

Application of M&V 2.0 techniques to identify and quantify non-routine events

JM Opperman



orcid.org/0000-0001-5608-0888

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

Supervisor: Dr K Nell

Co-Supervisor: Dr JF van Rensburg

Graduation: November 2022

Student number: 25933337

Abstract

Energy efficiency measures (EEM) are implemented on mines to reduce energy consumption. The impact of an EEM must be determined to evaluate if any energy savings were achieved. This is done by applying measurement and verification (M&V) principles. However, unanticipated operational changes known as non-routine events (NRE), inhibit accurate quantification of an EEM.

One challenge in industry is to rapidly identify an NRE. Further investigation on mines brought to light that NREs are detected and quantified through manual monitoring and strategy application. This process may take days or weeks and can lead to energy over expenditure or increased costs. Mines have implemented advanced metering infrastructure that continuously expand databases. To address the big data availability, M&V 2.0 emerged over the last decade. This field has seen little contribution in terms of anomaly detection in mining equipment. In this study, a data driven approach to rapidly identify NREs in mines is investigated.

The literature study identified the importance of data preparation to expose trends and gain insight. The methodology included the organisation of mining energy data according to Eskom time of use (TOU) periods. Changepoint detection strategies were then applied to each respective TOU category. Research stipulated that the cumulative sums (CUSUM), a prescribed M&V technique, was suitable to identify anomalies in large data sets. To contribute to the field of M&V 2.0, this study compared two changepoint detection algorithms to the CUSUM method—pruned exact linear time (PELT) and binary segmentation (BinSeg). A package called **changepoint**, developed in the R programming language, was applied to implement the BinSeg and PELT algorithms.

It was found that the BinSeg and PELT algorithms identified more changepoints than the CUSUM method. Overall, PELT was the most accurate. One flaw in the CUSUM method was that it did not identify changepoints in data with high variability. Whereas, the PELT and BinSeg algorithms may have been inaccurate if the parameters were not correctly specified for each data set. One such parameter was the minimum segment length, which stipulated the minimum duration of a valid changepoint. In this study, the minimum segment length was generalised to mining energy data as a whole and should have been specific to each data set.

In conclusion, the algorithm driven technique was proved to be effective in identifying NREs in mining energy data. Such algorithms can be applied to every data stream in the mine. The impact of this is more rapid identification of equipment failure or increased energy consumption. In this study, energy data from the national lockdown was used. During this period, increased demand in the gold plant energy data was identified. In a time where the mine could not generate revenue, the increased energy on the gold plant was estimated to cost R27000 daily.

Keywords: M&V 2.0; changepoint; CUSUM, BinSeg, PELT, mining

Acknowledgements

It was a unique opportunity to build a Masters degree on a real world application. I would like to thank Dr Johan Marais for aiding in the formulation of this study topic.

The guidance from Dr Kristy Nell was imperative to the successful completion of this thesis. Thank you sincerely for the hours of your personal time dedicated to improving this document.

Thank you, Prof Edward Mathews, for developing a program that promotes the advancement of academia relevant to pertinent environmental problems faced in South-African mines. This is also made possible by the financial support provided by ETA Operations Ltd and Enermanage (Pty) Ltd.

Contents

1	Background and literature review	1
1.1	Preamble	1
1.2	Non-routine event detection in mines and application of M&V 2.0 methods . . .	1
1.3	Data driven non-routine event detection applied in industry	4
1.3.1	Data quality and preparation for analysis	4
1.3.2	IPMVP-prescribed techniques for non-routine event detection	7
1.3.3	Algorithms for non-routine event detection	14
1.3.4	Change-point detection theory	18
1.4	Verification of rapid non-routine event detection	29
1.5	Problem statement, need and objectives	33
1.6	Outline of document	34
2	Methodology for rapid non-routine event detectio	35
2.1	Preamble	35
2.2	Preparation of mining-energy data	35
2.3	IPMVP anomaly detection	41
2.4	Change point detection algorithms	43
2.5	Verification	48
2.6	Conclusion	49
3	Rapid change point-detection implementation and results	50
3.1	Preamble	50
3.2	Case study explanation and data preparation	50
3.3	IPMVP anomaly detection	58
3.4	Change-point detection algorithms	60
3.5	Validation	65
3.6	Conclusion	67
4	Conclusion and recommendations	68
4.1	Preamble	68
4.2	Study overview and problem statement	68
4.3	Objectives addressed	69
4.4	Critical analysis of the study	69
4.5	Study impact	70
A	Statistical concepts	76
B	Further insight from literature and mining	77
C	Further insight on the case study data	78

D Change point detection results

82

List of Figures

1.1	Areal schematic of a deep level gold mine showing a network of tunnels [12] . . .	3
1.2	Daily demand trends of an office, furniture store, and bakery (top, centre, and bottom).	5
1.3	Demand profiles from Figure 1.2 split into peak (top line) and low (bottom line) demand periods.	6
1.4	Example of normally-distributed data showing outliers at more than three standard deviations from the mean.	9
1.5	Example of normally distributed data along with 10% and standard deviation test metrics.	9
1.6	Repeated example of normally-distributed data along with 10% and standard deviation test metrics	10
1.7	Example of a CUSUM plot showing acceleration and, therefore, higher rate of accumulation.	11
1.8	CUSUM of the distance between each point and the mean divided by the standard deviation ($\sigma = 1$). The highlighted area relates to $S_{H/L}$ in Table 1.2	12
1.9	Positive (S_H) and negative (S_L) accumulation of the standard deviation (Z) with sensitivity set to three standard deviations, $\omega = \pm 3\sigma$	13
1.10	Example of normally-distributed data along with 10% and standard deviation test metrics.	14
1.11	Example of statistical property changes showing change in mean (left) and variance (right).	15
1.12	Statistical change point at time τ , showing change in mean of arbitrary data. . .	17
1.13	Arbitrary time series data with standard deviation $\sigma \approx 1$	19
1.14	Plot of log likelihood for the mean of the data given in Figure 1.13.	19
1.15	Log likelihood plot showing the likelihood of the null hypothesis ($\mu = 1$ and its vertical distance from the alternative hypothesis.	20
1.16	Log likelihood curves for two sets of the data compared to that of the entire data set.	21
1.17	Top: log likelihood ratio Equation 1.21 with respect to τ . Bottom: arbitrary time series data with standard deviation $\sigma \approx 1$. The red line shows μ_1 and μ_2 as was discovered in Figure 1.17.	22
1.18	Visual representation of the BinSeg algorithm.	24
1.19	Arbitrary time series data showing two increases in the mean.	25
1.20	Summed negative log likelihood of two segments to identify a single change point in the arbitrary data of Figure 1.19 between time points 0.0 and 15.0.	26
1.21	Summed negative log likelihood of the previously identified change point MIN_A and the two segments within mini-segment $x_{11}-x_{25}$ to identify a single change point in Figure 1.19.	26

1.22	Arbitrary time series data showing two change points.	27
1.23	The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.	28
1.24	The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.	28
1.25	Illustration of the bubble sort algorithm.	30
1.26	Illustration of the heap sort algorithm.	30
1.27	Comparison of the bubble and heap sort algorithm complexities.	30
2.1	Power data from a mining compressor for August 2020.	37
2.2	Power data from a mining compressor for a selection of weekdays in August 2020.	37
2.3	Average hourly power on weekdays of a mining compressor in August 2020. The grey bars indicate Eskom peak tariff hours.	38
2.4	Standard deviation plot for a mining compressor's weekday peak time-of-use power output in August 2020.	40
2.5	Baseline data for a mining compressor's weekday power output during Eskom's peak periods in August 2020.	40
2.6	Scaled baseline data for a mining compressor's weekday power output during Eskom peak periods in August 2020.	41
2.7	Negative CUSUM applied to a mining compressor's weekday power output during Eskom peak periods in August 2020.	43
2.8	The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.	44
2.9	Change points discovered by the BinSeg algorithm for Eskom-delineated peak tariff data of a mining compressor's weekday power profile.	45
2.10	Change points discovered by the PELT algorithm for the Eskom-delineated peak tariff data of a mining compressor's weekday power profile.	45
2.11	Task manager showing the memory usage on the computer to be used for change point detection implementation.	47
3.1	Half-hourly power output of a rock mill.	51
3.2	Half-hourly power output of a compressor.	51
3.3	Half-hourly power output of a gold plant.	52
3.4	Half-hourly power output of a fridge plant.	52
3.5	End of Christmas break start-up period in January.	53
3.6	Scaled Summer weekday off-peak time-of-use plot of (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.	54
3.7	S_WKD_OFF standard deviation plots of February 2020 for (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.	55
3.8	S_WKD_OFF cleaned and scaled baseline plots of February 2020 for (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.	57

3.9	Top: S_WKD_OFF power output of a rock mill in March 2020. Bottom: S_L CUSUM result showing detection of anomalies.	58
3.10	Top: S_WKD_OFF power output of a compressor in March 2020. Bottom: S_L CUSUM result showing detection of anomalies.	59
3.11	S_WKD_OFF power output of a gold plant in March 2020.	59
3.12	S_WKD_OFF power output of a fridge plant in March 2020.	60
3.13	Elbow plots showing the number of change points along varying penalties in the PELT method for the March 2020 S_WKD_OFF power data for: (a) a rock mill, (b) a mining compressor, (c) a gold plant, and (d) a fridge plant.	61
3.14	A snapshot of the memory usage on the machine that was used to implement change-point detection algorithms.	62
3.15	(a) S_WKD_OFF power output in March 2020 of a mining rock mill; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.	63
3.16	(a) S_WKD_OFF power output in March 2020 of a mining compressor; (b) change-point detection results along with the scaled power data from plot (a).	63
3.17	(a) S_WKD_OFF power output of a mining compressor from 13-25 March 2020; (b) change-point detection results along with the scaled power data.	64
3.18	(a) S_WKD_OFF power output in March 2020 of a gold plant; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.	64
3.19	(a) S_WKD_OFF power output in March 2020 of a fridge plant; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.	65
A.1	Probability density curve of arbitrary time series data with $\mu = 1.5$ and standard deviation $\sigma \approx 1$	76
B.1	Sibanye Gold Kloof SCADA of pumps and water lines [29].	77
C.1	Summer weekday standard time of use plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.	78
C.2	Summer weekday peak time of use plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.	79
C.3	S_WKD_STD standard deviation plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.	80
C.4	S_WKD_PEAK standard deviation plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.	81
D.1	Top: S_WKD_OFF power output of a compressor in 2020. Bottom: S_L CUSUM result.	82
D.2	Elbow plots showing the number of changepoints with varying penalties in the PELT method for the March 2020 S_WKD_STD power data of: (a) rock mill, (b) mining compressor, (c) gold plant and (d) fridge plant.	83

D.3	Elbow plots showing the number of changepoints with varying penalties in the PELT method for the March 2020 S_WKD_PEAK power data of: (a) rock mill, (b) mining compressor, (c) gold plant and (d) fridge plant.	84
D.4	(a) S_WKD_STD power output in March 2020 of a rock mill. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph. . . .	85
D.5	(a) S_WKD_STD power output in March 2020 of the mining compressor. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.	85
D.6	Area graph of the scaled compressor data with the change point detection results overlaid.	86
D.7	(a) S_WKD_STD power output in March 2020 of the gold plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph. . . .	86
D.8	(a) S_WKD_STD power output in March 2020 of the fridge plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph. . . .	87
D.9	(a) S_WKD_PEAK power output in March 2020 of a rock mill. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph. . . .	87
D.10	(a) S_WKD_PEAK power output in March 2020 of the mining compressor. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.	88
D.11	(a) S_WKD_PEAK power output in March 2020 of the gold plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph. . . .	88
D.12	(a) S_WKD_PEAK power output in March 2020 of the fridge plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.	89

List of Tables

1.1	CUSUM example where x_i is the assessment period data, y_i is the CUSUM, $i = \{1, 2, \dots, n\}$, n is the number of data points and $y_0 = 0$	10
1.2	CUSUM example where the positive (S_H) and negative (S_L) accumulation in standard deviation is determined.	12
1.3	Summary of equations for the log likelihood hypothesis test.	21
1.4	Table of important factors to consider when confirming that a valid change point was identified.	31
2.1	Eskom assignment of vacation and holidays in 2020-2021 [23].	36
2.2	Eskom's assignment of time of use periods between peak, standard and off-peak times.	38
3.1	Baseline statistical parameters for CUSUM-anomaly detection.	56
3.2	Summary of maximum change point limits for the BinSeg algorithm.	62
D.1	Average runtime of the BinSeg and PELT algorithms.	89

List of Algorithms

1	CUSUM	42
2	Penalty Determination	46
3	Binary Segmentation	46
4	Pruned Exact Linear Time	47

Table of abbreviations

AMI	Advanced Metering Infrastructure
AM&V	Advanced Measurement and Verification (also known as M&V 2.0)
AMOC	At Most One Change
ASHRAE	American Society of Heating, Refrigerating and Air-conditioning Engineers
BCP	Bayesian Change Point
BinSeg	Binary Segmentation
CPD	Change Point Detection
CUSUM	Cumulative Sum
EEM	Energy Efficiency Measure
IPMVP	International Performance Measurement and Verification Protocol
LRT	Likelihood Ratio Test
MAD	Median Absolute Deviation
M&V	Measurement and Verification
NRE	Non-Routine Event
PELT	Pruned Exact Linear Time
SCADA	Supervisory Control and Data Acquisition
S_WKD_OFF	Summer Weekday Off-peak
S_WKD_PEAK	Summer Weekday Peak
S_WKD_STD	Summer Weekday Standard
TOU	Time Of Use

Glossary

Changepoint	A package written in R that contains statistical change detection algorithms.
Function	Lines of code that perform tasks, similar to the functions found in Excel such as SUM and AVERAGE [1].
Module	A directory of files with code that perform specific task, typically written in Python [2].
Package	Similar to “Module”, packages are typically written in R.
Python	Programming language.
R	Language and program designed to visualise and statistically process data [3].

Chapter 1 Background and literature review

1.1 Preamble

This chapter provides background on the research topic. The literature review aims to identify critical areas that will benefit from further investigation. From this, the problem statement is formulated. Thereafter, the objectives of this study are discussed. Furthermore, the research in this chapter forms the basis upon which the methodology is carried out in Chapter 2.

1.2 Non-routine event detection in mines and application of M&V 2.0 methods

Energy efficiency measures (EEMs) are frequently implemented for use in mining systems [4] with the aim to either reduce electricity consumption, or improve the efficiency of, a system [5]. In order to evaluate whether or not any savings were achieved as a result of EEM implementation, the impact of any EEMs must be determined — a process which can be complex given that energy consumption is often influenced by external factors, such as climate. The procedure used in quantifying the success of an EEM is known as “measurement and verification” (M&V).

