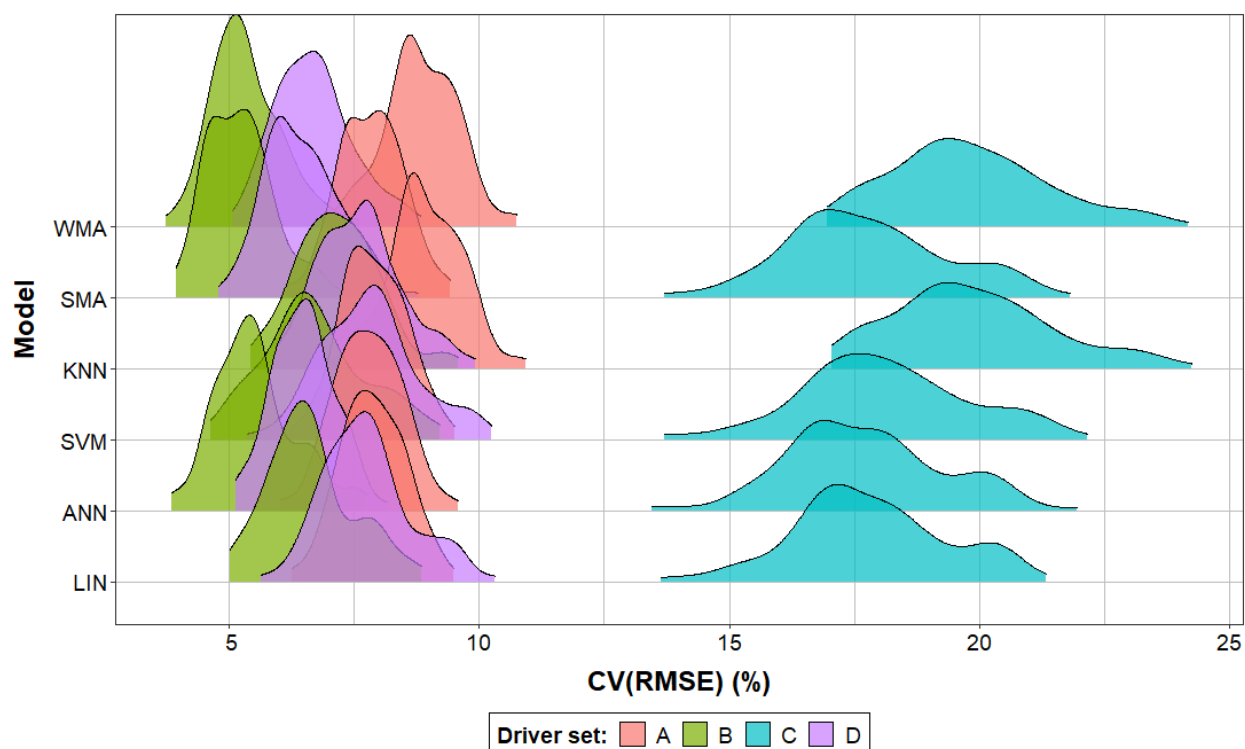


Figure C.38: The density plot of the CV(RMSE) for case study IV

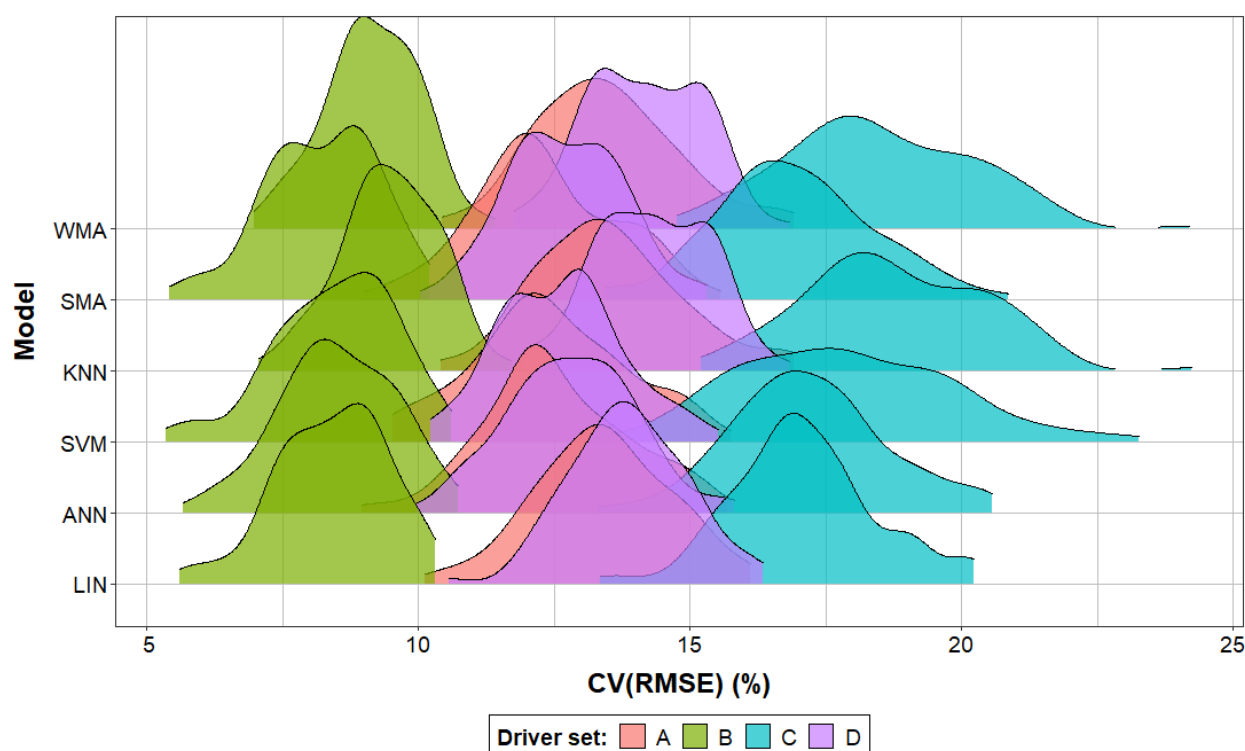


Figure C.39: The density plot of the CV(RMSE) for case study V

Appendix C7 – NMBE density plots

The density plots of the NMBE distributions for the respective case studies are shown in Figure C.40 to Figure C.43. In most of the cases close to normal distributions are formed with an average at zero, as with was the case when using the daily frequency for case study I. However, the widths of the distributions vary predominantly according to the driver set used.

The most noteworthy exceptions are:

Case study II (Figure C.40)

- (i) The LIN model type using driver set C;
- (ii) The SVM model using driver set C;
- (iii) Consequently, the SMA model type using driver set C.

Case study III (Figure C.41)

- (iv) The LIN model type when using driver set A;
- (v) Consequently, the SMA model type when using driver set A,
- (vi) The KNN model type using driver sets B and D.

Case study IV (Figure C.42)

- (vii) All the model types when using driver set C, except for the KNN model type and consequently the WMA model type;
- (viii) The SVM model type using driver set D;
- (ix) The KNN model type using driver sets B and D.

Case study V (Figure C.43)

- (x) The SVM model type when using driver sets A and C.

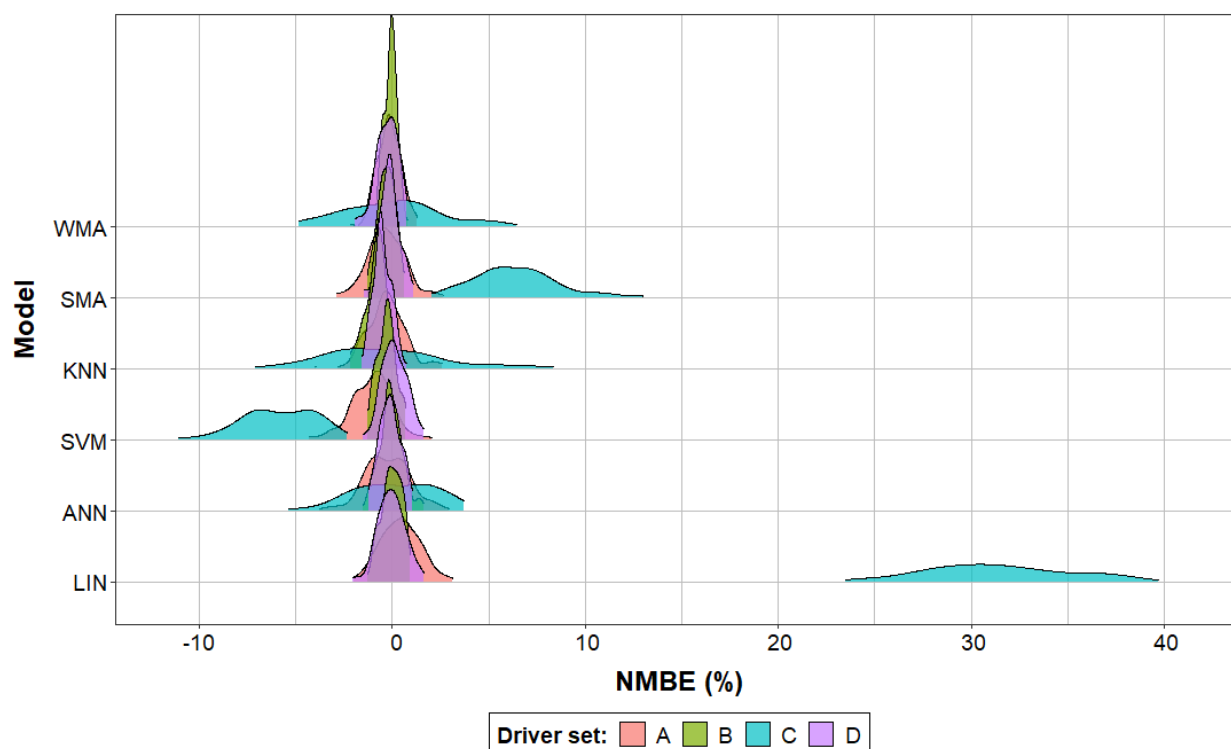


Figure C.40: The density plot of the NMBE for case study II

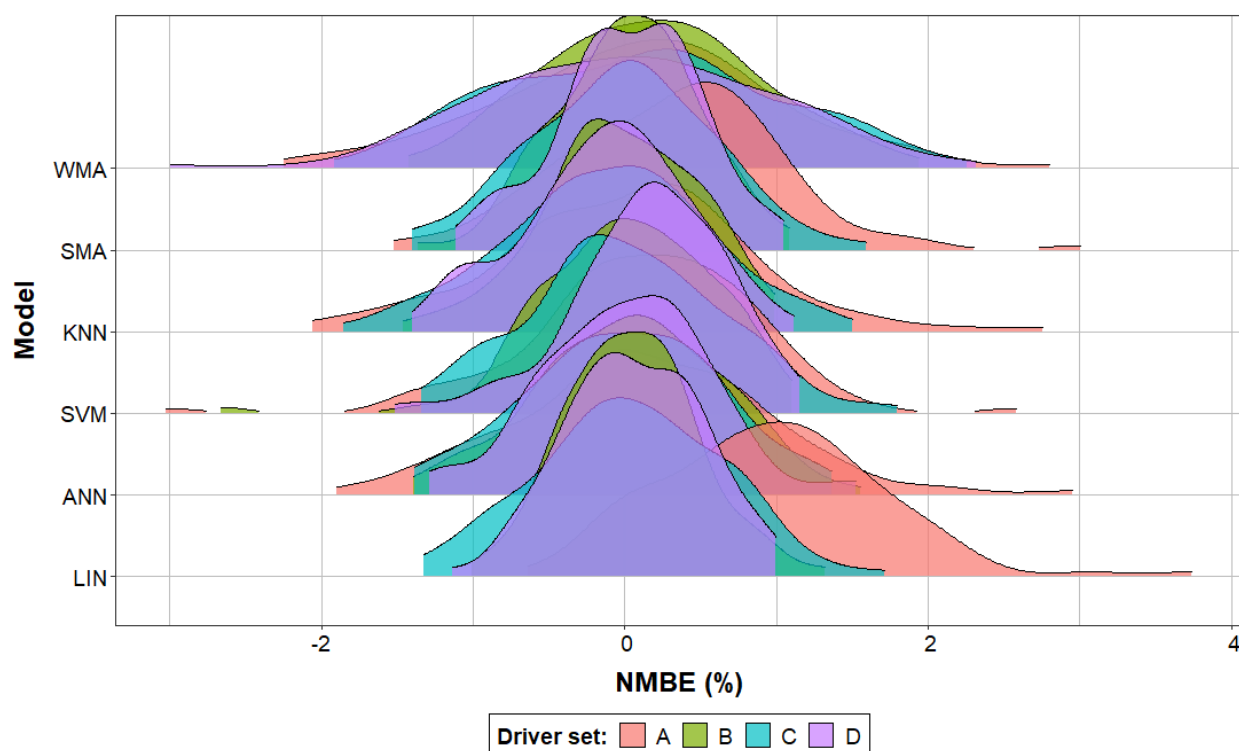


Figure C.41: The density plot of the NMBE for case study III

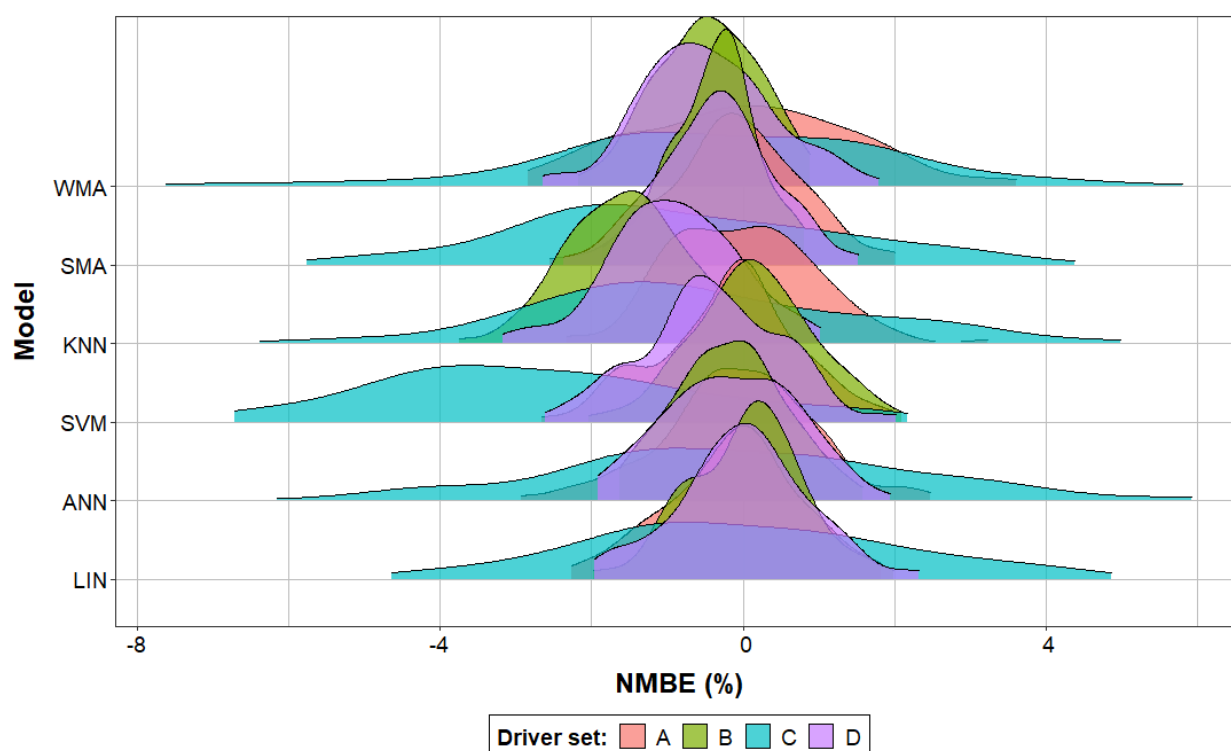


Figure C.42: The density plot of the NMBE for case study IV

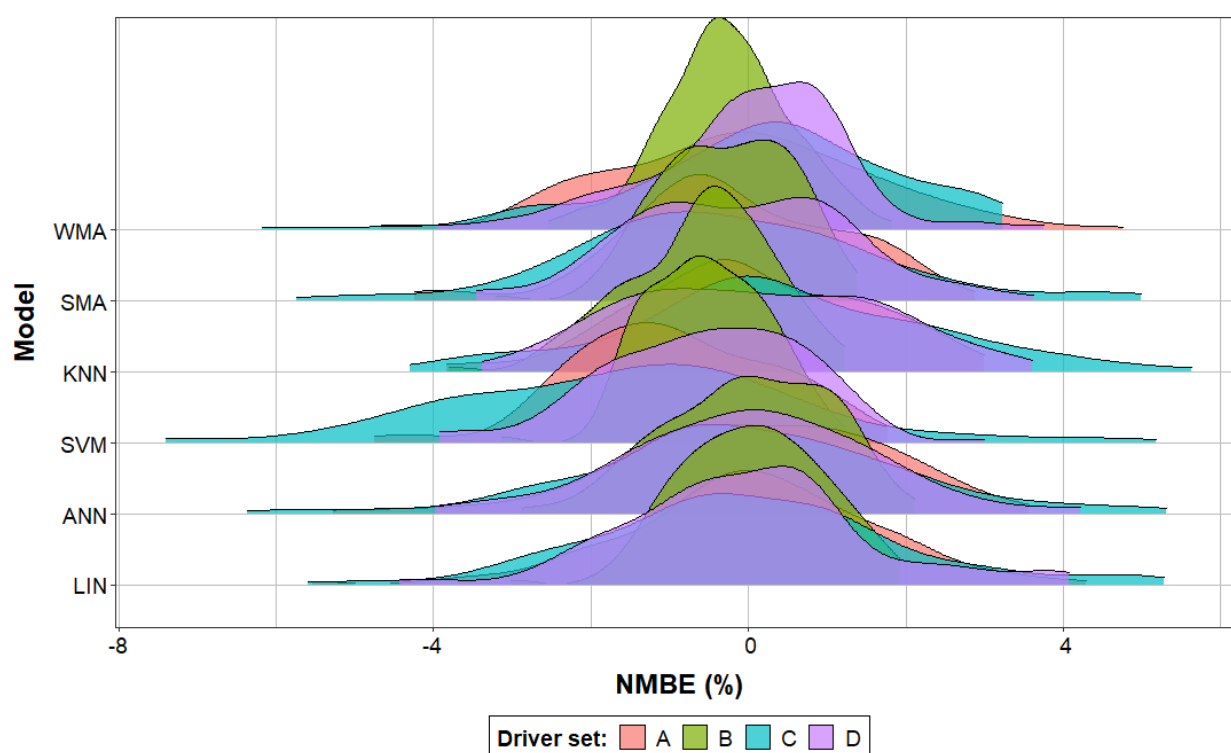


Figure C.43: The density plot of the NMBE for case study V

Appendix C8 – The error scatter plots

The scatter plots of the NMBE vs the CV(RMSE) for both the all-data approach and the sample-data approach for the respective case studies are shown in Figure C.44 to Figure C.55. The following general observations are made:

- (i) As with the daily frequency for case study I, the error values tend to group according to the different driver sets and especially so for the sample-data approach;
- (ii) As shown by the model-scoring, the SMA model type fall amongst the most accurate model types based on the CV(RMSE) when using the sample-data approach;
- (iii) As with the daily frequency for case study I, the KNN model type is amongst the most accurate model types when using the all-data approach, but amongst the least accurate model types when using the sample-data approach (based on the CV(RMSE)) values.