M&V methodology generally follows the International Performance for Measurement and Verification Protocol (IPMVP). To calculate energy savings, this standardised procedure is summarised by Equation 1.1 [5].

$$savings = baseline - assessment \pm adjustments \quad (1.1)$$

In this instance, the baseline represents the system’s energy consumption before the EEMs are implemented. The assessment period energy is the energy consumed while an energy efficiency project (EEM) is ongoing. As can be derived from Equation 1.1, savings are achieved when less energy is consumed in the latter period.

In Equation 1.1, adjustments are applied for numerous reasons. One such reason is that of a non-routine event (NRE), as characterised by unforeseen changes in the operating state of the system — such as by equipment failure. An unplanned-for event that has caused lower energy consumption cannot, thus, be claimed to be an energy-saving measure [5]. Conversely, if a system consumes more energy for reasons that could neither be foreseen nor prevented by the EEM, then the project should not be penalised as a result. An example of such an outcome is the presence of newly-installed equipment that was not present during the baseline period [6].

According to the literature [4][7], NREs are mostly detected and quantified through manual monitoring. In the context of a mine, this is a time-consuming procedure, as the individual who suspects the occurrence of an NRE must first contact mining personnel so as to retrieve the needed context relevant to the equipment in question. Furthermore, the classification of an event as an NRE is a subjective decision made by the individual performing the investigation. Therefore, a standardised technique for use in mines and similar environments would be an improvement [6]. Ideally, the system should be monitored closely so as to rapidly notice any unplanned changes. Industry has become more reliant on statistical analyses being developed in the advanced M&V (M&V 2.0) field [6]. In the mining industry in particular, this dependency is enforced by the scale at which mines continue to exist, with some shafts as deep as 3.8 km [8].

Figure 1.1 shows the layout of a gold mine in South Africa [9]. In this instance, complex and intricate electrical metering, piping systems, and networks of machinery that have been developed over the mine’s lifetime are present on site [10]. A breakdown in any one of these areas may influence another working area [11], making the manual monitoring process of every section, as well as the ability to rapidly notice anomalies, increasingly challenging.

Numerous power meters are in place to measure the mine’s energy consumption [13]. The capability of metering equipment has improved since the late 1970s, when data logging capabilities were developed. However, the widespread and global adoption of such technologies was only witnessed from 1990-onward [14]. It is possible to extract and store data from these power meters in half-hour intervals, a development known as advanced metering infrastructure (AMI) [4]. Thus, AMI’s proliferation has increased the resolution and volume of energy data that can be obtained. Aspects concerning resolution was improved upon given that AMI consistently logs data at frequent intervals (such as daily, hourly, or every second) [15]. This has created the need for more sophisticated techniques in order to process the large quantity of data now attainable. To address the evolved metering environment, M&V 2.0 has subsequently emerged over the last decade in response [5]. The principles of M&V 2.0 are the same as M&V, however there is greater focus on automated methods and data analysis in addressing the expansion of data now potentially available. For example, high-resolution data specifically require adequate preparation beyond simple outlier mitigation [16].

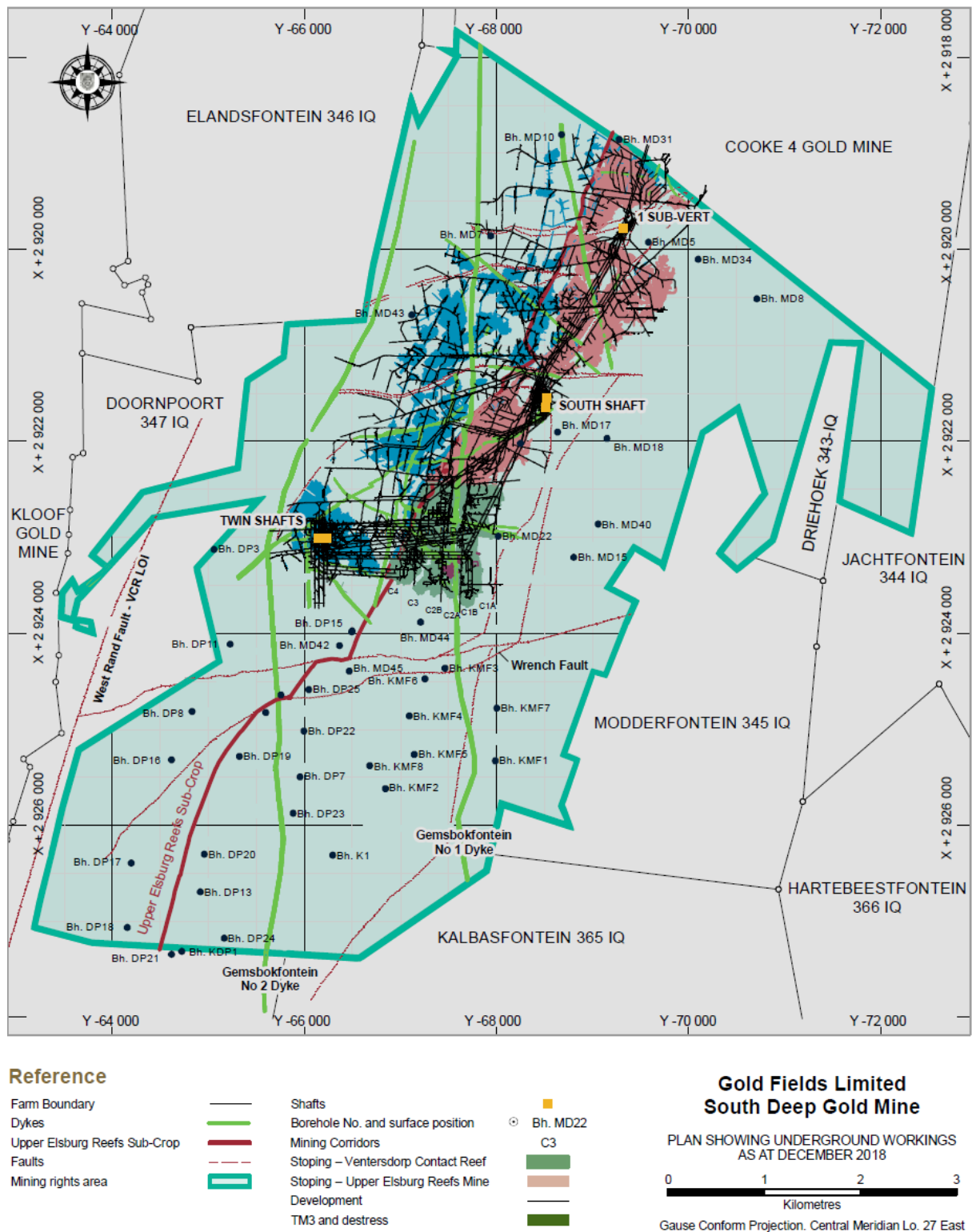


Figure 1.1: Areal schematic of a deep level gold mine showing a network of tunnels [12]

1.3 Data driven non-routine event detection applied in industry

1.3.1 Data quality and preparation for analysis

Data quality must be ensured when performing any data analysis. Data loss, measurement instrument reliability, and descriptions (such as of units) all fall under the umbrella of data quality [17]. Poor data quality will lead to incorrect conclusions, potentially resulting in undesirable and unforeseen repercussions [18]. A data quality assessment follows a series of checks in identifying inconsistencies — such as that concerning data loss and outliers. The points most relevant to the quality of energy data are given below [19]:

1. *Accuracy*: the data should be in the correct order of magnitude that is specified by the source. For example, if a pump has a power rating of 45 kW, then the logged data should not exceed this value.
2. *Reliability*: This refers to the trustworthiness of measurement equipment and whether they were calibrated correctly.
3. *Completeness*: If data loss occurred, measures must be put in place to account for, or estimate, the missing values where possible.
4. *Accessibility*: This defines how easy it is to obtain the data.
5. *Timeliness*: The period in which the data are relevant. Data that are out of date may skew results.

It is important to note that the points listed above also represent the reasons for the emergence of anomalies and may, as such, interfere with NRE investigations. For example, data loss may lead to misdiagnosed equipment failure. Therefore, rapid NRE detection process used in this study includes highlighting poor data quality. The next step in data preparation is statistical analysis. To expose patterns and trends, the data should be grouped according to descriptive metrics [20]. These are known as categorical variables that describe each data point. Examples include season, date and location.

Measurement resolution should be considered in parallel with categorical variables; for a monthly analysis, using hourly data points might contribute to noise and obscure change points [7]. In a study [21] where M&V 2.0 modeling strategies were tested on building energy consumption, it was found that subdividing the raw energy data into weekly periods improved the model's accuracy. Unique models were created for each time of week corresponding to when a specific load could be expected. Figure 1.2 has been extracted from a similar study and clearly shows how resampling to lower resolution can reveal trends [22].

Figure 1.2 shows load profiles of an office building (top), a furniture store (centre), and a bakery (bottom). The daily profiles were created by taking the average of each hour in the month — for example, the average of each measurement taken at 12:00h on Mondays, and so forth. Doing so revealed periods of higher and lower demand [22].

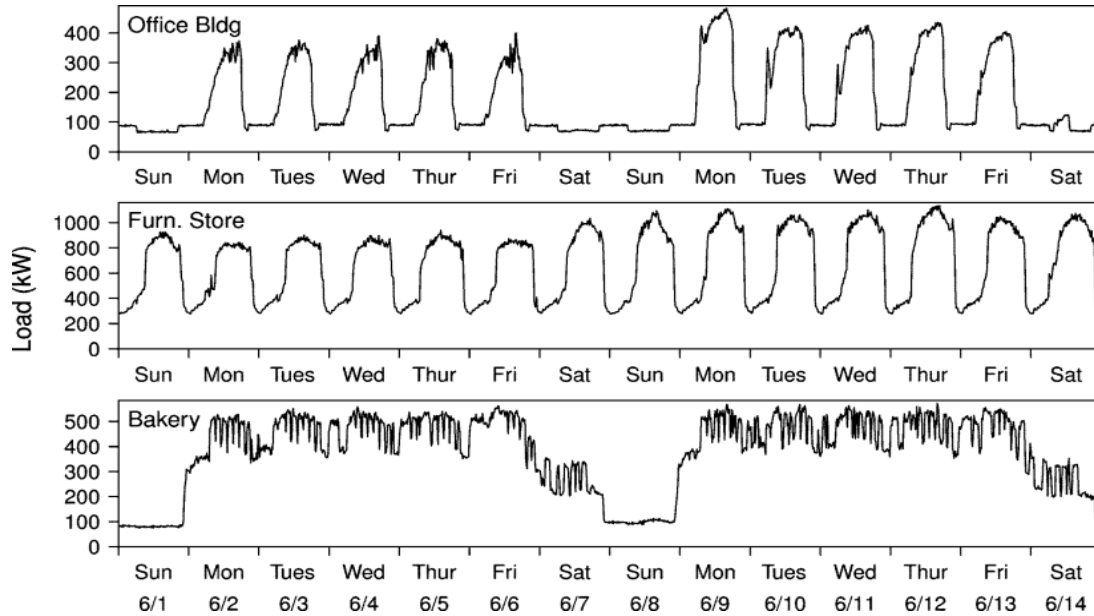


Figure 1.2: Daily demand trends of an office, furniture store, and bakery (top, centre, and bottom).

The results are shown in Figure 1.3. For each of the three plots, the apexes of the curve correlate to periods of high demand, whereas the corresponding troughs indicate periods of low demand. The dotted/dashed lines indicate the 15th, 50th and 85th percentiles of the high demand data. The dips appear on holidays.

Figure 1.3 shows that: in the office building, a peak demand of 300 to 500 MW can be expected; the furniture store’s peak demand ranges from 800 to 1000 MW; and that the bakery’s peak demand mostly falls between 500 and 600 MW.

This consistency enables higher model accuracy when both the time of day and the day of the week can be used as inputs [22]. Therefore, data preparation via the selection of the correct resolution and categorical variables is necessary in order to gain insights. In South Africa the national electricity provider, Eskom, has specific tariff schedules for high-demand facilities, such as ”megaflex” on certain mines [23]. Eskom applies a higher electricity cost during certain hours of the day — the tariff book [23] also shows specific tariffs for summer, winter, and Saturdays and Sundays. Furthermore, the tariff book also includes a calendar whereby holidays and vacations that fall on a weekday are classified as Saturdays or Sundays. Time of use periods, according to the Eskom tariffs, will affect operation in the mine and may produce similar trends to those shown in Figure 1.3 [24]. To further support the importance of categorical variables, [25] states that improved simulation accuracy can be achieved when they are accounted for.

In M&V, simulations and EEM performance are built with, or relative to, baseline data.

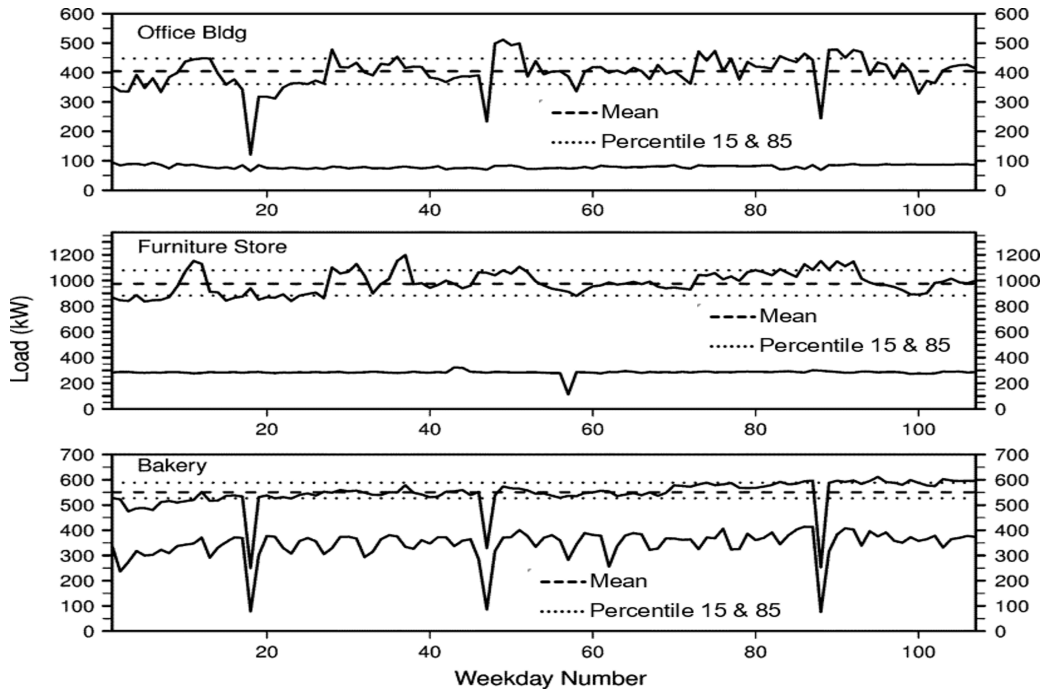


Figure 1.3: Demand profiles from Figure 1.2 split into peak (top line) and low (bottom line) demand periods.