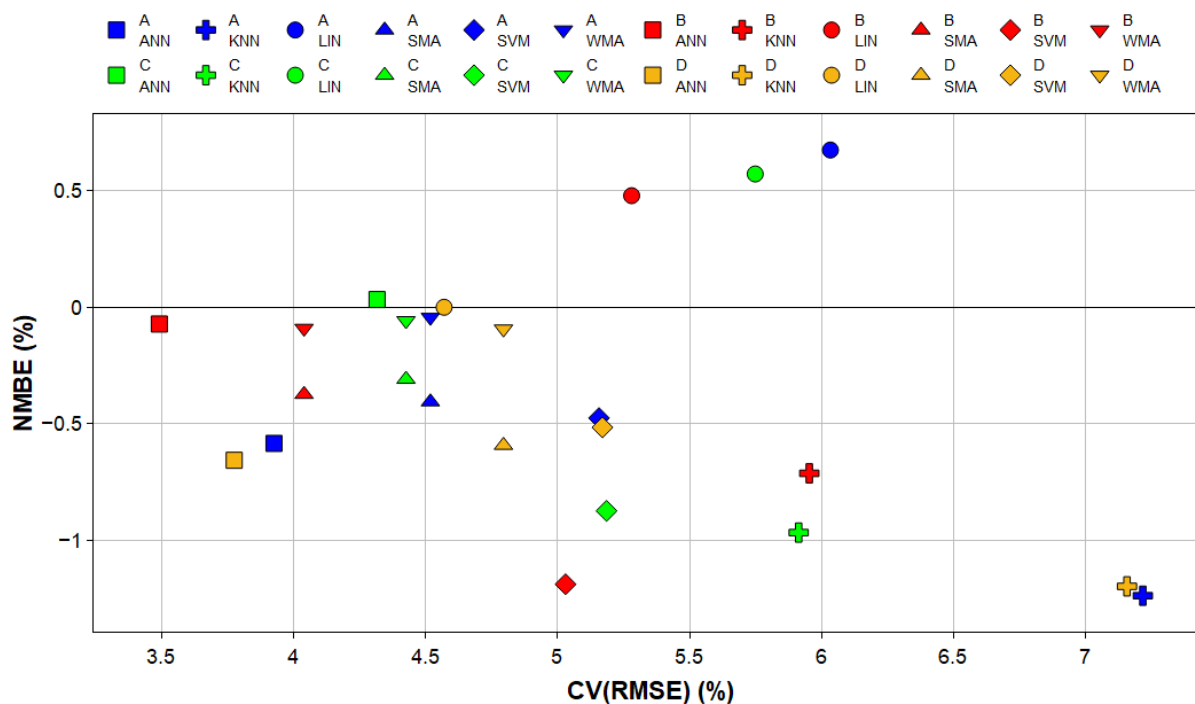


Figure C.44: NMBE vs CV(RMSE) for the forward models when using all the training data for the weekly frequency of case study I

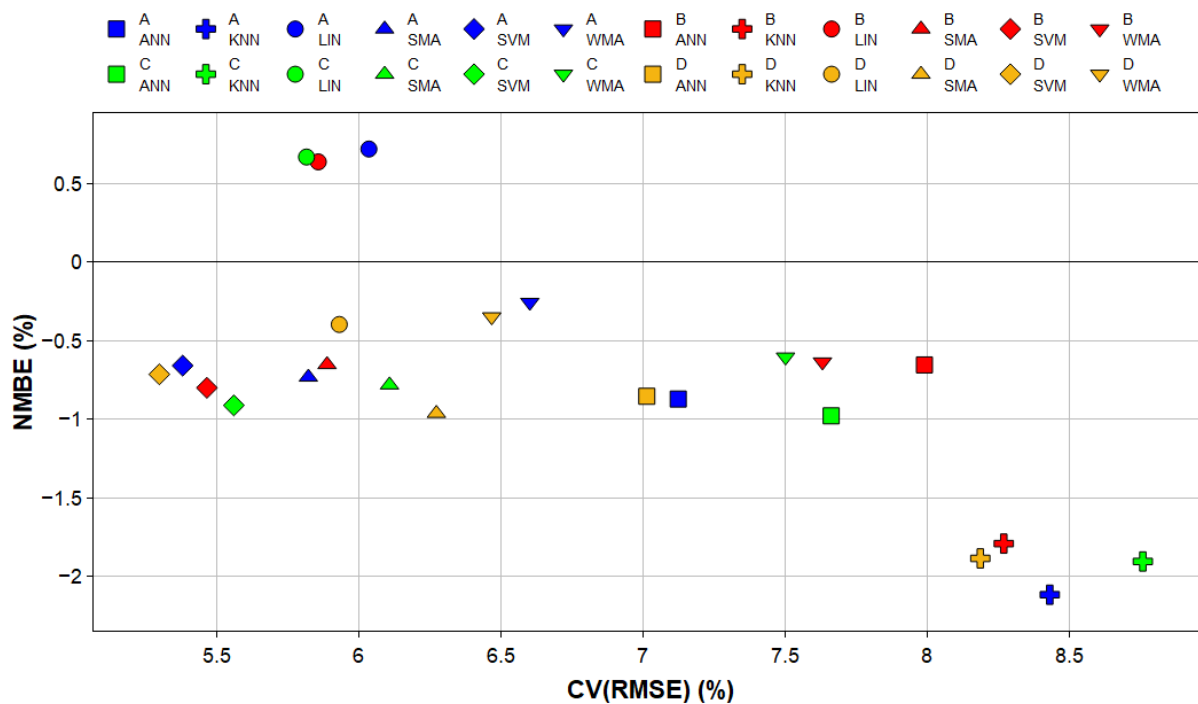


Figure C.45: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for the weekly frequency of case study I

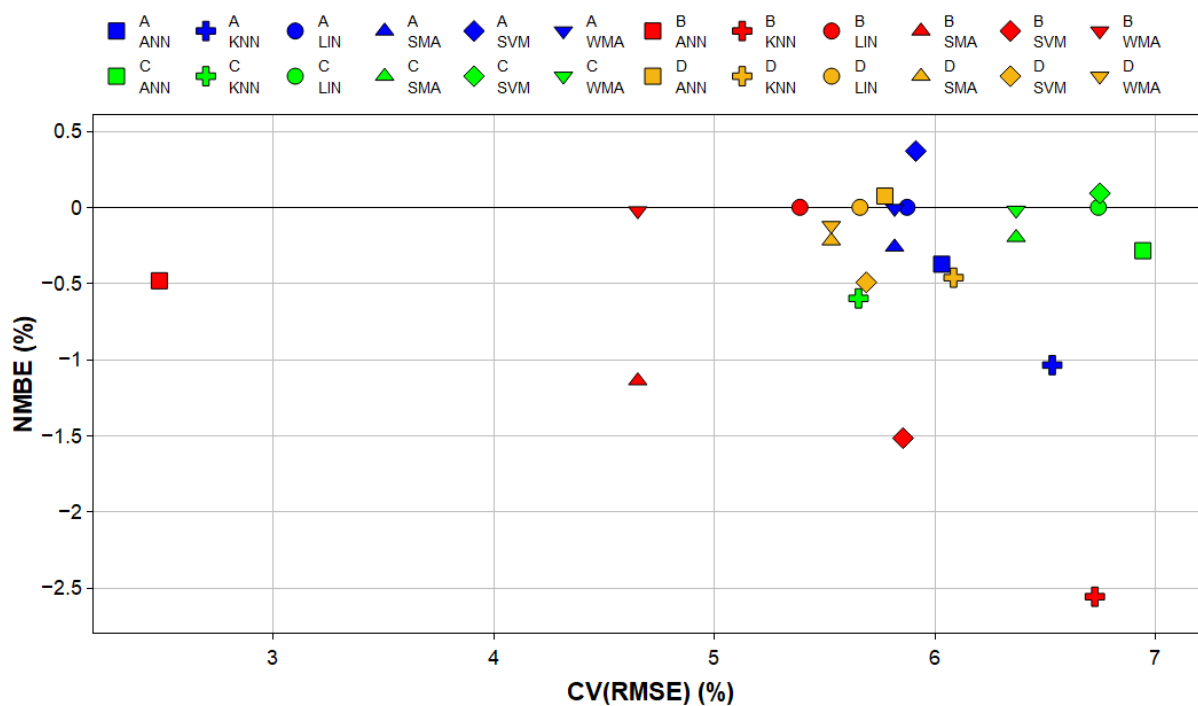


Figure C.46: NMBE vs CV(RMSE) for the forward models when using all the training data for the monthly frequency of case study I

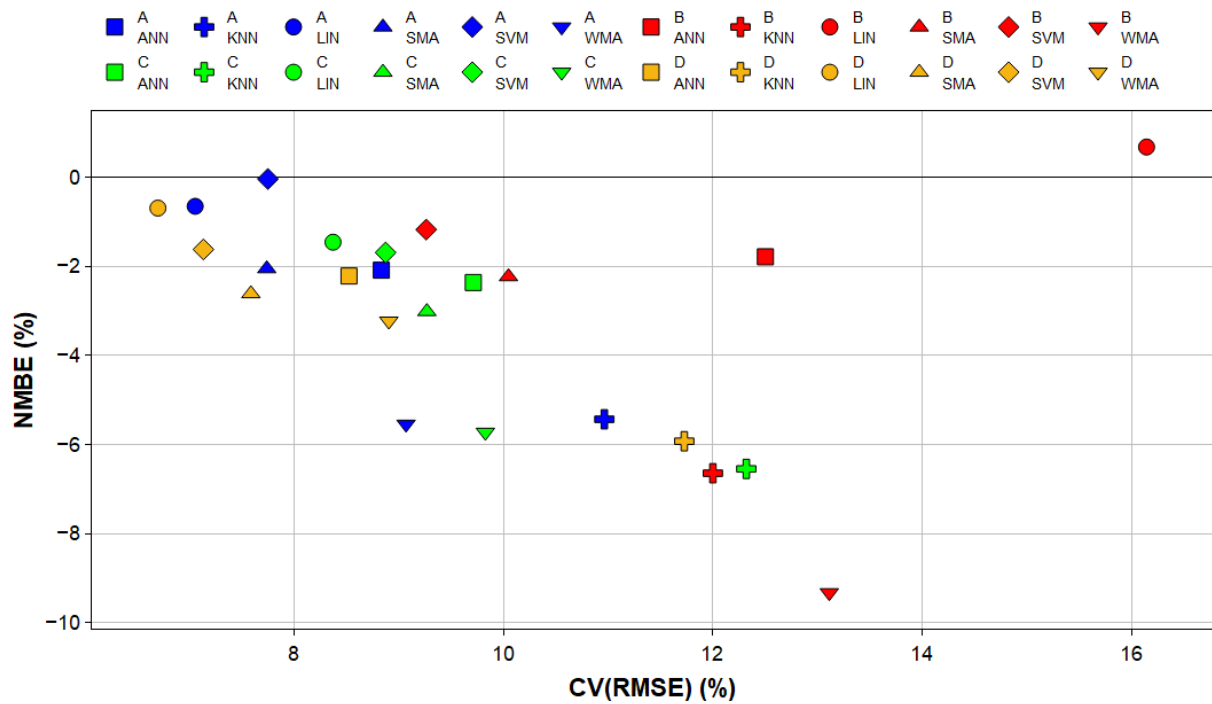


Figure C.47: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for the monthly frequency of case study I

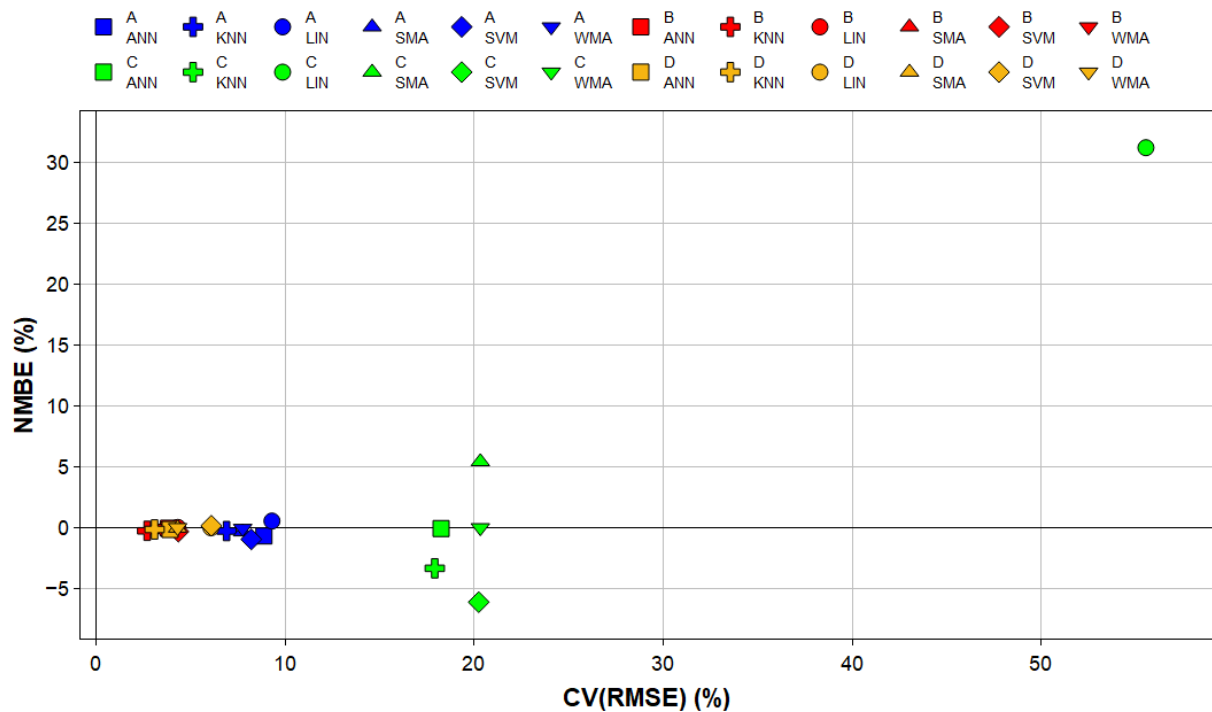


Figure C.48: NMBE vs CV(RMSE) for the forward models when using all the training data for case study II

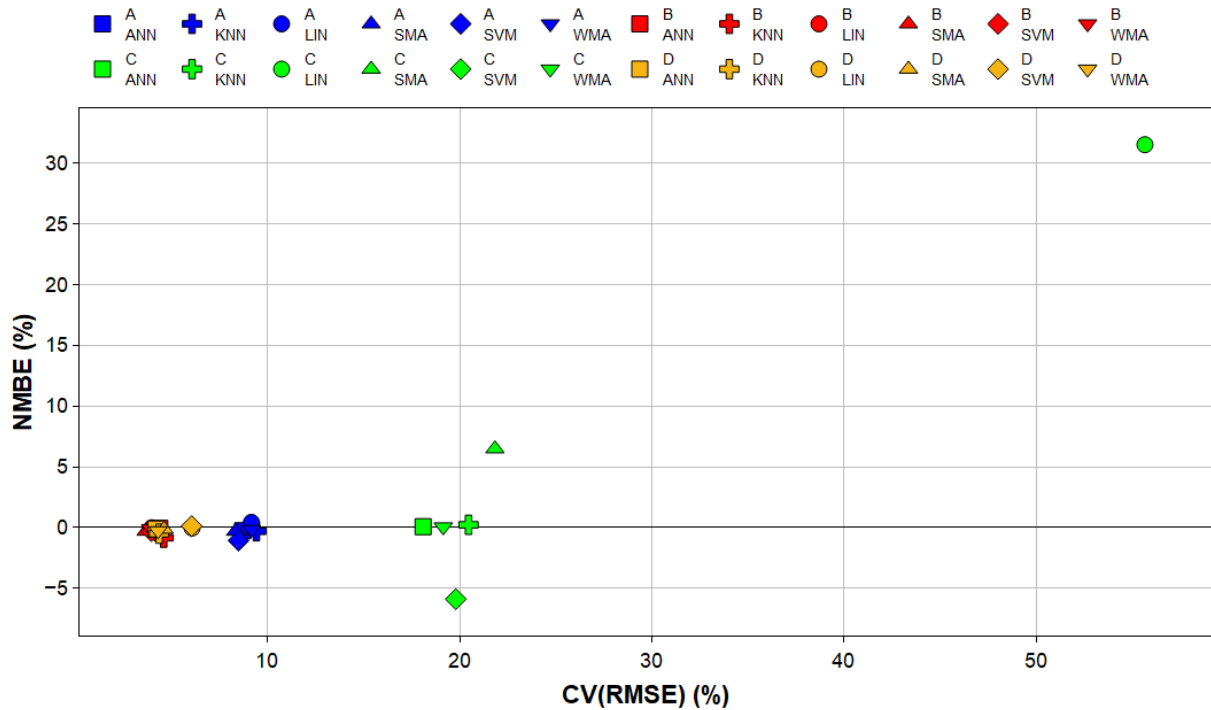


Figure C.49: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for case study II

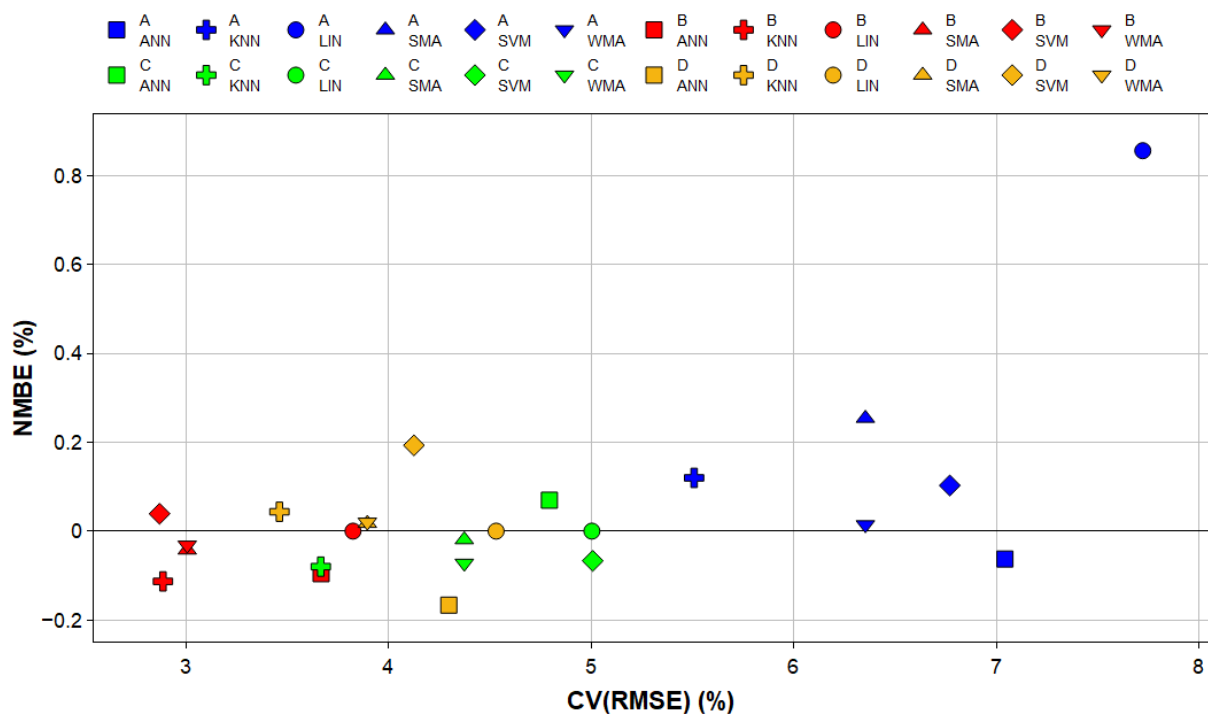


Figure C.50: NMBE vs CV(RMSE) for the forward models when using all the training data for case study III

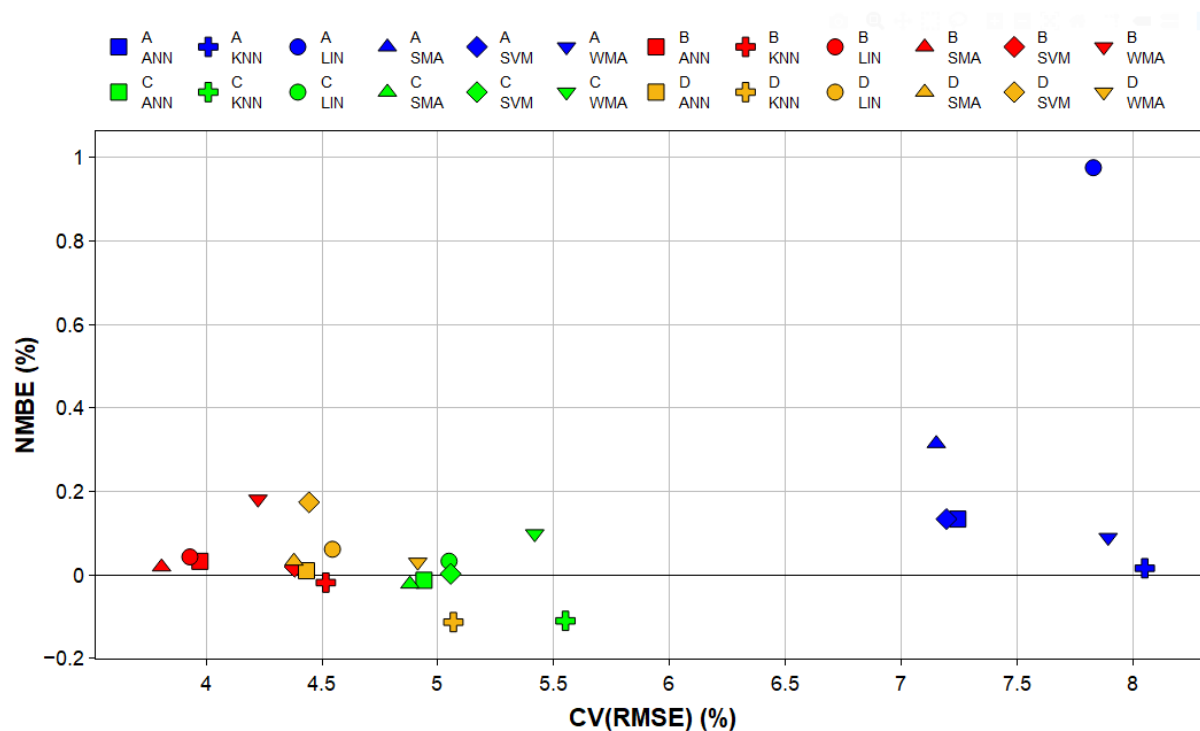


Figure C.51: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for case study III