For mine EEMs, energy data are recorded by an extensive network of power meters, with the data verified via a process of comparison to the monthly Eskom bill [24]. In a study [24] which developed energy budgets for deep-level mines in South Africa, it was found that the baseline data should be categorised according to the following periods: weekdays, Saturdays, Sundays, public holidays, the Christmas break, and Easter weekend. These periods provide baseline profiles that can be used to standardise the expected energy consumption on similar days across the assessment period. Few data quality checks are performed in South African mines other than by comparing the total measured energy consumption to the bills provided by Eskom [24]. When selecting baseline data for mine energy consumption, the following data preparation steps should be performed:

1. Identify large periods of “zero” consumption — if it is abnormal then remove the corresponding date and time from the data set [24] [26].
2. For missing data, remove the corresponding date and time from the data set [24].
3. Highlight data points that are more than two standard deviations from the mean [24].
4. Evaluate whether these points are outliers; if so, then remove them.
5. Determine if any known NREs occurred during the baseline period. If so, this data should be replaced with the average consumption during normal operation [24].

In summary, the following was drawn from literature. Data quality must be ensured by analysis and cleaning[17][18][19]. It is important to consider specific operational aspects of a

system when preparing the data [16]. This can be done by grouping the data into periods where constant conditions are expected [7][22][20]. Anomaly-detection algorithms may then more successfully locate abnormal data points [25]. In the IPMVP guide on NRE detection, it is suggested that monitoring changes in time-of-use periods is a possible way to quickly identify an NRE [21][27]. For mining energy data, this can be achieved by categorising the data according to Eskom time of use periods [23][24].

1.3.2 IPMVP-prescribed techniques for non-routine event detection

The International Performance Measurement and Verification Protocol (IPMVP) prescribes numerous options for retrospective NRE detection. It is recommended that these methods be applied periodically and as soon as possible once an EEM has been implemented, as it will ensure early detection and mitigation of NREs. The most simple approach to identifying NREs is to monitor areas that are expected to remain constant. This method is known as “identifying changes in static factors” [27]. The IPMVP core concept defines static factors below:

***“Static factors:** those characteristics of a facility which affect energy consumption and demand within the defined measurement boundary that are not expected to change and were therefore not included as independent variables in the energy model. If they change, Non-Routine Adjustments need to be calculated to account for these changes.... Note: Those characteristics may include fixed, environmental, operational and maintenance characteristics.”*[27]

An example of a static factor is an increase in temperature expected during summer — periods of abnormally cold days may affect cooling equipment that use temperature for the setpoint. Another example is equipment installation/removal or their failure which may directly affect the EEM. Static factors may vary with time, weather, or other independent variables, such as production [7]. Most commonly, regular feedback is required from site personnel. By involving site personnel, the precedent is set for energy efficiency management and will aid in the continuation of the EEM [27].

The process of solely relying on reports from site personnel may not be sufficient. The IPMVP application guide strongly suggests the use of data analysis as an additional metric. Energy data should be investigated to verify the personnel’s findings [27], thus requiring data logging equipment such as flow and power meters. Many mines rely on SCADA systems — like the one in Figure B.1 — that allow for remote data gathering [28] [29]. This method is suggested for EEMs that do not use a mathematical model in order to predict energy consumption. This method is particularly advantageous when facilities have high level and sub-level metering. The sub-level metering is installed on energy consuming equipment and makes it possible to pinpoint the location of an NRE [27].

As previously stated, AMI produces large quantities of data, rendering attempts at manual

monitoring unfeasible. More efficient strategies, such as visualisation tools, must thus be used. In this regard, the IPMVP suggests the use of: heats maps, surface charts, time-series plots, and overlays of daily load profiles. However, all of the visual techniques are limited by the number of data points and the relevant time period. According to literature [30], visual techniques are limited to two dimensions, few variables, and few observations. Too many points of observation rapidly clutter the plots and prohibit conclusions to be drawn. Given that manual monitoring is already proven to be too slow and insufficient, this study will focus on other more-automated approaches.

SCADA applications have alarm systems. These alarms are connected to sensors that can detect “on/off” or if equipment exceeds any preset limits [31]. This method of outlier detection can be based on thresholds or statistical standard deviation. According to ASHRAE guideline 14, acceptable thresholds are 10 % of the maximum and minimum [32]. However, the IPMVP recommends establishing an expected range. For normally distributed data, the sample standard deviation can be used to define limits [27].

In normally distributed data, many of the observations are near the mean, the distribution curve is bell shaped, and data are dispersed symmetrically about the mean [33]. This is useful given that the z-score method can then be applied to present the data in accordance to the number of standard deviations (σ) from the mean [34]. An example of a normal distribution with the z-score on the x-axis is shown in Figure 1.4. From the figure, this representation provides insights independent of units or magnitude.

The z-score is the standardised distance between each point and the mean [34]. In Equation 1.2, Z_i is the distance between point x_i and the mean in terms of number of standard deviations (σ).

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (1.2)$$

Standard deviation (σ) is the square root of a data set’s variance (σ^2). Variance is the distance between each data point and the mean as shown in Equation 1.3. Where n = number of data points, μ is the mean and x_i is each individual data point. The difference, $x_i - \mu$, is squared to produce a positive value.

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1} \quad (1.3)$$

Simply put, the variance gives an indication of the spread of the data [35]. A large variance suggests that many of the data points are far from the mean.

It is important to use the standard deviation of either the baseline period or from the data used to create the model [27]. The IPMVP suggests that outliers are those data points considered

to be more than three standard deviations away from the mean, as shown in Figure 1.4 [27].

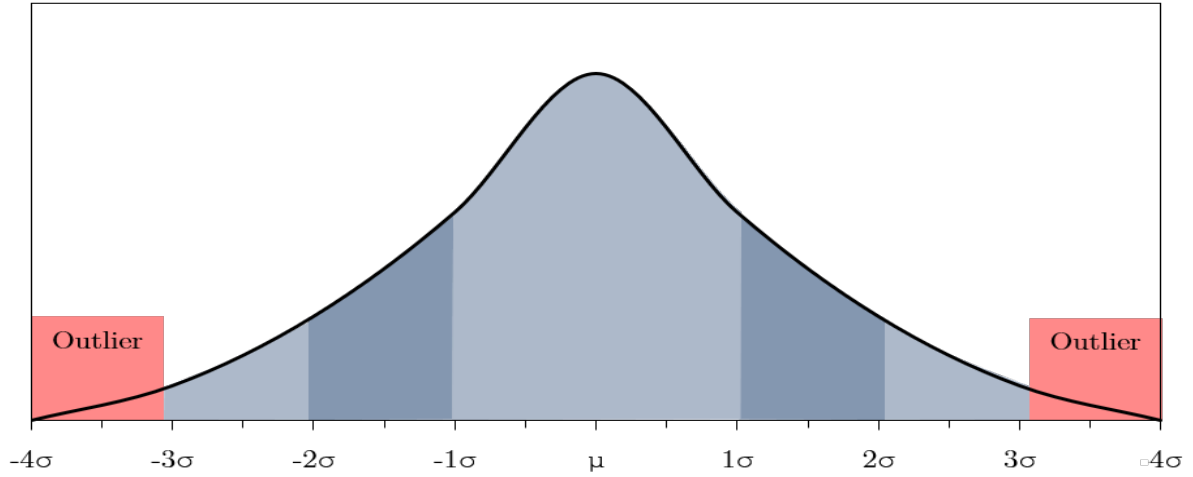


Figure 1.4: Example of normally-distributed data showing outliers at more than three standard deviations from the mean.

However, literature [24] suggests that the threshold should be reduced to two standard deviations for mining-energy data, given that it is a more strict criterion necessary for the magnitude of demand from mining equipment. This strategy requires baseline period data from “consistent operations” [27], indicating that the data should not include days when the equipment had been turned off, or weekends/holidays where no production occurred. Figure 1.5 shows the data from Figure 1.4, along with the limits as prescribed by the IPMVP. These limits are shown in the top legend as: maximum plus 10%, minimum less 10%, and $\pm 3\sigma$. The baseline data on the left hand side of Figure 1.5 was randomly generated with a mean of zero and standard deviation of one. The assessment period shows that the arbitrary data exceed the upper threshold towards the end of the time period.

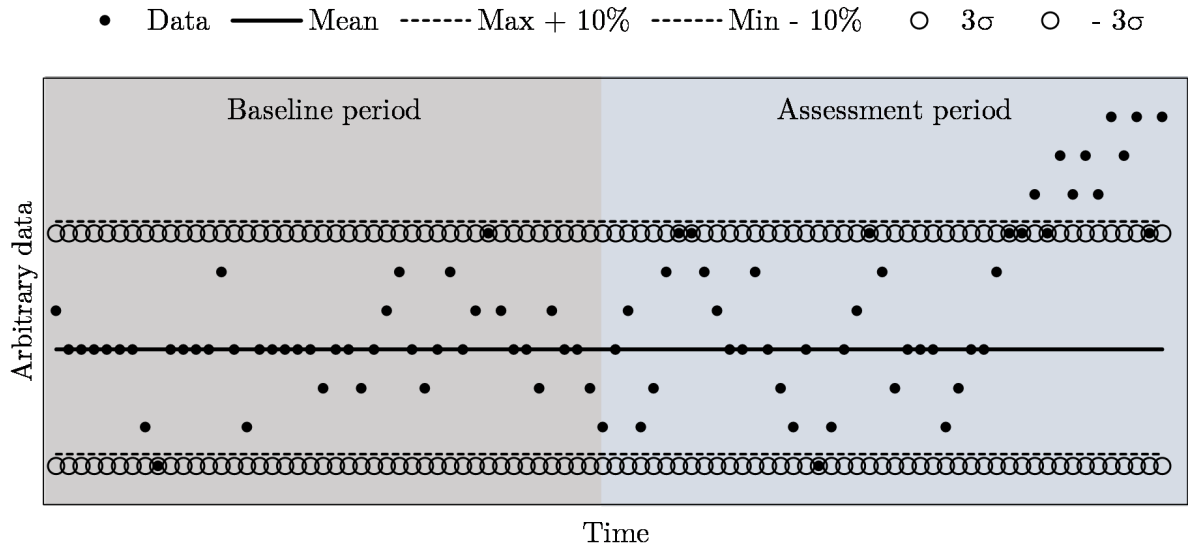


Figure 1.5: Example of normally distributed data along with 10% and standard deviation test metrics.

The advantage of a normally-distributed data set is that outliers can be spotted with simple

thresholds set up via the SCADA [27]. Figure 1.5 shows that the assessment period data exceed the limit towards the end. The system must *exceed* the limits to activate an alarm. Unless more-strict thresholds are set, this method is only reactive, and not proactive. A more proactive approach would be to monitor the rate of change in the data [27]. Figure 1.6 highlights where the increased rate of change may indicate a warning sign of the system approaching a limit.

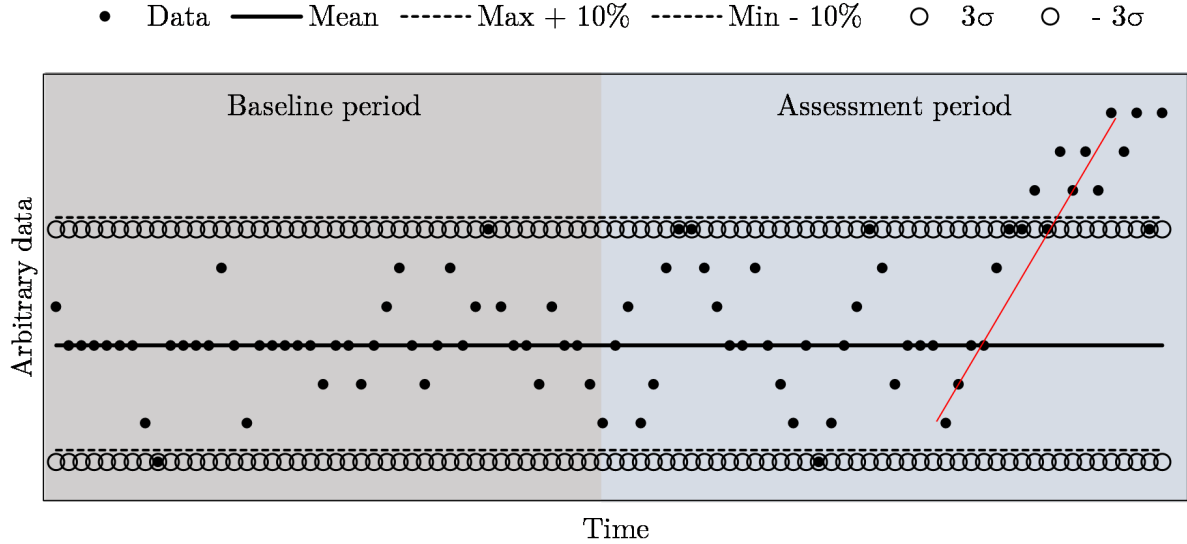


Figure 1.6: Repeated example of normally-distributed data along with 10% and standard deviation test metrics

The IPMVP suggests tracking this with cumulative sums [27]. As suggested by the name, a cumulative sum (CUSUM) chart plots the accumulation of values. That is, the sum of the values from start to present (if the data are time-series). The calculation is demonstrated in Table 1.1. CUSUM charts are used to identify change [36].

Table 1.1: CUSUM example where x_i is the assessment period data, y_i is the CUSUM, $i = \{1, 2, \dots, n\}$, n is the number of data points and $y_0 = 0$.

i	x_i	$y_i = y_{(i-1)} + x_i$
1	3	3
2	3	6
3	2	8
n	$\vdots$	$\vdots$

Figure 1.7 demonstrates this principle with the same data from Figure 1.6. Keeping in mind that the mean is zero, some of the data points are negative and, therefore, Figure 1.7 initially does not show accumulation. Towards the end, it is clear that the trend had changed, corresponding to the system exceeding threshold. Changes in a CUSUM chart may consist of a change in slope or direction [36].

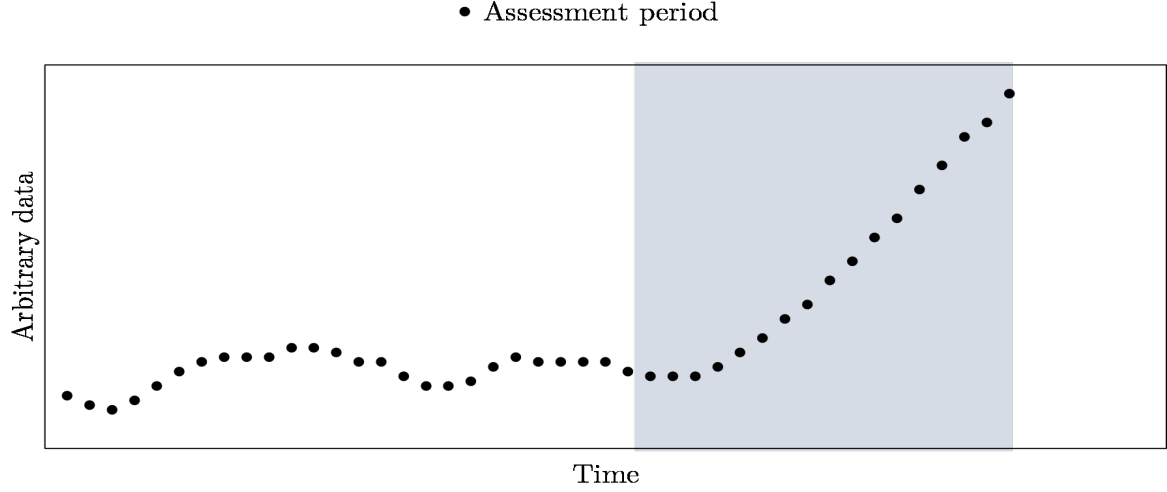


Figure 1.7: Example of a CUSUM plot showing acceleration and, therefore, higher rate of accumulation.

More formally, the CUSUM calculation is defined by Equations 1.4 and 1.5 [36]. Variables S_H and S_L represent positive and negative accumulation, respectively. The optional term ω is a parameter for adjusting the strictness of the test [37].

$$S_{H(i+1)} = \max(0, S_{Hi} + x_i - \omega) \quad (1.4)$$

$$S_{L(i+1)} = \min(0, S_{Li} + x_i - \omega) \quad (1.5)$$

This method does not necessarily require visual monitoring as it can be combined with the threshold technique (for example, the CUSUM of standard deviations can be used). To generate the values in Table 1.2, x_i in the CUSUM formulas above was replaced with Z_i from Equation 1.2.

Table 1.2 shows that S_H reaches 3σ when $i = 28$ — the data have a standard deviation of one and, therefore, when $Z_i > 3$ the distance between the data points and the mean may be too big. This is an indication that the system has possibly reached the upper threshold. The S_L column shows that the system had already exceeded the lower threshold (-3σ).

Table 1.2: CUSUM example where the positive (S_H) and negative (S_L) accumulation in standard deviation is determined.

i	Z_i	S_H	S_L
25	-1	0	-7
26	0	0	-7
27	1	1	-6
28	2	3	-4
29	0	3	-4
30	-1	2	-5

The points in Table 1.2 are highlighted in Figure 1.8. Upon visual inspection, it is clear that the positive change as induced via CUSUM provides early warning that the system may exceed its limits, but perhaps too early. Figure 1.8 implies that the system exceeded three standard deviations in the middle of the time period. However, when comparing this with the data in Figure 1.6, it is apparent that the system only exceeded the threshold near the end of the time period.

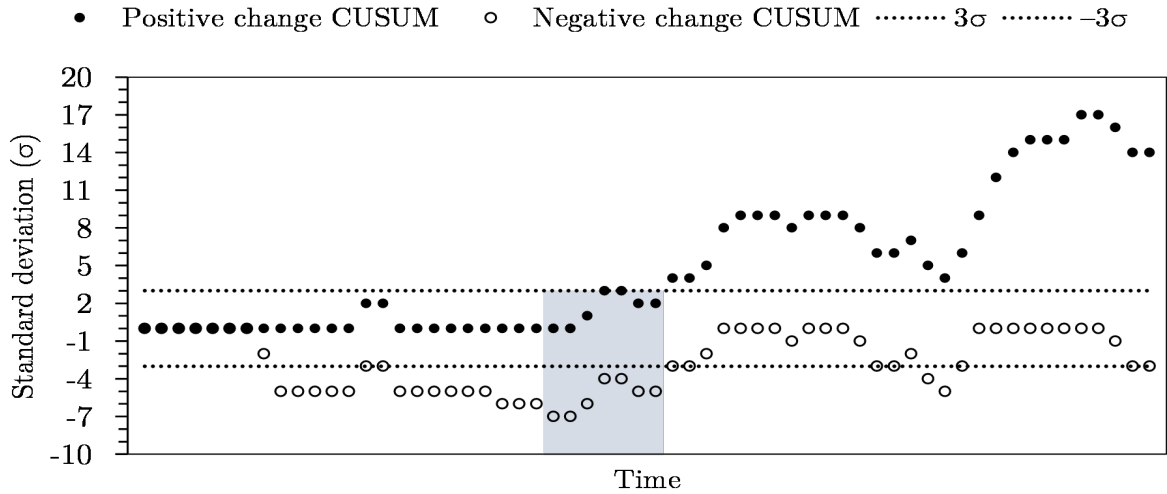


Figure 1.8: CUSUM of the distance between each point and the mean divided by the standard deviation ($\sigma = 1$). The highlighted area relates to $S_{H/L}$ in Table 1.2

In Figure 1.8, the negative change CUSUM shows that the system exceeds the lower (-3σ) limit near the beginning. However, when comparing this to Figure 1.6, it is clear that the data points which lie slightly below the mean decrease S_L to prematurely exceed the set boundary — the converse is true for S_H . This causes the data in Table 1.2 to wrongfully indicate out of bounds activity.

The importance of ω is now evident as it can be used to decrease the sensitivity of the CUSUM calculation [37]. According to McNeese et al. [37] [38], ω is typically set to 0.5σ . After

some iteration, in Equations 1.4 and 1.5, ω was set to $\pm 3\sigma$ to reduce the sensitivity. In summary, Equations 1.2 to 1.8 are repeated below with values assigned to all variables:

$$Z_i = \frac{x_i - \mu}{\sigma} \text{ where } \mu = 0 \text{ and } \sigma = 1 \quad (1.6)$$

$$S_{H(i+1)} = \max(0, S_{Hi} + Z_i - \omega) \text{ where } \omega = 3\sigma \quad (1.7)$$

$$S_{L(i+1)} = \min(0, S_{Li} + Z_i - \omega) \text{ where } \omega = -3\sigma \quad (1.8)$$

These equations were used to generate the values in Figure 1.9. For ease of comparison, Figure 1.5 is repeated in Figure 1.10.

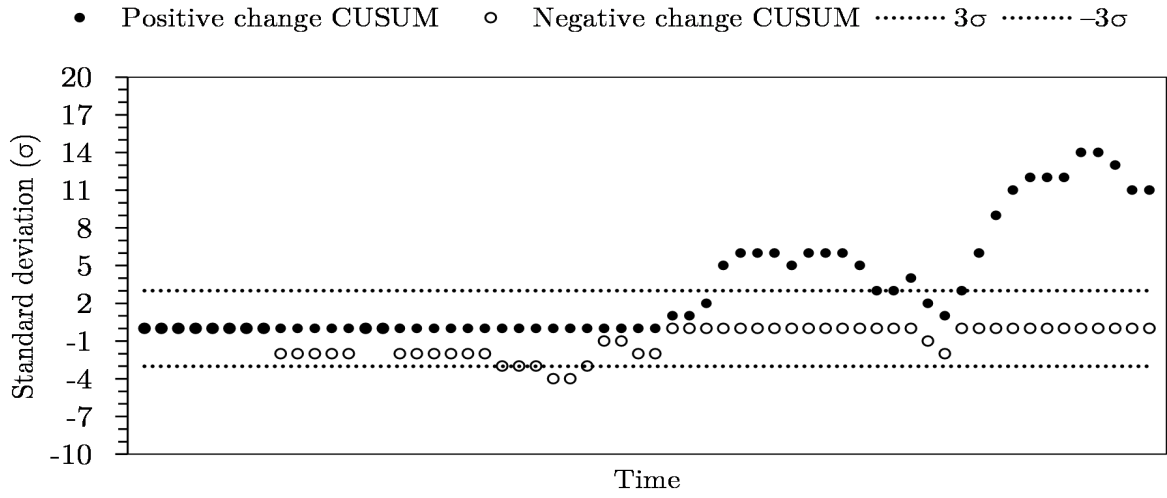


Figure 1.9: Positive (S_H) and negative (S_L) accumulation of the standard deviation (Z) with sensitivity set to three standard deviations, $\omega = \pm 3\sigma$.

Figure 1.9 shows that the system exceeds the upper bound near the end of the time period. This coincides with the data in Figure 1.10, where the system increased rapidly towards the end. Figure 1.9 also shows that the use of standard deviation CUSUM can provide an early warning. The upper limit in Figure 1.9 was crossed before the system went out of bounds in Figure 1.10

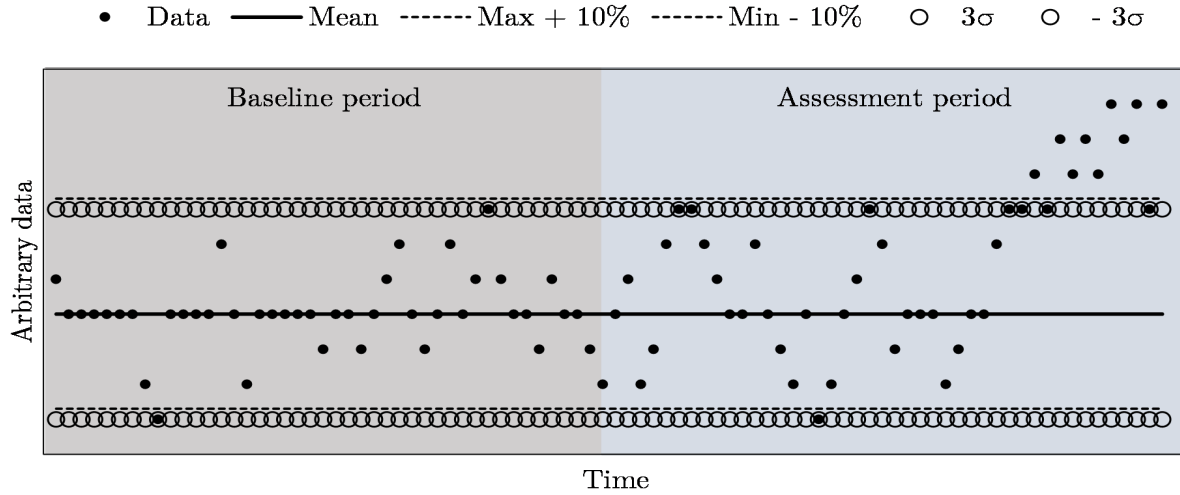


Figure 1.10: Example of normally-distributed data along with 10 % and standard deviation test metrics.

The following was derived from literature. An IPMVP approach to NRE detection makes use of static factors [27]. These are processes or conditions that are expected to remain constant for the foreseeable future [7]. Currently, in large scale industry static factors are generally monitored manually [27]. To locate outliers, statistical thresholds can be set up on visual monitoring (SCADA) systems[28][29][32][31]. However, visual monitoring has space limitations and few data streams can be plotted to prevent clutter[30]. According to Benoit and Petitgas [39], CUSUM calculations hold the potential to provide early warning signs; these may indicate if the system is approaching its limits. This is one of the methods applied in advanced M&V (AM&V or M&V 2.0) which has emerged over the past decade [4]. The movement was facilitated by the development of AMI and is characterised by:

1. High resolution data logging (by the hour, or more frequently).
2. Using statistical or analytical methods to process data for rapid feedback and calculations.

The following section discusses research surrounding algorithmic NRE detection.

1.3.3 Algorithms for non-routine event detection

The analytical approach to NRE detection is an up-and-coming area of research. Currently, the approach mostly consists of manual feedback and a review of CUSUM charts [4]. In Section 1.2, Granderson et al. [4] and Touzani et al. [6] confirmed that manual feedback is inconsistent and not fast enough. To improve on this, analytical methods can standardise NRE detection, making it a straightforward and consistent procedure [4]. As a possible solution, Touzani et al. [6] suggest investigating an analytical method based on statistical properties. This method aims to detect fluctuations in data trends [40].

The application of statistical-change detection in building energy analysis was investigated

by Touzani et al. [6]. Statistical change can be defined as “the point at which the statistical properties of a sequence of observations change” [41]. A sequence is an ordered list that contains observations [42], with observations, in turn, referring to the data points. Statistical properties are descriptors such as mean and variance [43]. Figure 1.11 shows an example of statistical change using arbitrary data. The left- and right-hand side figures show change in mean and variance respectively.

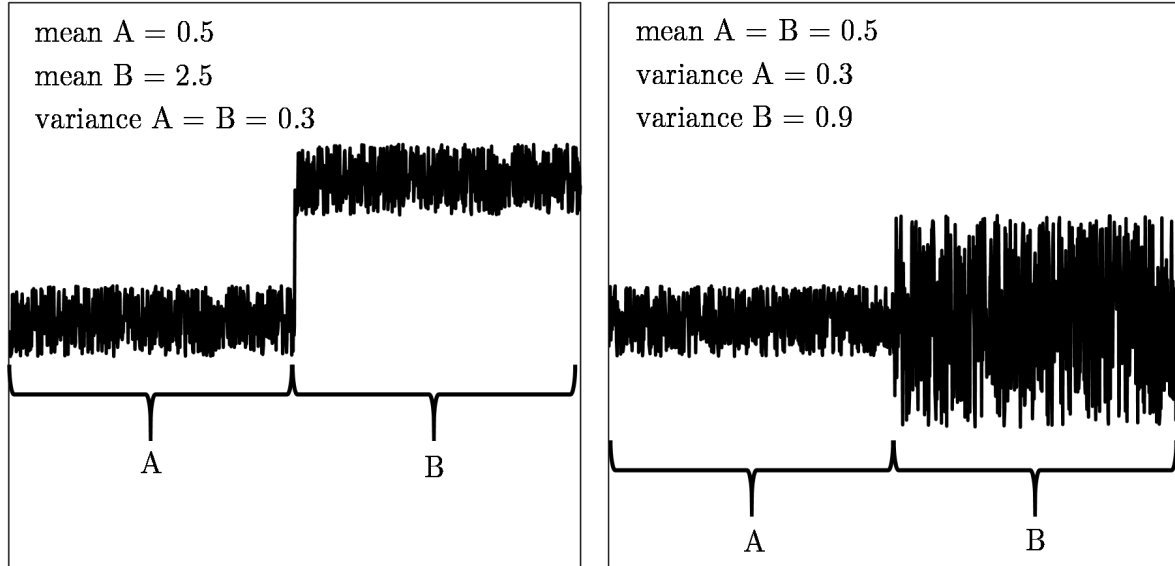


Figure 1.11: Example of statistical property changes showing change in mean (left) and variance (right).