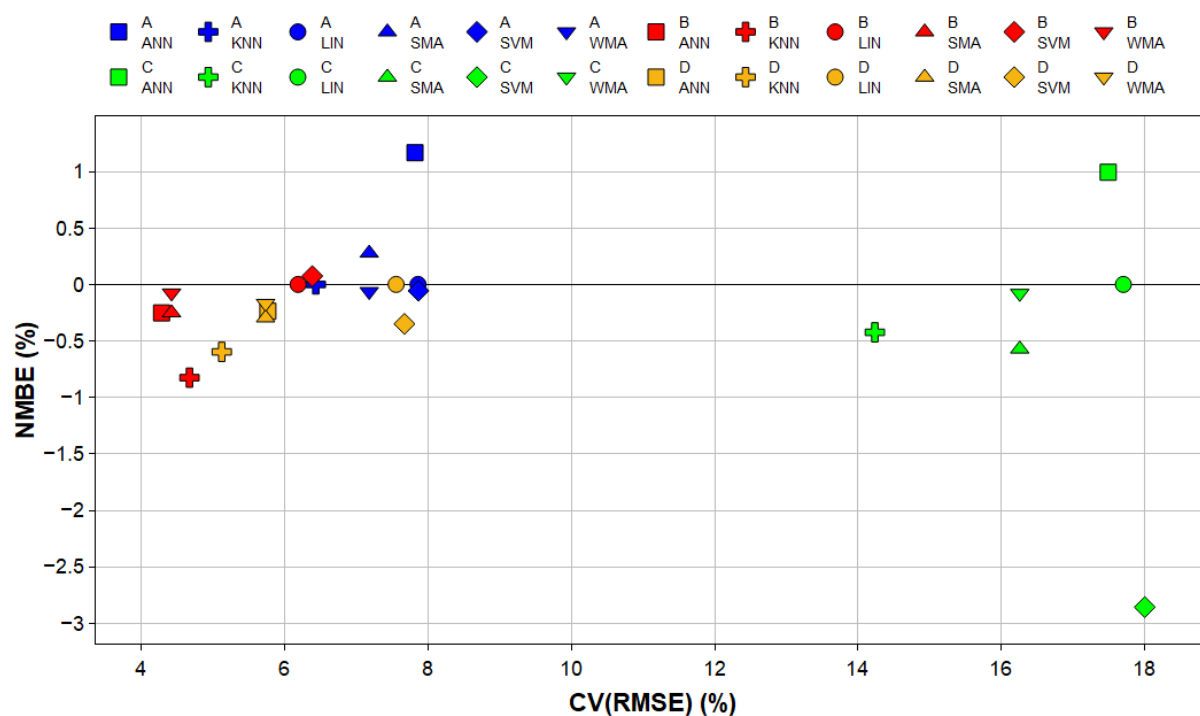


Figure C.52: NMBE vs CV(RMSE) for the forward models when using all the training data for case study IV

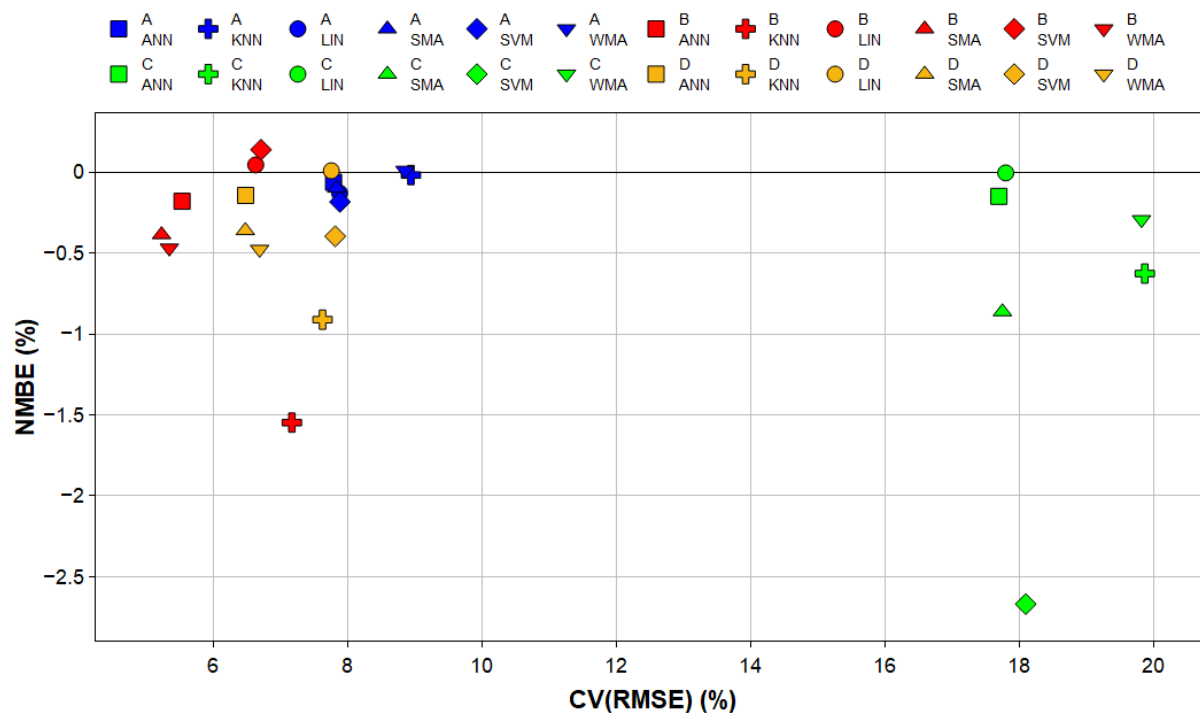


Figure C.53: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for case study IV

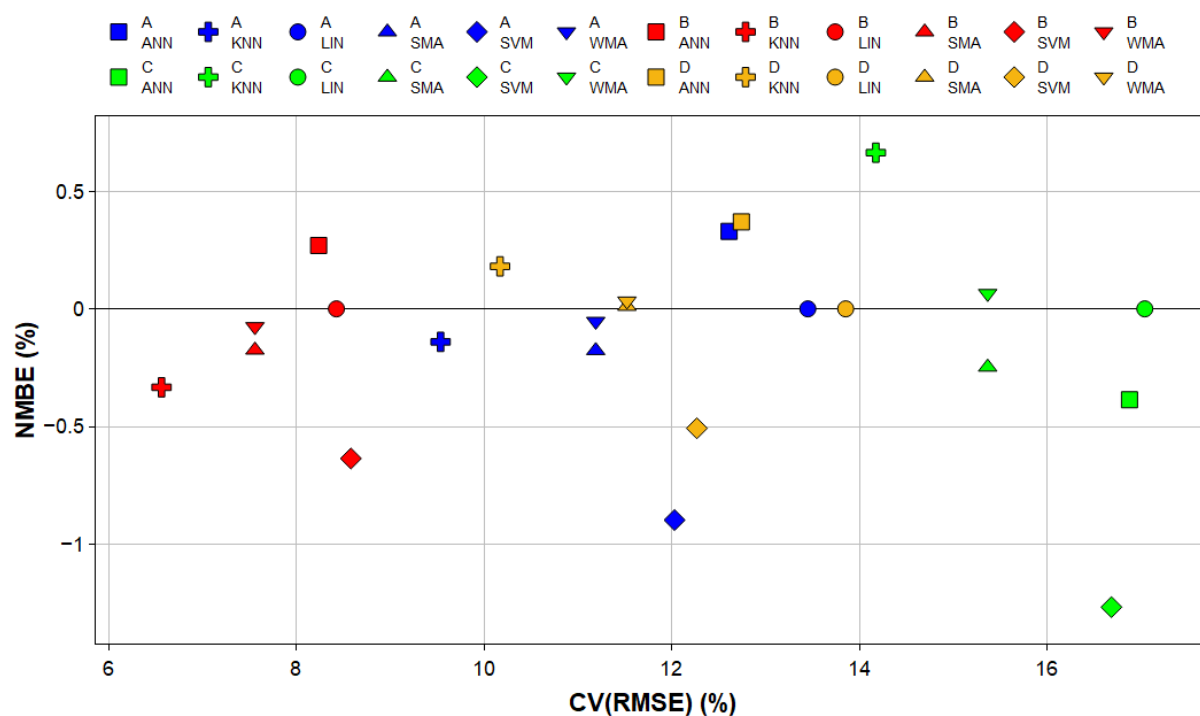


Figure C.54: NMBE vs CV(RMSE) for the forward models when using all the training data for case study V

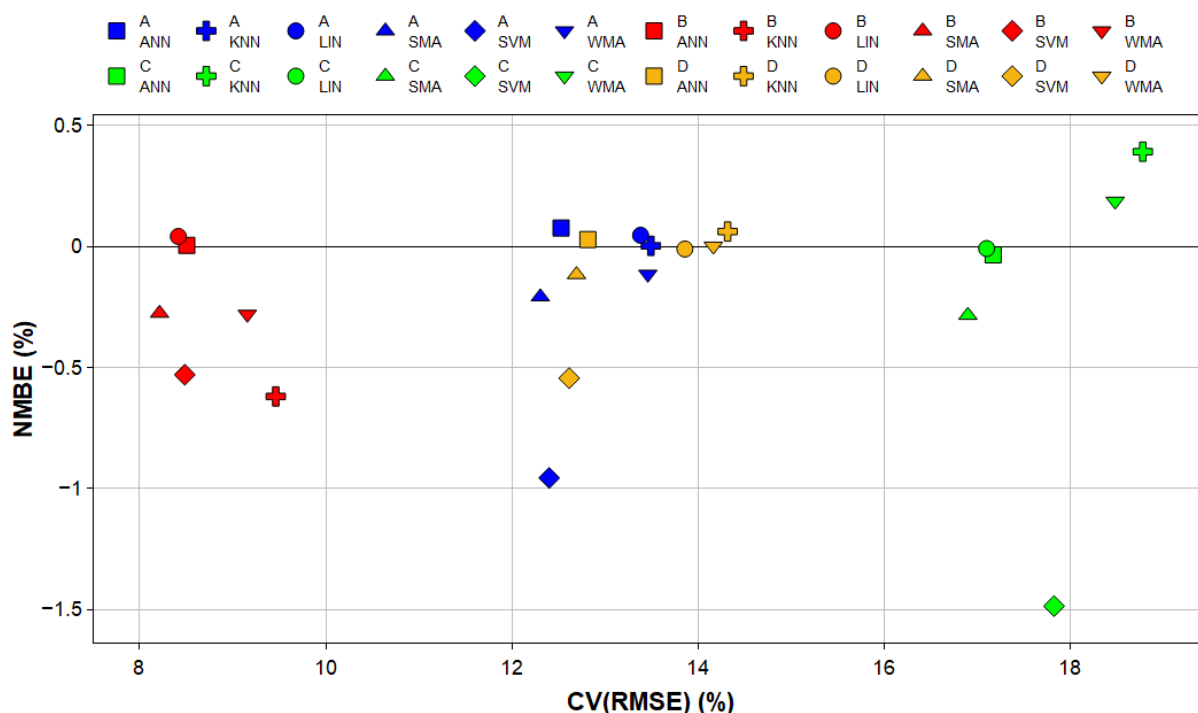


Figure C.55: NMBE vs CV(RMSE) for the sample-data approach in the forward direction (averages) for case study V