One study [6] used a change point detection algorithm to locate NREs in various EEMs applied to building energy. The study was implemented in R, a programming language and platform readily used for statistical analysis and visualisation — the platform is also open source, meaning that no licence is required for its use and that anyone can contribute useful solutions via the platform. The research in this dissertation will be conducted in R. The algorithm tested by Touzani et al. [6] successfully identified 75 % of the change points in the study’s energy data. This study recommends further investigation of change point detection (CPD) algorithms so as to locate anomalous energy usage and equipment failure.

Numerous CPD algorithms exist. The choice of which to apply to mining time-series data is discussed in this section. According to Killick and Eckley [41], the differentiation of algorithms should be based on how many change points are expected. Many algorithms are excellent when there is only one change to be found, with a study conducted by Otto [44] also having compared the most common algorithms for change point detection. The following algorithms, as used in R, are not suitable when multiple change points are expected [44]:

1. *AMOC* (at most one change), as in the name, finds the location of a single change point.
2. *BCP* (bayesian analysis of change points) uses probability to divide the data into blocks. BCP is an iterative process whereby different groupings are tested. Each iteration is reflected upon in order to determine which combination of blocks have the smallest variance.

However, this algorithm was designed for normally-distributed data and often only detects one change.

3. *Segmented regression* can detect more than one change point, but the number of change points must be known beforehand.
4. A *regression tree* splits the data in two and calculates the deviation from the mean in each segment. It moves the split point until minimum deviation in each segment is obtained. Then it splits the two new segments and repeats the previous step. Knowledge of the minimum number of change points is required, often too many change points are found.

The package **changepoint** contains algorithms that showed promise when multiple change points are expected [44]. Studies [6][44] have suggested that the *pruned exact linear time* (PELT) algorithm be investigated — available in **changepoint**. Before delving into the principles of PELT, it is important to describe change points, terminology, and algorithms in general.

In time-series analysis, the change point refers to a point in time where the properties of the data change. Properties include mean, variance, magnitude and gradient, among others. The fundamental assumption of some change point algorithms is that the properties remain constant both before and after the change point [45]. To define a change point mathematically, consider the following definitions:

$y_{1:n}$ is an ordered sequence such as,

$$y_{1:n} = (y_1, y_2, \dots, y_n) \quad (1.9)$$

τ is a location in time such that,

$$\tau \in \{1, 2, \dots, n - 1\} \quad (1.10)$$

Let y_a and y_b be two consecutive sets in $y_{1:n}$ such as:

$$y_a = (y_1, \dots, y_\tau) \text{ and } y_b = (y_{\tau+1}, \dots, y_n) \quad (1.11)$$

A change point is present at location τ when the statistical properties of y_a differ from y_b . The concept is presented visually in Figure 1.12. For the case of change in mean, Equations 1.9 to 1.11 can be summarised by Equation 1.12 [41].

$$\mu_t = \begin{cases} \mu_1 & \text{if } 1 \leq t \leq \tau_1 \\ \vdots & \vdots \\ \mu_{m+1} & \text{if } \tau_m < t \leq \tau_{m+1} = n \end{cases} \quad (1.12)$$

where m is the number of change-points and μ_1 is the mean in segment 1. Notice in Figure 1.12 that when there is only one change point, then $m = 1$, but there are two segments and thus μ_1

and μ_2 . This explains the last condition in Equation 1.12, where μ_2 is described as the mean of the segment after τ_m , the last change point. In this representation of change points, the last data point (n) in the sequence is denoted as change point τ_{m+1} .

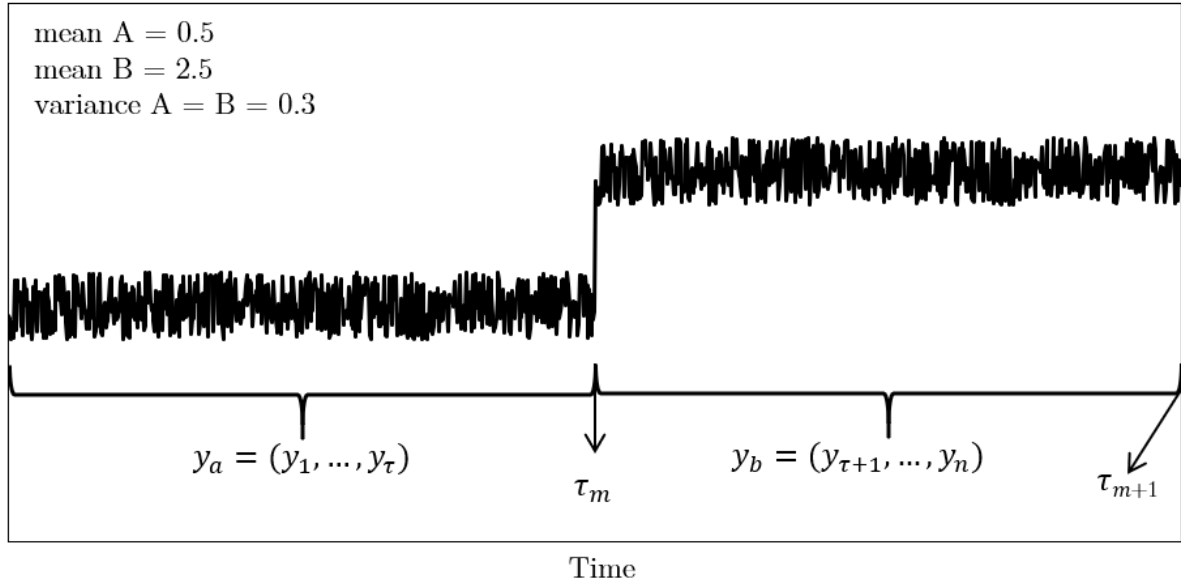


Figure 1.12: Statistical change point at time τ , showing change in mean of arbitrary data.

Change point detection can be applied either online or offline. When applying offline change detection, the entire data set is evaluated to find changes retrospectively. Online methods aim to identify changes as new data are received. How rapid an online algorithm generates results depends on the type of data. It is important to note that “online” should not be interpreted as “immediate”. Online change-point detection is a novel field that requires highly complex algorithms [46]. Therefore, this study will investigate the application of offline models.

An offline model does, nonetheless, have the potential to locate change points rapidly enough. Consider the current time frame of NRE detection through manual monitoring; it involves surveys, interviews, and physical audits of the equipment. An offline model can be run daily and is able to analyse the previous day’s data. As such, the aim is to improve the current method in a manner that is easy to understand and implement.

The main points drawn from this research segment are as follows. Manual monitoring of EEMs are not sufficient [4][6]. It is suggested that an automated approach be investigated [40].

This study will implement offline algorithms capable of locating more than one change point. However, to demonstrate the principles of multiple change point detection, it is necessary to explain the fundamentals of single change point detection.

1.3.4 Change-point detection theory

This section will describe the CPD algorithms tested throughout this study. As an introduction, the mathematical description of single change detection is discussed. On a fundamental level, the algorithms in `changepoint` determine τ with log likelihood test statistics. This is different from probability, which is defined in Appendix A. Although, it is important to first determine what characteristics of the data are required.

Killick et al. [41], recommends scaling the data. It is assumed that a change in mean may be present in the data. Therefore it is uncertain what the standard deviation (σ) would be for each segment. Scaling ensures that $\sigma \approx 1$, and it is sensible to scale with respect to the median [47]. This can be done with the median absolute deviation (MAD) scaling method shown in Equation 1.13 [48].

$$x_{i-scaled} = \frac{(x_i - median)}{median(|x_i - median|)} \quad (1.13)$$

The MAD is an indicator of variability because it involves the distance between each data point and the median [49]. As can be derived from Equation 1.13, when most data points are near the median, the MAD will be low. A high MAD indicates sporadic data because many of the data points are located far from the median. Now consider the likelihood function given by Equation 1.14, with $\sigma = 1$ substituted. This states what the likelihood is that the mean of the data set is μ , given the observed data point x_i .

$$L(\mu|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(x_i - \mu)^2}{2\sigma^2} \right] \quad (1.14)$$

The point x_i will be part of a sequence $x = (x_1, x_2, \dots, x_n)$. The standard deviation and mean can be inferred from a sample. With the assumption that samples are independent and identically distributed, Equation 1.15 shows the likelihood of μ given x_1 and x_2 [50].

$$L(\mu|x_1 \text{ and } x_2) = L(\mu|x_1) \times L(\mu|x_2) \quad (1.15)$$

To determine the likelihood of μ given a sequence $x_{1:n}$, Equation 1.15 can be linearised by taking the log on both sides. The linearisation steps are explained in literature [50] and results in Equation 1.16 (continuing with the assumption that $\sigma = 1$).

$$\log[L(\mu|x_{1:n})] = -\frac{1}{2}n \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \quad (1.16)$$

Given the arbitrary values for $x_{i:n}$ shown in Figure 1.13, Equation 1.16 is plotted in Figure 1.14 with a range of possible values of μ . Upon inspecting the data, it was estimated that the mean may lie somewhere between the values of 0 and 3.0.

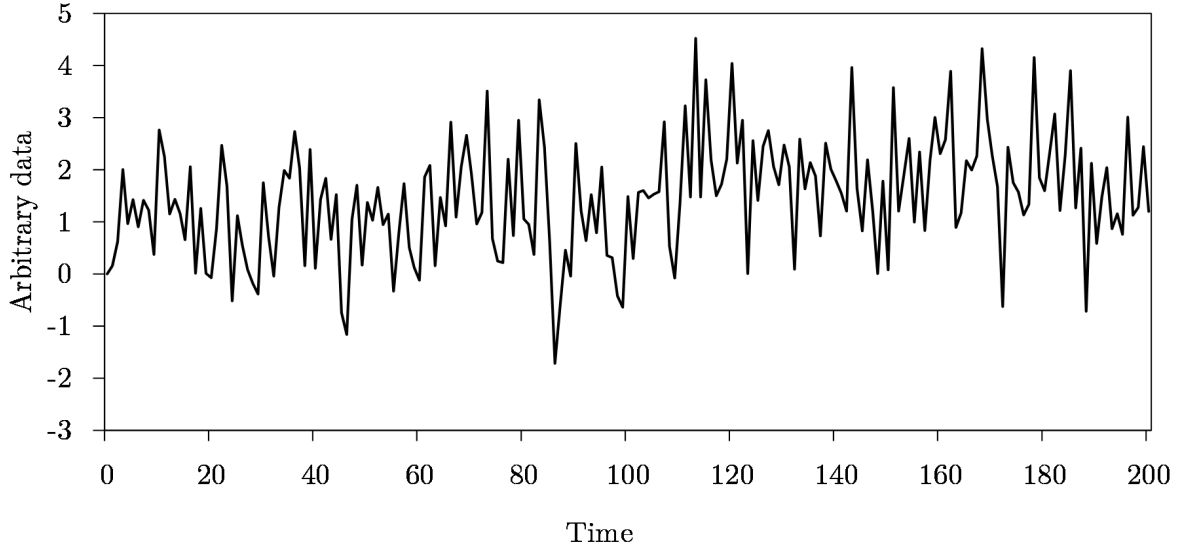


Figure 1.13: Arbitrary time series data with standard deviation $\sigma \approx 1$.

Figure 1.14 shows μ as the independent variable on the x-axis. The maximum point on the curve implies that the mean of the arbitrary data set is most likely ≈ 1.5 .

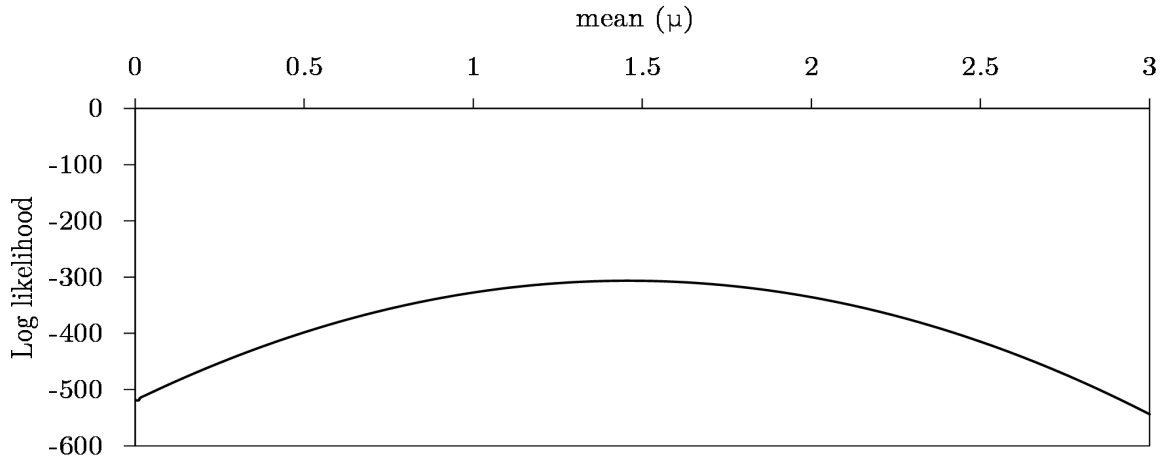


Figure 1.14: Plot of log likelihood for the mean of the data given in Figure 1.13.

The next building block in probabilistic change detection is the likelihood ratio test [50] which measures how different the likelihood of an occurrence is from the null hypothesis [51]. Continuing with the example data in Figure 1.13, let the null hypothesis be $\mu_0 = 1$. When substituting $\mu = 1$ into Equation 1.16 (or by tracing the curve in Figure 1.14) then likelihood of μ_0 given the data $x_{1:n}$ is:

$$\log[L(\mu_0|x_{1:n})] = -328 \quad (1.17)$$

Refer to Figure 1.15, which shows the likelihood of the null hypothesis (Equation 1.17). The curve was plotted with the data from Figure 1.13 as x_i by ranging μ from 0.5 to 2.0 in Equation 1.16. When the sequence $x_{1:n}$ is observed, the alternative hypothesis is created (H_1), stating that the maximum likelihood is -305 when $\mu_1 = 1.5$.