Appendix C9 – Selected hyperparameters for the different case studies

The selected hyperparameters for the different model types are given in Table C.1 to Table C.8. The following general patterns appear:

- (i) The numerical hyperparameters vary in range across different case studies and driver sets;
- (ii) The most optimal ANN according to the k-fold cross-validation did not always use the maximum number of nodes available and in some cases as few as one node was used in the hidden layer;
- (iii) The chosen kernel functions for the SVM and KNN model types, the distance calculation method and the use of an intercept tend to agree across driver sets for a certain case study with some exceptions.

Table C.1: Selected hyperparameters for case study I (daily)

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	Without intercept	Without intercept	Without intercept	With intercept
ANN				
<i>Nodes</i>	11	8	11	10
SVM				
<i>Kernel function</i>	linear	linear	linear	linear
<i>Gamma</i>	2	1	2	3
<i>Cost</i>	10	1	2	2
KNN				
<i>k-Value</i>	3	5	3	3
<i>Method</i>	Euclidean	Euclidean	Euclidean	Euclidean
<i>Kernel function</i>	uniform	uniform	uniform	uniform
<i>Lambda</i>	1	1	1	1

Table C.2: Selected hyperparameters for case study I (weekly)

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	Without intercept	Without intercept	Without intercept	With intercept
ANN				
<i>Nodes</i>	10	4	3	9
SVM				
<i>Kernel function</i>	Linear	Linear	Linear	Linear
<i>Gamma</i>	1	1	1	1
<i>Cost</i>	1	2	10	2
KNN				
<i>k-Value</i>	5	3	3	5
<i>Method</i>	Euclidean	Euclidean	Euclidean	Euclidean
<i>Kernel function</i>	Uniform	Uniform	Uniform	Uniform
<i>Lambda</i>	1	1	1	1

Table C.3: Selected hyperparameters for case study I (monthly)

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	Without intercept	Without intercept	Without intercept	With intercept
ANN				
<i>Nodes</i>	11	8	11	10
SVM				
<i>Kernel function</i>	Linear	Linear	Linear	Linear
<i>Gamma</i>	2	1	2	3
<i>Cost</i>	10	1	2	2
KNN				
<i>k-Value</i>	3	5	3	3
<i>Method</i>	Euclidean	Euclidean	Euclidean	Euclidean
<i>Kernel function</i>	Uniform	Uniform	Uniform	Uniform
<i>Lambda</i>	1	1	1	1

Table C.4: Selected hyperparameters for case study II

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	Without intercept	Without intercept	Without intercept	With intercept
ANN				
<i>Nodes</i>	9	11	1	11
SVM				
<i>Kernel function</i>	Radial	Linear	Linear	Linear
<i>Gamma</i>	2	2	2	3
<i>Cost</i>	10	10	1	10
KNN				
<i>k-Value</i>	3	3	3	3
<i>Method</i>	Euclidean	Euclidean	Euclidean	Mahalanobis
<i>Kernel function</i>	Triangular	Triangular	Triangular	Triangular
<i>Lambda</i>	1	1	1	1

Table C.6: Selected hyperparameters for case study III

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	Without intercept	With intercept	With intercept	With intercept
ANN				
<i>Nodes</i>	11	4	10	10
SVM				
<i>Kernel function</i>	Radial	Radial	Linear	Radial
<i>Gamma</i>	2	1	3	2
<i>Cost</i>	10	10	10	10
KNN				
<i>k-Value</i>	3	3	3	3
<i>Method</i>	Euclidean	Euclidean	Euclidean	Euclidean
<i>Kernel function</i>	Triangular	Triangular	Triangular	Triangular
<i>Lambda</i>	1	1	1	1

Table C.7: Selected hyperparameters for case study IV

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	With intercept	With intercept	With intercept	With intercept
ANN				
<i>Nodes</i>	8	9	7	11
SVM				
<i>Kernel function</i>	Linear	Linear	Linear	Linear
<i>Gamma</i>	1	1	1	3
<i>Cost</i>	4	4	2	2
KNN				
<i>k-Value</i>	3	3	3	3
<i>Method</i>	Euclidean	Euclidean	Euclidean	Euclidean
<i>Kernel function</i>	Triangular	Triangular	Triangular	Triangular
<i>Lambda</i>	1	1	1	1

Table C.8: Selected hyperparameters for case study V

Hyperparameter	Driver set			
	A	B	C	D
LIN				
<i>Intercept</i>	With intercept	With intercept	With intercept	With intercept
ANN				
<i>Nodes</i>	3	2	1	2
SVM				
<i>Kernel function</i>	radial	linear	radial	radial
<i>Gamma</i>	2	1	3	2
<i>Cost</i>	10	10	4	4
KNN				
<i>k-Value</i>	3	3	3	3
<i>Method</i>	euclidean	euclidean	euclidean	euclidean
<i>Kernel function</i>	Triangular	Triangular	Triangular	Triangular
<i>Lambda</i>	1	1	1	1

Appendix C10 – The calculated metrics

The metrics calculated for each case study for the model-scoring is given in Table C.9 to Table C.13

Table C.9: The metrics calculated for case study I (daily frequency)

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
Error - CV(RMSE)	A	7.66	7.46	7.52	8.17	7.32	8.05
	B	7.50	7.01	7.19	7.45	6.79	6.97
	C	8.86	7.68	7.98	8.24	7.60	8.18
	D	7.50	7.51	7.51	8.44	7.43	8.33
	<i>Average</i>	7.88	7.41	7.55	8.08	7.28	7.88
Error - NMBE	A	0.29	0.12	0.10	0.30	0.01	0.27
	B	0.30	0.13	0.17	0.51	0.13	0.23
	C	0.85	0.15	0.39	0.30	0.00	0.34
	D	0.18	0.25	0.07	0.40	0.22	0.16
	<i>Average</i>	0.40	0.16	0.18	0.37	0.09	0.25
Savings std. dev. across driver sets	NA	0.11	0.58	0.20	0.50	0.60	0.28
Savings std. dev. for sensitivity repetitions	A	0.12	0.23	0.20	0.20	0.14	0.18
	B	0.42	0.44	0.34	0.18	0.26	0.32
	C	0.15	0.26	0.16	0.19	0.15	0.19
	D	0.12	0.28	0.15	0.16	0.13	0.16
	<i>Average</i>	0.20	0.30	0.21	0.18	0.17	0.22

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
CV(RMSE) std. dev. for sensitivity repetitions	A	0.65	0.65	0.58	0.72	0.59	0.69
	B	0.64	0.70	0.55	0.80	0.60	0.74
	C	0.73	0.69	0.58	0.68	0.61	0.68
	D	0.61	0.67	0.60	0.88	0.65	0.88
	Average	0.66	0.68	0.58	0.77	0.61	0.75
NMBE std. dev. for sensitivity repetitions	A	0.85	0.94	1.02	0.94	0.88	0.87
	B	0.79	0.87	0.79	0.77	0.72	0.65
	C	0.99	0.91	0.98	0.86	0.86	1.00
	D	0.86	0.89	0.90	0.87	0.83	0.93
	Average	0.87	0.90	0.92	0.86	0.82	0.86
Consistency difference between forward & backward	A	9.21	5.68	19.92	24.59	9.66	20.24
	B	44.58	16.90	40.93	9.84	30.05	12.04
	C	0.41	7.44	4.07	3.59	22.39	13.37
	D	7.66	3.10	20.57	23.27	10.55	28.20
	Average	15.46	8.28	21.37	15.32	18.16	18.46

Table C.10: The metrics calculated for case study II

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
Error - CV(RMSE)	A	9.16	8.71	8.49	9.42	8.34	9.17
	B	3.98	4.39	3.95	4.60	3.69	3.92
	C	55.66	18.10	19.80	20.46	21.84	19.15
	D	6.08	4.20	6.06	4.36	4.62	4.31
	Average	18.72	8.85	9.58	9.71	9.62	9.14
Error - NMBE	A	0.42	0.27	1.08	0.31	0.31	0.17
	B	0.02	0.07	0.29	0.84	0.31	0.09
	C	31.48	0.07	5.89	0.22	6.47	0.05
	D	0.04	0.12	0.11	0.51	0.14	0.27
	Average	7.99	0.13	1.84	0.47	1.81	0.14
Savings std. dev. across driver sets	NA	12.97	12.73	15.41	11.32	12.19	13.01
Savings std. dev. for sensitivity repetitions	A	0.29	1.30	0.98	0.74	0.60	0.74
	B	0.21	1.10	0.28	0.39	0.30	0.97
	C	0.72	0.67	0.57	0.96	0.51	1.34
	D	0.43	0.92	0.44	0.30	0.36	0.31
	Average	0.41	1.00	0.57	0.60	0.44	0.84
CV(RMSE) std. dev. for sensitivity repetitions	A	1.35	1.16	1.19	0.90	1.03	0.95
	B	1.88	1.98	1.94	1.31	1.59	1.49
	C	3.03	2.63	3.14	2.95	1.86	2.56
	D	1.01	1.51	1.00	1.10	1.20	1.13
	Average	1.82	1.82	1.82	1.56	1.42	1.53
	A	1.02	1.19	1.09	1.04	1.00	0.62