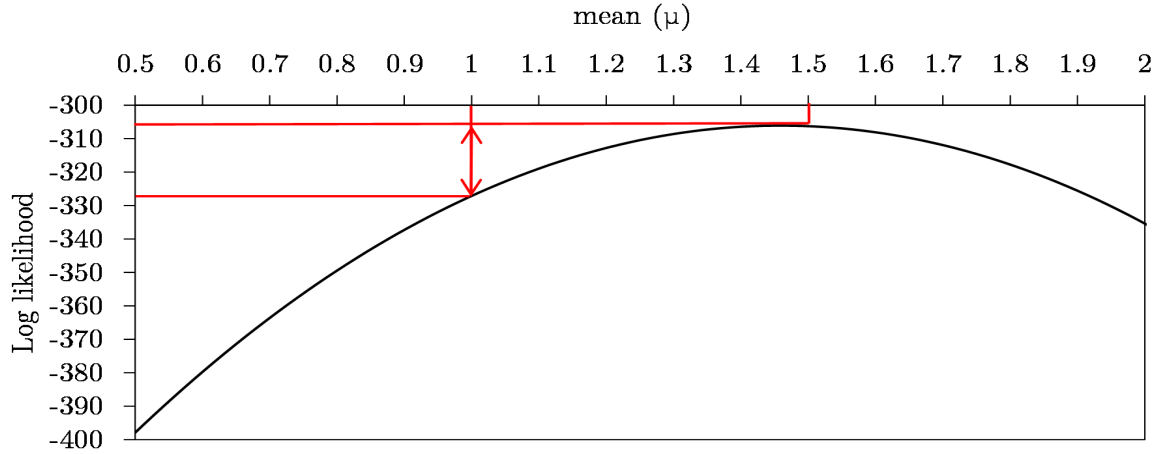


Figure 1.15: Log likelihood plot showing the likelihood of the null hypothesis ($\mu = 1$ and its vertical distance from the alternative hypothesis).

The difference in the likelihood of H_0 and H_1 is known as the likelihood ratio test (LRT). This is shown in Equation 1.18 [50].

$$LR = \frac{L(\mu_1|x_{1:n})}{L(\mu_0|x_{1:n})} \quad (1.18)$$

Equation 1.18 above can be linearised by taking the log of each side. This will then achieve the expression shown in Equation 1.19.

$$\log[LR] = 2(\log[L(\mu_1|x_{1:n})] - \log[L(\mu_0|x_{1:n})]) \quad (1.19)$$

Figure 1.15 shows the LRT to approximate the mean of a single data set. To translate the use of LRTs to change point detection, it is applied to more than one data set [41]. Consider the example data in Figure 1.13. Splitting the data at time τ would produce two segments, $x_{1:\tau}$ and $x_{\tau:n}$. For this example, τ was set to split the data midway. Now let H_0 be the hypothesis that there is no change point (solid line) and H_1 is the hypothesis that there is a single change point (dotted and dashed lines). The test is summarised in Table 1.3 and plotted in Figure 1.16.

Table 1.3: Summary of equations for the log likelihood hypothesis test.

Hypothesis	Data set	Equation 1	Equation 2
H_0 no change point	$x_{1:n}$	$\log[L(\mu_0 x_{1:n})]$	
H_1 one change point	$x_{1:\tau}$ & $x_{\tau:n}$	$\log[L(\mu_1 x_{1:\tau})]$	$\log[L(\mu_2 x_{\tau:n})]$

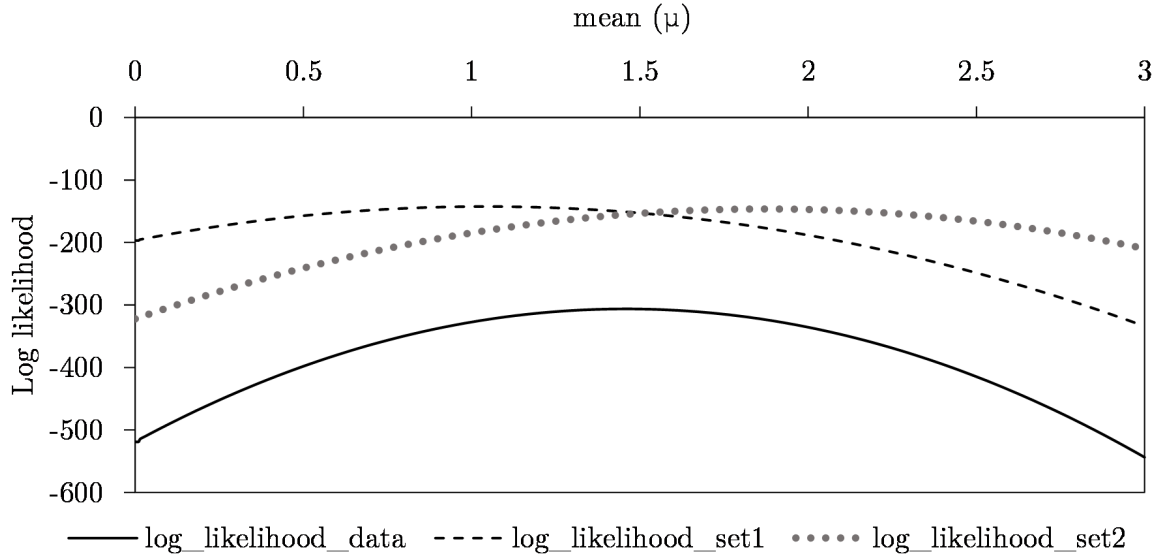


Figure 1.16: Log likelihood curves for two sets of the data compared to that of the entire data set.

According to Killick et al. [41], the maximum likelihoods of the alternative hypothesis corresponds to the mean of each segment. Therefore, let ML be the maximum likelihood:

$$ML(x_{1:n}) = \max \log[L(\mu|x_{1:n})] \quad (1.20)$$

Then the likelihood ratio in Equation 1.19 can be written as,

$$\log[LR] = 2[ML(x_{1:\tau}) + ML(x_{\tau:n}) - ML(x_{1:n})] \quad (1.21)$$

A change point at time τ occurred if Equation 1.21 is larger than some test metric (or cost c) [41]. For the case of one change point and small data sets, Killick et al. [41] suggests ranging τ from 1.0 to n . This incrementally splits the data into segments x_0 & $x_{1:n}$, $x_{0:1}$ & $x_{2:n}$, ... $x_{1:10}$ & $x_{11:n}$, etc. Then, by iteratively calculating values for Equation 1.15, we can plot Figure 1.17. The maximum point on this curve shows in which two segments the respective means (μ_1 and μ_2) are the most different from the mean of the entire data set (μ_0). At this point, the null hypothesis is rejected and a change point exists at τ . Figure 1.17 has a maximum when $\tau = 109$. This indicates a change in the mean in Figure 1.13 at the 109th data point.

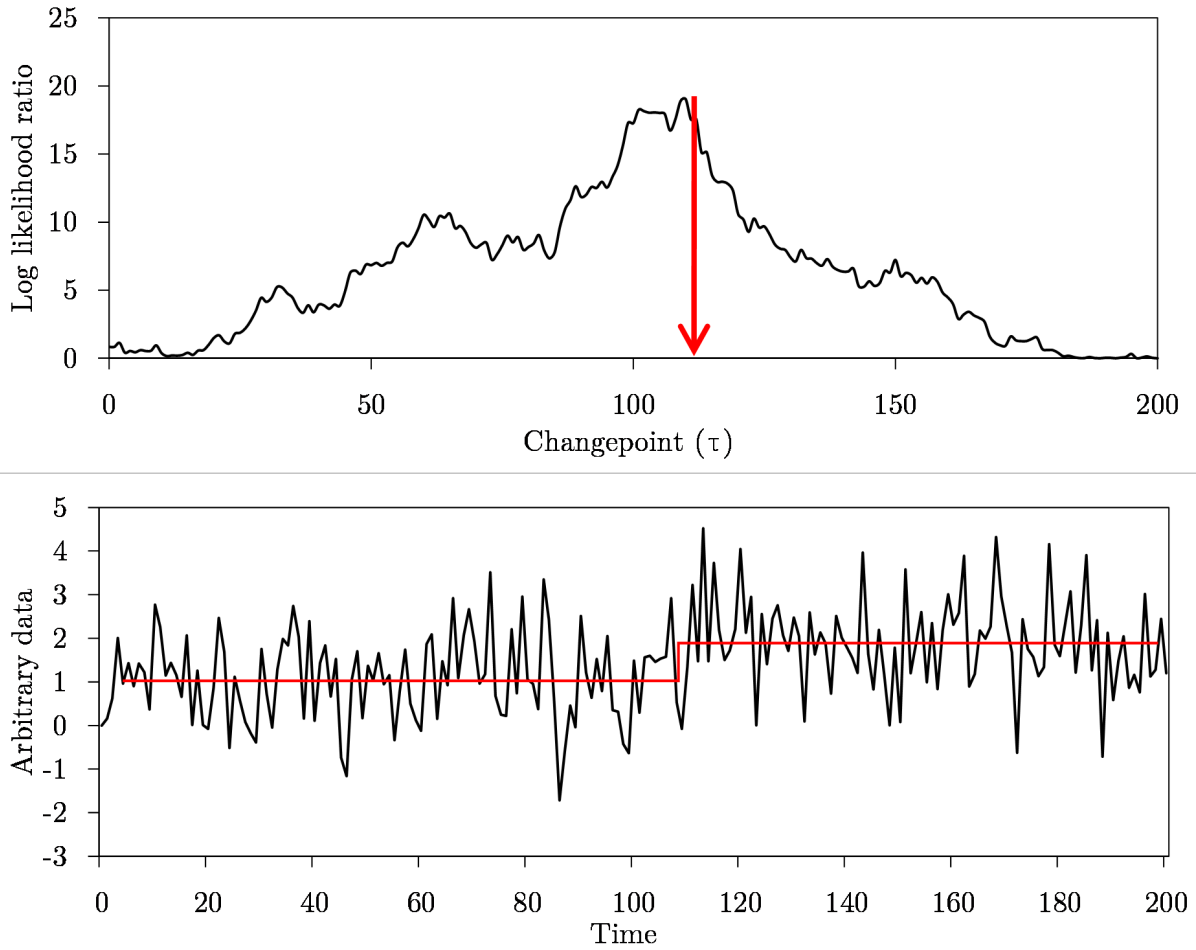


Figure 1.17: Top: log likelihood ratio Equation 1.21 with respect to τ . Bottom: arbitrary time series data with standard deviation $\sigma \approx 1$. The red line shows μ_1 and μ_2 as was discovered in Figure 1.17.

This method is not efficient when multiple change points are being searched for within a large data set [41]. It is suggested that a suitable value for c be chosen; however, methods for this are currently being researched and will not be focused upon in this study [41]. To locate multiple change points, it is possible to sum the maximum likelihoods (Equation 1.20) for various subsets in the data. The accompanying problem with this method, however, is deciding how to “search” for the optimal combination of subsets that will produce the maximum likelihood ratios. A popular method is to minimise a cost function [41]. Let m be the number of segments, if the data were partitioned into two segments then $m = 2$. A cost function (C) can then be calculated for each segment and summed (note that the cost function is different from the test metric c). This can be repeated with different segments to find the minimum sum. Equation 1.22 below is the mathematical representation of this process.

$$\sum_{i=1}^{m+1} [C(x_{(\tau_{i-1}+1):\tau_i})] \quad (1.22)$$

where τ_i is the location of a change point and x is the data. The algorithms discussed hereafter

use the negative of Equation 1.20 as the cost function C [41]. Substituting $C = -ML$ into Equation 1.22 would produce:

$$\sum_{i=1}^{m+1} -[ML(x_{(\tau_{i-1}+1):\tau_i})] \quad (1.23)$$

One of the most-widely-applied methods to minimise Equation 1.23 is that of Binary Segmentation (BinSeg) [41], as developed by Edwards and Cavalli-Sforza [52]. This method first applies an LLR test on the whole data set so as to find a single change point, shown in the second plot of Figure 1.18. To find the single change point, Equation 1.21 is used as a cost function and the test metric is then set to maximise the cost function [53]. In other words, $c = \max(LLR)$ in the BinSeg algorithm. If a change point is detected, the data are then divided at this point (third plot) and the same calculation is repeated on the two new segments (fourth plot). The process is repeated on each new segment until no change points are detected [41]. This method is not exact given that each new change point is affected by the previous change point [41].

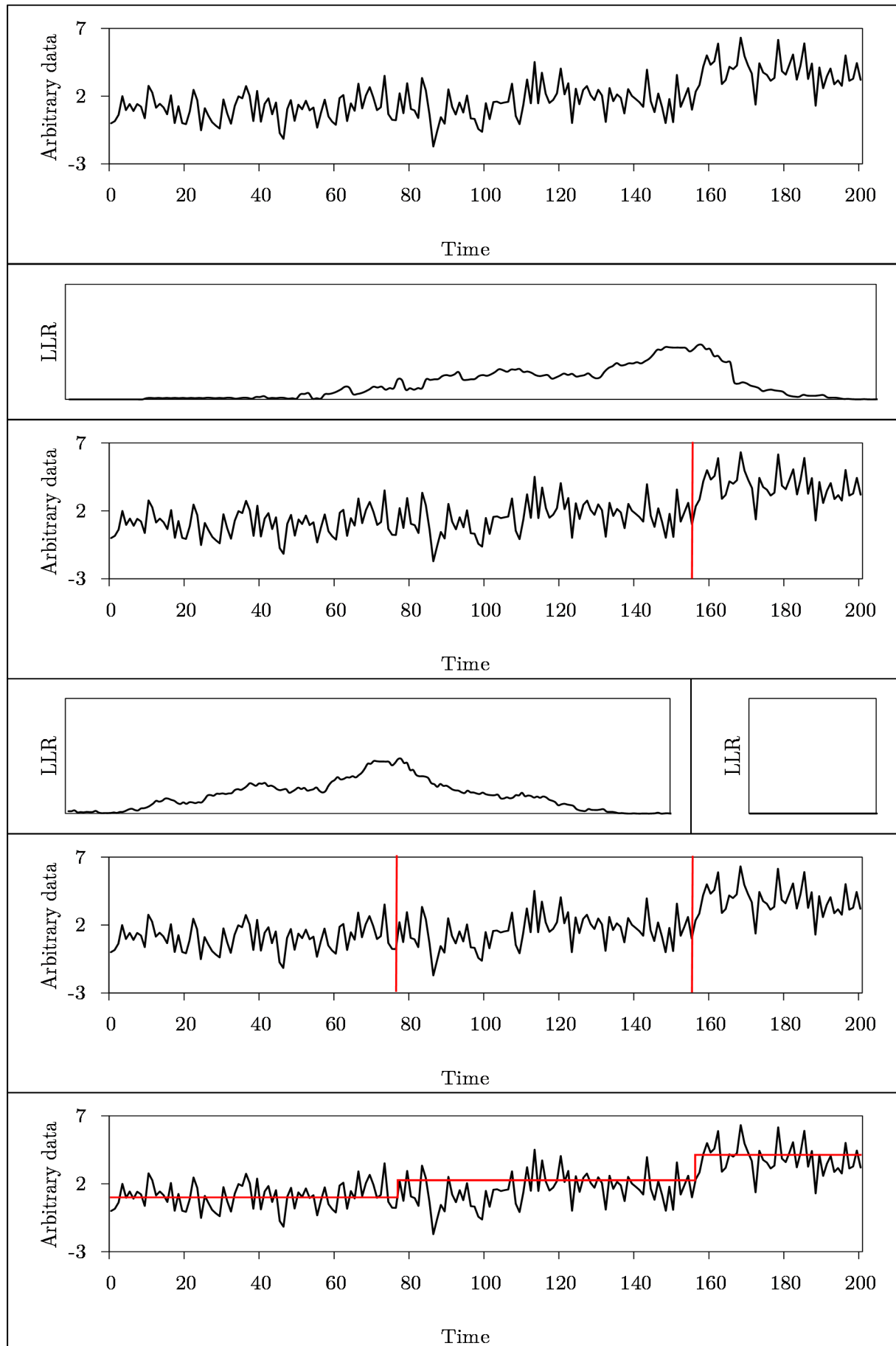


Figure 1.18: Visual representation of the BinSeg algorithm.

For more accurate multiple change point detection, Killick et al. [41] suggest the use of an optimal partitioning algorithm — the following explanation is adapted from the work by Killick et al. [41] and Killick, Fearnhead, and Eckley [53]. So as to explain this algorithm via an example, Figure 1.19 presents arbitrary data with two steps where the mean increases. This algorithm aims to minimise Equation 1.24).

$$\sum_{i=1}^{m+1} -[ML(x_{(\tau_{i-1}+1):\tau_i})] \quad (1.24)$$

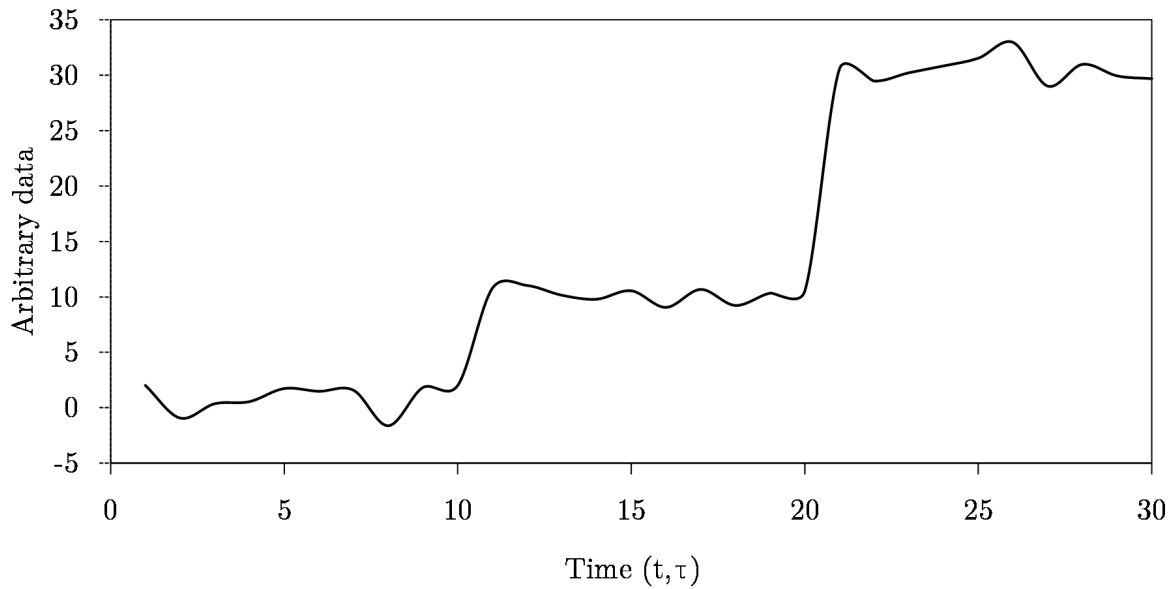


Figure 1.19: Arbitrary time series data showing two increases in the mean.

In each step, a single change point is located by computing the minimum of Equation 1.24 for a small subset of the data [53]. For example, the mini-segment may range from x_0 to x_{15} in Figure 1.19, Keeping in mind that $\tau_0 = 0$ and $\tau_2 = 15$, which are the first and last data points in the mini-segment. Then a single change point is searched for by minimising:

$$singlechange\ point = \min[-ML(x_{(\tau_0+1):\tau_1}) - ML(x_{(\tau_1+1):\tau_2})] \quad (1.25)$$

Equation 1.25 is simply the expansion of Equation 1.24 for two segments, or a single change point. These two segments are now inside of the mini-segments. If one were to plot term 1 and term 2 in Equation 1.25, their curves would be similar to those of set 1 and set 2 in Figure 1.16. However, the curves would be concave because the negative log likelihood is plotted [41]. The lowest points for term 1 and 2 are provided by the minimum of Equation 1.25 and indicate a possible change point — let this step be MIN_A . Visually, this minimisation is presented in Figure 1.20, with a clear indication that a possible change point exists at time = 10.

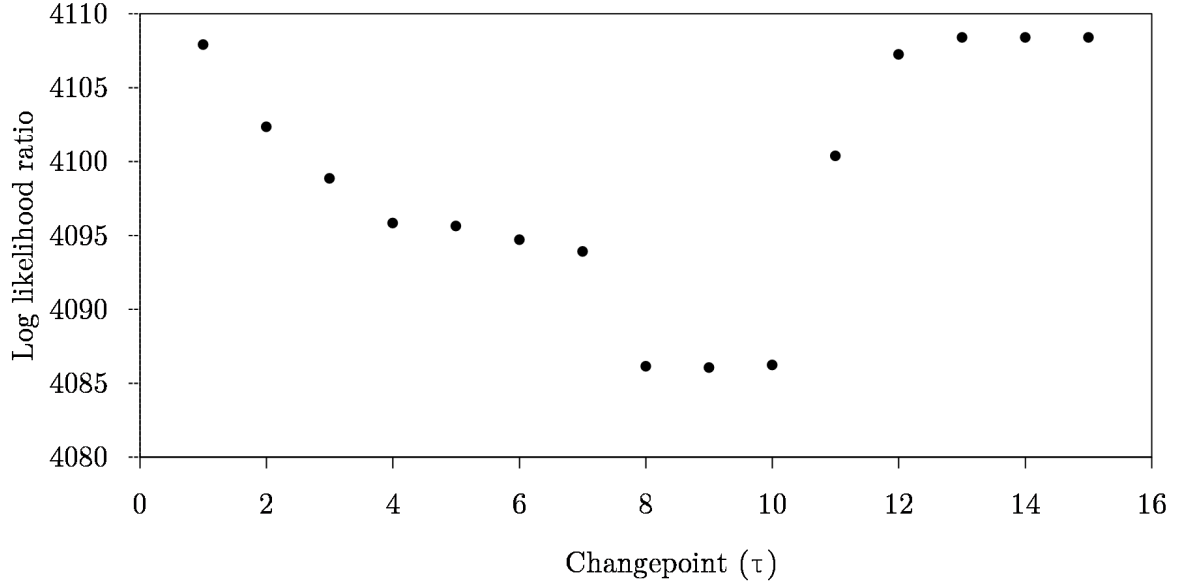


Figure 1.20: Summed negative log likelihood of two segments to identify a single change point in the arbitrary data of Figure 1.19 between time points 0.0 and 15.0.

This change point is denoted τ^* and is known as “the last change point” [53]. The minimum of Figure 1.20 (MIN_A) is stored. The starting point (x_{11}) for the next mini-segment is determined by τ^* . Let the length of the segment be 15.0 once again. The minimisation of Equation 1.25 is repeated for this new mini-segment and added to MIN_A . This is shown in Equation 1.26 and plotted in Figure 1.21.

$$MIN_A - [ML(x_{(\tau^*+1):\tau_1}) - ML(x_{(\tau_1+1):\tau_2})] \quad (1.26)$$

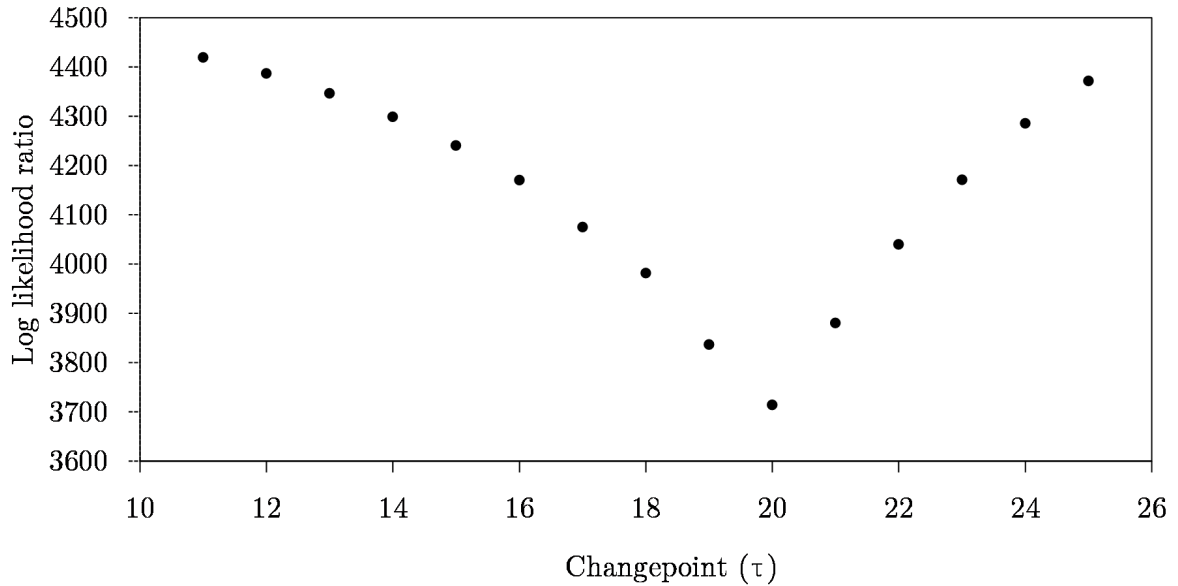


Figure 1.21: Summed negative log likelihood of the previously identified change point MIN_A and the two segments within mini-segment $x_{11}-x_{25}$ to identify a single change point in Figure 1.19.

The new minimum point in Figure 1.21 indicates another change point. The formal definition

of optimal partitioning is presented in Equation 1.27. This is written in terms of the cost function C , which will be the negative log likelihood (-ML) summation for two segments within a mini-segment (such as that depicted in Figure 1.20).

$$F(n) = \min_{\tau^*} \left\{ \min_{\tau|\tau^*} \sum_{i=1}^m C[y_{(\tau_{i-1}+1):\tau_i}] + C[y_{(\tau^*+1):n}] \right\} \quad (1.27)$$

The PELT algorithm uses optimal segmentation. Pruning involves eliminating the time points that cannot possibly be change points [53]. Optimal partitioning is an exact method, meaning that all of the possible change points are correctly identified. This substantially reduces the search space and reducing the optimal partitioning algorithm's running time, hence the term "linear time" — as discussed in the next section. Optimal partitioning may be *too* sensitive and may identify each data point as a change point. To prevent this over fitting, a penalty term β is added to Equation 1.27, as shown in Equation 1.28 [41].

$$F(n) = \min_{\tau^*} \left\{ \min_{\tau|\tau^*} \left(\sum_{i=1}^m C[y_{(\tau_{i-1}+1):\tau_i}] + \beta \right) + C[y_{(\tau^*+1):n}] \right\} \quad (1.28)$$

An iterative approach can be taken to identify β . By varying *beta* from some integer a to b in Equation 1.28, a range of change points can be calculated. The number of change points are then plotted against β [54]. This was applied to arbitrary data shown in Figure 1.22. The penalty was varied from 1.0 to 20.0 and resulted in the elbow plot in Figure 1.23.

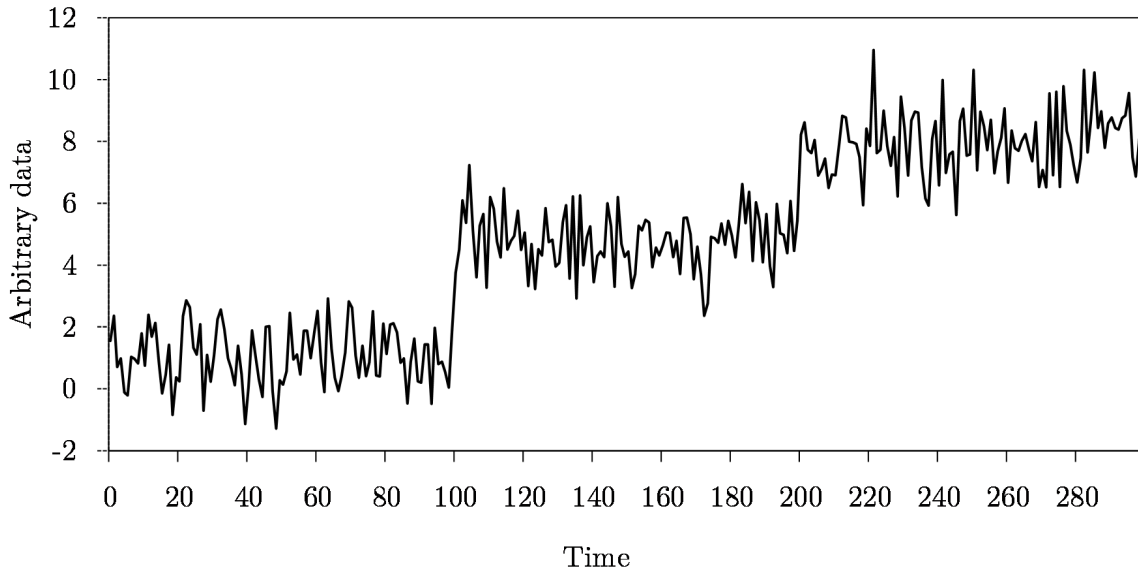


Figure 1.22: Arbitrary time series data showing two change points.