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
<i>NMBE std. dev. for sensitivity repetitions</i>	B	0.51	0.60	0.49	0.53	0.46	0.33
	C	3.77	2.10	1.95	4.08	2.24	2.59
	D	0.71	0.55	0.64	0.46	0.51	0.58
	Average	1.50	1.11	1.04	1.53	1.05	1.03
<i>Consistency difference between forward & backward</i>	A	11.45	107.48	37.11	51.99	68.81	11.73
	B	79.63	77.62	74.75	311.06	106.57	142.08
	C	136.61	7.27	1.85	2.04	39.75	10.40
	D	323.95	126.79	538.10	17.02	321.15	7.07
	Average	137.91	79.79	162.95	95.53	134.07	42.82

Table C.11: The metrics calculated for case study III

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
<i>Error - CV(RMSE)</i>	A	7.83	7.24	7.19	8.05	7.15	7.89
	B	3.93	3.97	4.38	4.52	3.81	4.22
	C	5.05	4.94	5.06	5.55	4.88	5.42
	D	4.55	4.43	4.44	5.07	4.38	4.91
	Average	5.34	5.15	5.27	5.80	5.05	5.61
<i>Error - NMBE</i>	A	0.98	0.13	0.13	0.02	0.31	0.09
	B	0.04	0.03	0.02	0.02	0.02	0.18
	C	0.03	0.01	0.00	0.11	0.02	0.10
	D	0.06	0.01	0.17	0.11	0.03	0.03
	Average	0.28	0.05	0.08	0.06	0.10	0.10
<i>Savings std. dev. across driver sets</i>	NA	3.65	3.16	3.09	2.98	3.07	3.19
<i>Savings std. dev. for sensitivity repetitions</i>	A	0.18	0.39	0.40	0.28	0.19	0.27
	B	0.36	0.59	0.99	0.34	0.34	0.70
	C	0.15	0.26	0.21	0.27	0.16	0.25
	D	0.10	0.25	0.20	0.22	0.15	0.22
	Average	0.20	0.37	0.45	0.28	0.21	0.36
<i>CV(RMSE) std. dev. for sensitivity repetitions</i>	A	0.59	0.46	0.92	0.51	0.49	0.50
	B	0.33	0.34	1.31	0.44	0.46	0.98
	C	0.34	0.32	0.34	0.39	0.31	0.37
	D	0.31	0.35	0.72	0.44	0.39	0.44
	Average	0.39	0.37	0.82	0.44	0.41	0.57
<i>NMBE std. dev. for sensitivity repetitions</i>	A	0.74	0.82	0.75	0.83	0.72	0.92
	B	0.43	0.60	0.57	0.51	0.46	0.70
	C	0.60	0.63	0.63	0.67	0.59	0.89
	D	0.45	0.55	0.51	0.53	0.46	0.94
	Average	0.55	0.65	0.62	0.63	0.56	0.86
	A	4.74	30.76	4.56	16.71	1.73	15.35

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
Consistency difference between forward & backward	B	943.16	412.42	82.52	481.50	256.29	148.40
	C	30.75	38.75	20.53	138.17	115.65	21.90
	D	2.46	6.68	0.09	6.69	3.49	4.00
	Average	245.28	122.15	26.92	160.77	94.29	47.41

Table C.12: The metrics calculated for case study IV

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
Error - CV(RMSE)	A	7.88	7.79	7.89	8.94	7.84	8.85
	B	6.63	5.53	6.71	7.17	5.23	5.34
	C	17.80	17.70	18.10	19.87	17.75	19.83
	D	7.76	6.48	7.81	7.63	6.48	6.69
	Average	10.02	9.38	10.13	10.90	9.32	10.18
Error - NMBE	A	0.13	0.06	0.19	0.02	0.10	0.01
	B	0.04	0.18	0.14	1.55	0.39	0.47
	C	0.01	0.15	2.67	0.63	0.86	0.29
	D	0.01	0.15	0.40	0.91	0.36	0.48
	Average	0.05	0.14	0.85	0.78	0.43	0.31
Savings std. dev. across driver sets	NA	3.60	3.33	4.47	2.31	2.72	3.32
Savings std. dev. for sensitivity repetitions	A	0.33	0.70	0.41	0.40	0.37	0.42
	B	0.75	0.98	0.61	0.56	0.48	0.57
	C	0.74	1.37	0.82	1.15	0.85	1.17
	D	0.39	0.66	0.47	0.57	0.38	0.48
	Average	0.56	0.93	0.58	0.67	0.52	0.66
CV(RMSE) std. dev. for sensitivity repetitions	A	0.65	0.68	0.65	0.73	0.65	0.73
	B	0.87	0.82	1.04	0.84	0.68	0.71
	C	1.54	1.58	1.69	1.62	1.62	1.58
	D	0.89	0.62	0.97	0.83	0.75	0.76
	Average	0.99	0.92	1.09	1.00	0.93	0.95
NMBE std. dev. for sensitivity repetitions	A	0.87	0.98	0.91	0.94	0.87	1.43
	B	0.74	0.72	0.79	0.77	0.57	0.67
	C	2.07	2.43	2.12	2.17	2.10	2.33
	D	0.86	0.85	0.91	0.83	0.75	0.86
	Average	1.14	1.25	1.18	1.18	1.07	1.32
Consistency difference between forward & backward	A	5.74	7.74	12.28	5.32	4.98	10.25
	B	22.59	5.37	59.98	53.84	17.26	11.48
	C	11.94	11.58	11.75	44.26	50.25	11.99
	D	162.44	181.71	119.41	15.46	68.85	103.79
	Average	50.68	51.60	50.85	29.72	35.34	34.38

Table C.13: The metrics calculated for case study V

Parameter	Driver set	Models					
		LIN	ANN	SVM	KNN	SMA	WMA
Error - CV(RMSE)	A	13.38	12.53	12.40	13.49	12.31	13.46
	B	8.42	8.51	8.49	9.46	8.21	9.16
	C	17.10	17.17	17.83	18.78	16.90	18.48
	D	13.86	12.82	12.62	14.32	12.70	14.17
	<i>Average</i>	13.19	12.76	12.83	14.01	12.53	13.82
Error - NMBE	A	0.05	0.08	0.96	0.00	0.21	0.12
	B	0.04	0.00	0.53	0.62	0.28	0.28
	C	0.01	0.04	1.49	0.39	0.28	0.19
	D	0.01	0.03	0.54	0.06	0.12	0.00
	<i>Average</i>	0.03	0.04	0.88	0.27	0.22	0.15
Savings std. dev. across driver sets	NA	2.27	2.09	2.57	0.93	1.20	1.94
Savings std. dev. for sensitivity repetitions	A	0.36	0.58	0.29	0.35	0.33	0.35
	B	0.31	0.52	0.18	0.28	0.25	0.32
	C	0.42	0.58	0.78	0.56	0.48	0.54
	D	0.36	0.52	0.32	0.43	0.34	0.44
	<i>Average</i>	0.36	0.55	0.39	0.41	0.35	0.41
CV(RMSE) std. dev. for sensitivity repetitions	A	1.24	1.34	1.41	1.36	1.31	1.34
	B	1.02	1.07	1.12	0.90	1.04	0.87
	C	1.35	1.44	1.93	1.65	1.45	1.74
	D	1.07	1.16	1.12	1.03	1.05	1.05
	<i>Average</i>	1.17	1.25	1.40	1.24	1.21	1.25
NMBE std. dev. for sensitivity repetitions	A	1.60	1.65	1.29	1.44	1.40	1.69
	B	0.90	1.02	0.76	0.93	0.85	0.81
	C	1.94	2.03	2.24	1.95	1.93	1.75
	D	1.55	1.56	1.28	1.59	1.40	1.27
	<i>Average</i>	1.50	1.56	1.39	1.48	1.40	1.38
Consistency difference between forward & backward	A	1.00	4.68	19.27	0.24	6.94	2.12
	B	586.87	639.95	729.59	29.69	7.87	132.13
	C	0.57	0.40	41.35	19.20	2.84	3.99
	D	0.69	0.60	7.94	10.94	8.89	7.89
	<i>Average</i>	147.28	161.41	199.54	15.02	6.63	36.53

Appendix C11 – The baseline test predictions

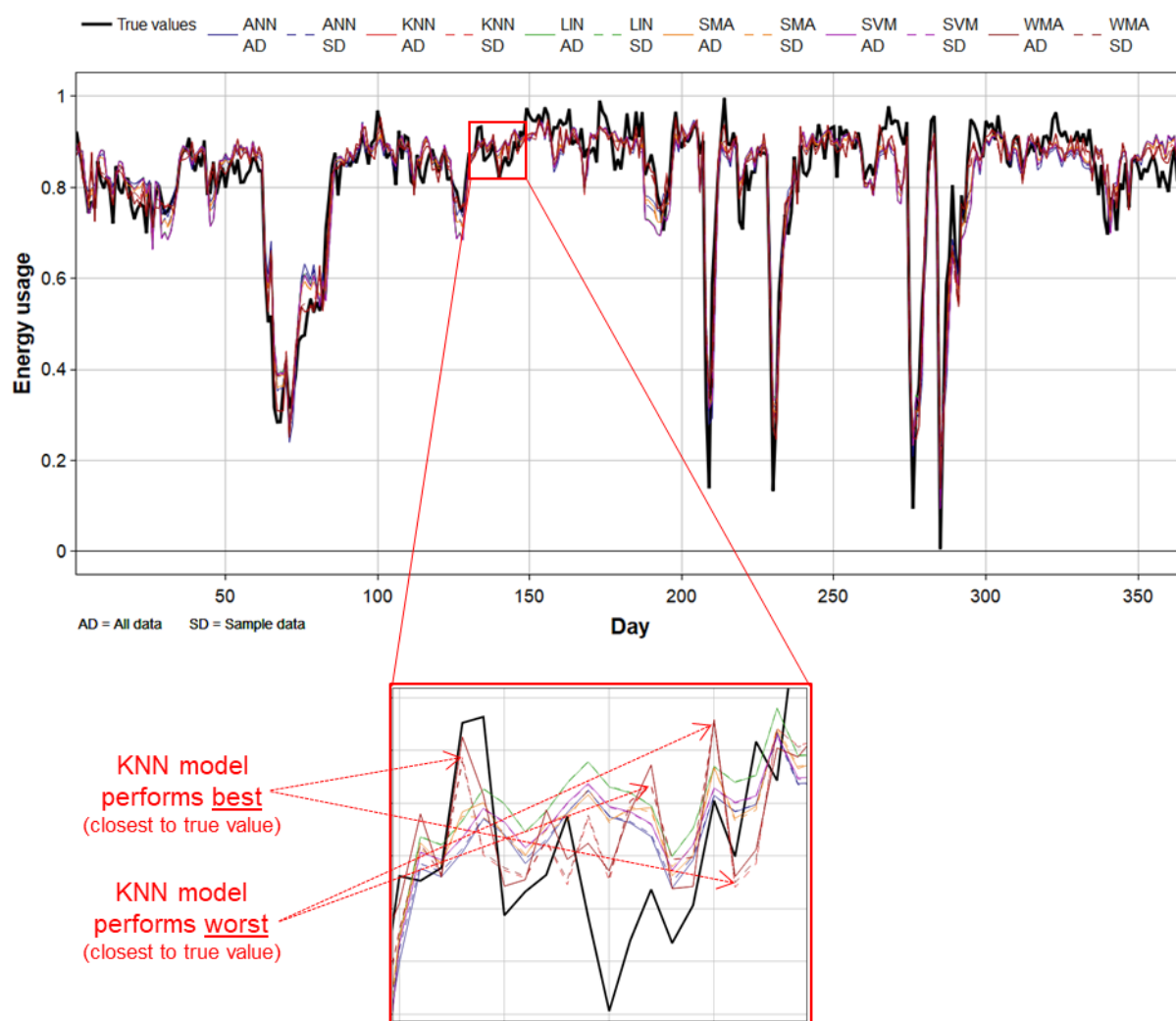


Figure C.56: The baseline test predictions for case study I using driver set A for case study I

This section only looks at the daily frequency of case study I as an example. Figure C.56 shows the predictions of the baseline models for both the all-data (AD) and sample-data (SD) approaches for the baseline dataset when using driver set A and the true energy usage (solid black line). I.e., it shows how well the respective models can match the training data. All the models tend to match the predictions of one another more than the true energy usage. However, there is still some variation, as can be seen when enlarging a portion of the graph (in the red block below the graph).

It can be understood why the SMA model type yields improved accuracy (as discussed in section 3.6) on average (based on the CV(RMSE)) when considering the enlarged portion of the

graph. The KNN model type can be used as an example to illustrate this. It should be remembered that the SMA models are only influenced by the single model types (LIN, ANN, SVM and KNN). For some of the data points, the predictions of the KNN model type are the closest to true values for both approaches while for other datapoints it is the furthest away. The same argument can be applied to each of the model types. This explains why SMA gives the best result in terms of absolute accuracy.

The same arguments presented here are also true for the other driver sets of case study I and for the remaining case studies. The baseline prediction plots for case study I for driver sets B to D are shown in Figure C.57 to Figure C.59.

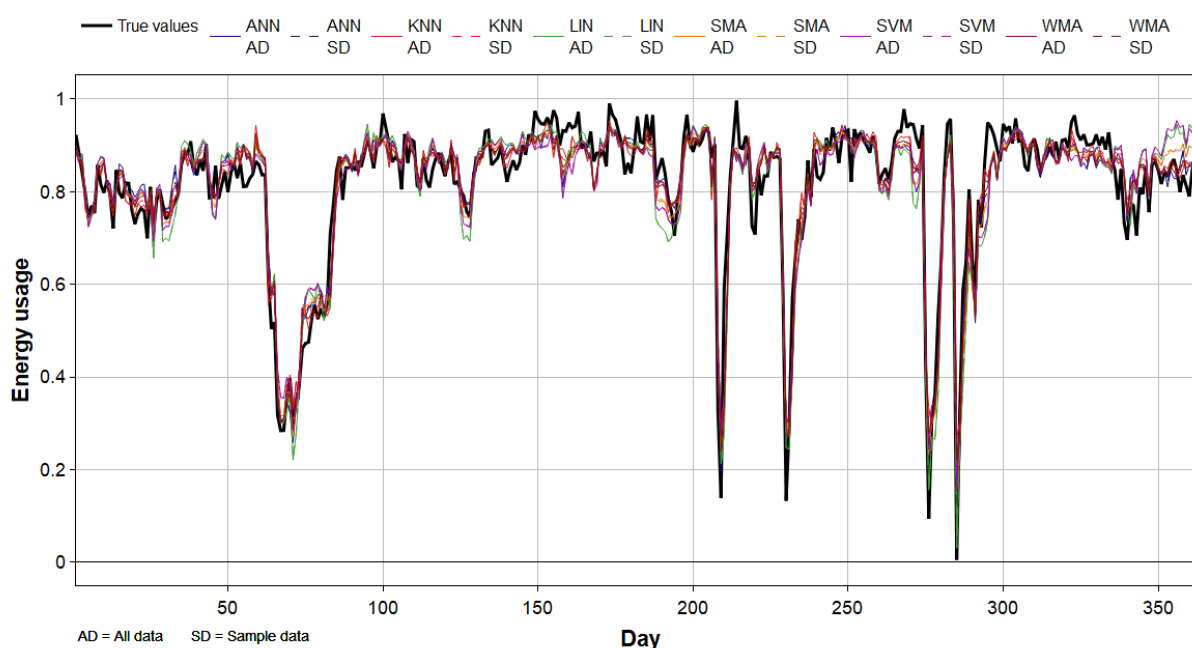


Figure C.57: The baseline test predictions for case study I using driver set A for case study I

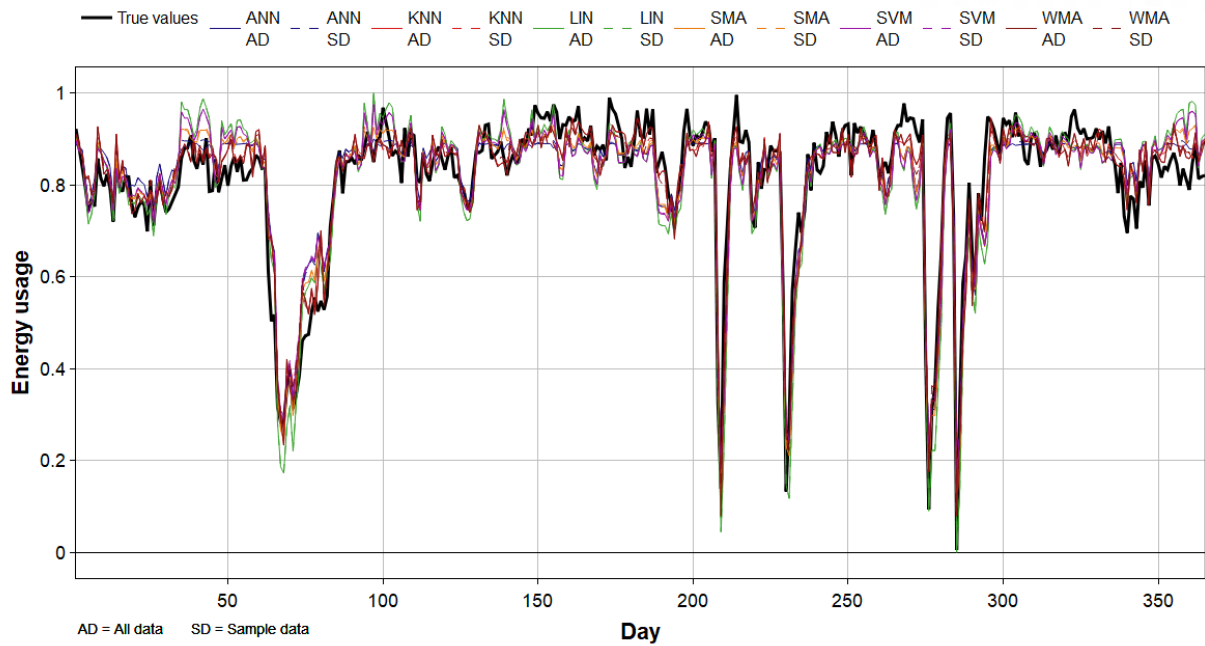


Figure C.58: The baseline test predictions for case study I using driver set A for case study I

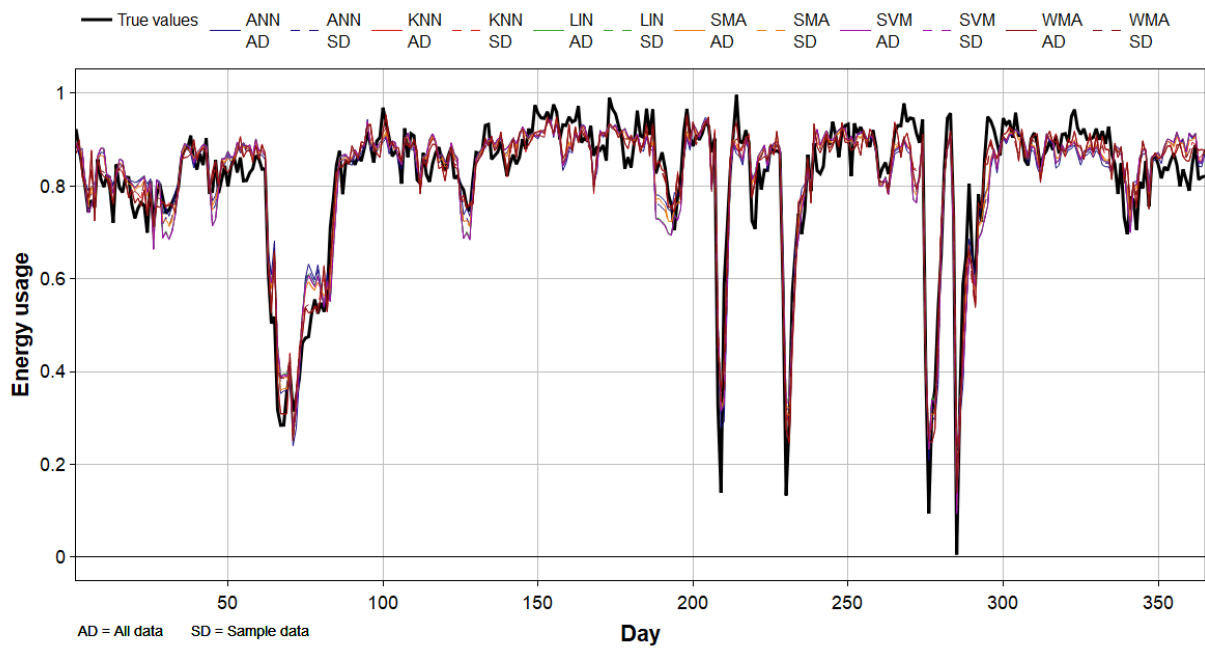


Figure C.59: The baseline test predictions for case study I using driver set A for case study I