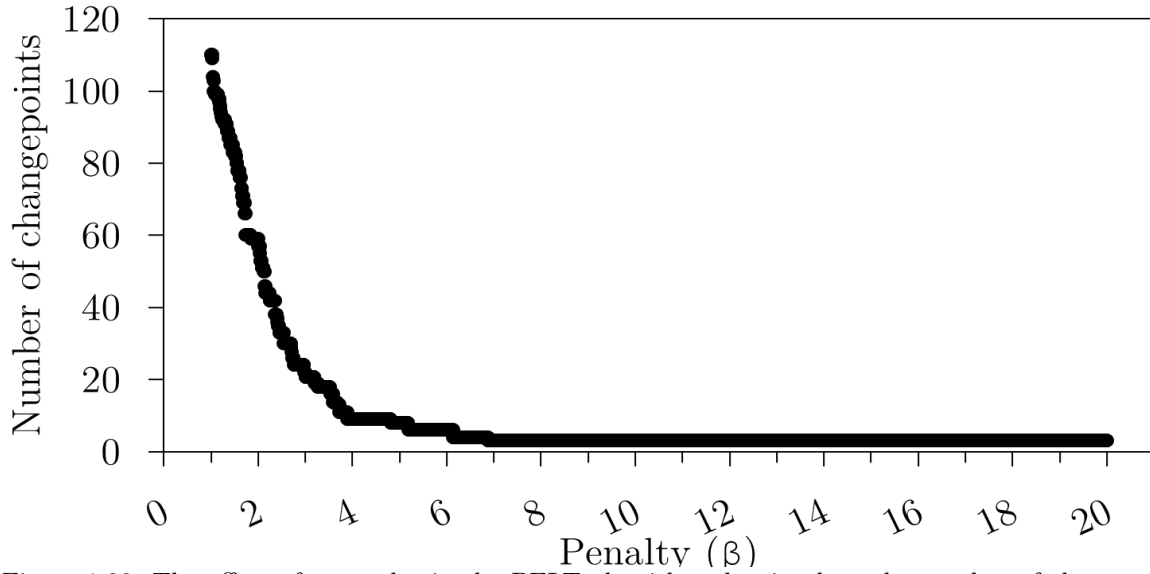


Figure 1.23: The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.

As a rule, the apex of the angle in the elbow plot can be used to estimate a suitable value for β [41]. It seems as though the elbow curve converges to zero as the penalty reaches 8.0. A penalty should be chosen that enables the identification of a realistic number of change points. The test was repeated with a penalty range spanning from 1.0 to 10.0. From this plot, Figure 1.24 shows that a penalty between 4.0 and 6.0 is reasonable for this data set and that a realistic number of change points is less than 10.0.

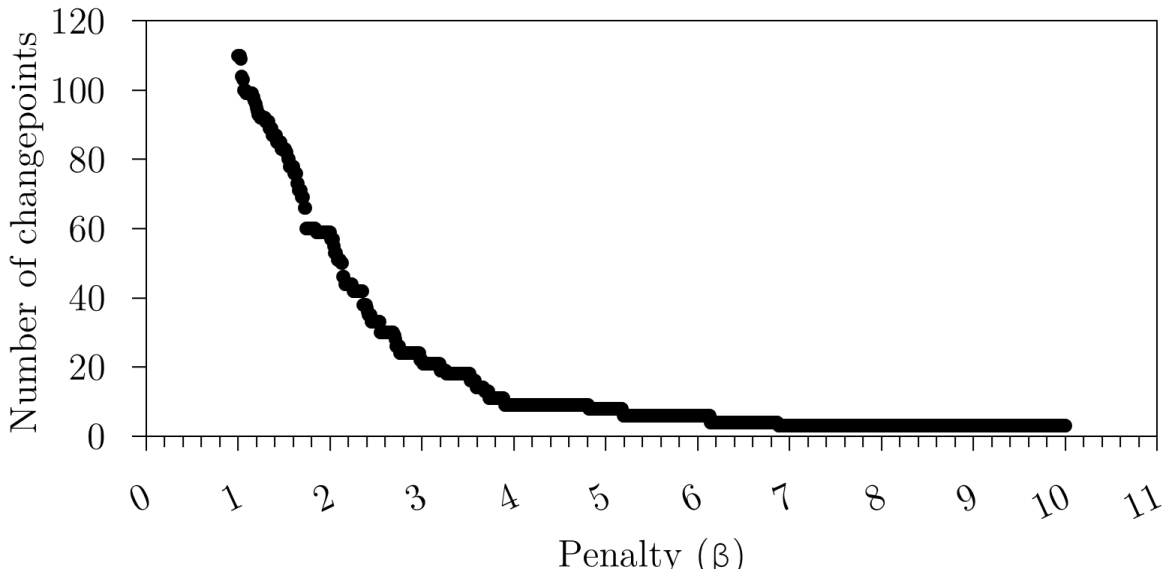


Figure 1.24: The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.

In change point detection analysis, the selection of an appropriate penalty remains an unresolved and open research question [53]. Although the penalty does improve the identification of true change points, it does not guarantee that all of these changes constitute an NRE. In the following section, verification metrics that scrutinise the success of an NRE detection method will be discussed.

1.4 Verification of rapid non-routine event detection

In Section 1.3.1, the importance of data preparation for NRE detection was explained. This constitutes the first methodology verification step: to ensure the raw data is correctly organised so as to aid in successful NRE detection. Data preparation must account for changes in static factors [27] and must be reorganised to avoid expected change points that can be attributed to changes in season or peak demand patterns.

In studies that test algorithms as possible solutions to some research problems, verification is defined as “the task that determines whether an algorithm will be suitable and applicable to the problem at hand” [55]. One suitability test for CPD algorithms is that of runtime, whereby “the algorithm must remain computationally efficient when applied to a large quantity of high resolution data.”

An algorithm’s runtime can be determined by measuring the start and ending timestamps and finding the difference between the two points. To compare two algorithms, it is important to ensure consistent and equal testing conditions. Algorithms often compute in milliseconds. On such a small scale, measurement noise is often present — as such, the algorithm should be executed several times, with the average runtime calculated as a result. As presented in a paper by Shastri, Manostru, and Sharma [56], the average runtime was computed by executing each algorithm 10 times. One must be aware that other processes on the computer, such as updates and malware protection, are most likely running concurrent to the execution of the algorithm. These processes share some of the memory that is needed when running an algorithm. Therefore, it is suggested that most unnecessary processes are paused while testing is being conducted. This ensures consistency when comparing the performance of algorithms with one another [56][57].

Time complexity is expressed in “big-O” notation [58]. Given the number of data points n , an algorithm may have a complexity of $O(g(n))$. This reads as “the complexity is big O of some function g of n ” [58] and relates runtime to the number of operations that an algorithm performs. It is important to state that big-O establishes the worst case runtime, therefore an algorithm could be faster than its big-O function [58]. To explain further, consider two sorting algorithms, bubble sort and heap sort. Given an unsorted list of integers, bubble sort iteratively compares each element to the next one and checks which is bigger. If the second element is bigger, then it switches their positions in the list and starts over, reading the list from the beginning. This is illustrated in Figure 1.25.

The heap sort method simply identifies blocks of already sorted numbers, as illustrated in Figure 1.26. It then iteratively slots the rest of the unsorted numbers into their correct positions.

The bubble sort complexity is $O(n^2)$, with heap sort’s complexity being $O(n\log(n))$. Heap sort has lower time complexity than that of bubble sort. This is shown in Figure 1.27, where it is clear that the bubble sort runtime (solid line) rapidly increases with input (n). When the list

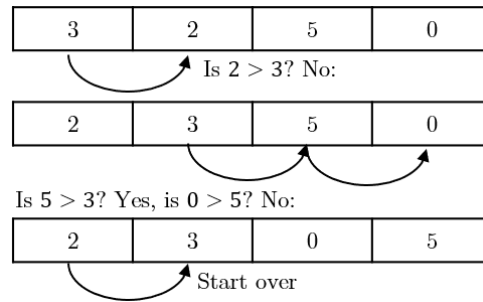


Figure 1.25: Illustration of the bubble sort algorithm.

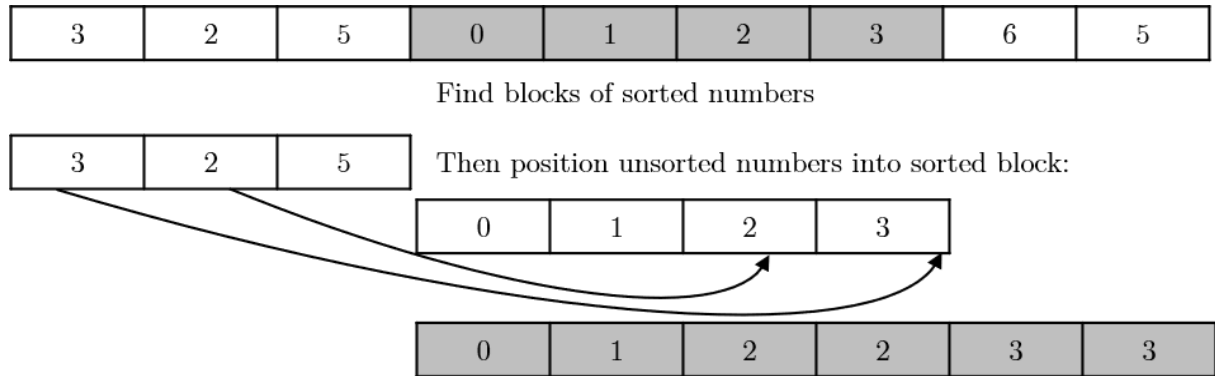


Figure 1.26: Illustration of the heap sort algorithm.

has twenty elements, for example, the bubble sort algorithm performs 400 operations to sort the list. The time in seconds varies based on the computer it is run on and should be timed. This becomes quite significant with larger data sets and would render the bubble sort algorithm too sluggish for use.

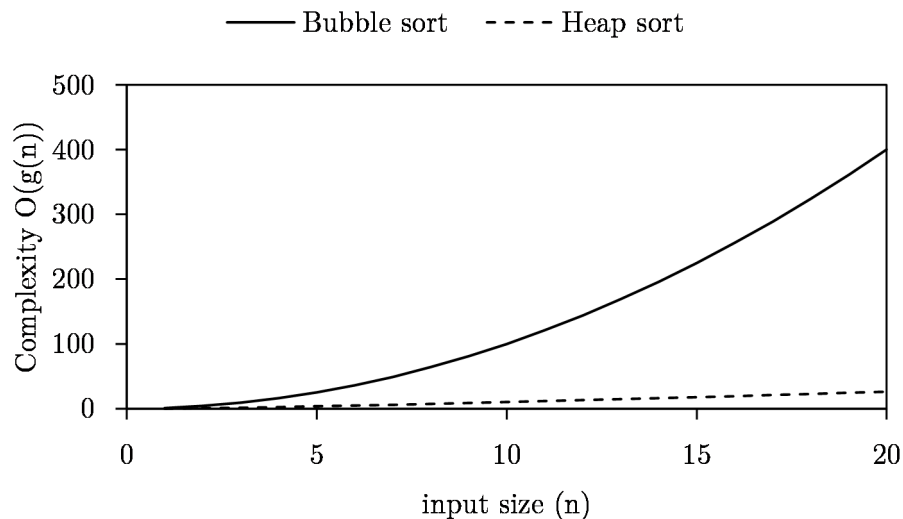


Figure 1.27: Comparison of the bubble and heap sort algorithm complexities.

The computational complexity of the CUSUM method is negligible given that it performs a single calculation per data point [59]. Therefore, it is not necessary to quantify. Complexity

is important when many operations are performed on multiple data points. Briefly, the BinSeg method searches for a single change point in the entire data set. If one is found, the data set is split at that point. A single change point is subsequently searched for in each subset. Because operations are performed on subsets of the data, it improves computational efficiency [41]. BinSeg has complexity $O(n \log n)$ [53], which is inexpensive, but which is not exact. To improve accuracy while keeping complexity low, Killick et al. [53] developed PELT, which has been proven to have a linear computational cost $O(n)$ [53]. The “pruning” removes the locations where it is impossible that a change point would occur, reducing the search space and the number iterations needed to find the optimal combination of change-point locations.

According to Pedersen et al. [60], a validation procedure stipulates the performance indicators that a mathematical model must have in order to be successful. In terms of CPD, Touzani et al. [6] state that a CPD algorithm must be capable of both correctly identifying a change point and detecting multiple changes in time series data.

An NRE detection algorithm can be validated by applying it to data with a previously confirmed NRE. If the algorithm identifies the confirmed NRE as a change point, then it passes the validation check. As pointed out by Earni and Therkelsen [7], it is worth noting that momentary spikes or dips in the energy do not constitute a valid change point. Therefore, a reasonable minimum segment length should be specified so as to prevent momentary spikes or dips being identified as change points. General guidelines and factors that should be accounted for to validate an NRE algorithm are summarised in Table 1.4 [61]:

Table 1.4: Table of important factors to consider when confirming that a valid change point was identified.

Metric	Example	Description
Magnitude	Set limits	The most important factor to classify a valid change point. This should be set according to acceptable limits of the system. As discussed on page 8 in Section 1.3.2, possible thresholds are 3σ and 10 % of min/max [27]. For mining-energy data, the threshold is 2σ [24].
Duration	Short Long	Anomalous data points may not be noteworthy if the duration is seconds/minutes. Duration in combination with magnitude should more accurately identify a significant change point [7]. In mining-energy data, events are considered significant if the duration is several hours in length [24].

Frequency	Sample rate	This refers to the resolution. Whether data points are logged hourly, daily or monthly will affect what time frame can be monitored [7]. Page 4 in Section 1.3.1 discussed that high resolution can be a noisy data set that obscures trends.
Period	Baseline Assessment Transition	The transition (in between) period just after the EEM is implemented is expected to show changes [27].
Type	Demand Schedule Variance	Mines implement load shifting according to Eskom's tariffs. In more expensive periods, equipment is cut back upon or turned off. The peak period times (schedules) are different in the summer and winter. When equipment is not cut on back during this scheduled time, it can be a possible NRE [24].
Independent variables	Temperature Weekday Production Season	A change is not considered an NRE if it is brought about by cyclical changes (such as that of a seasonal change). However, if the system is directly affected by production, and a sudden increase of production is initiated by the mine, then the resulting energy increase is an NRE [24].

In summary, the change point detection methodology and selection of algorithms can be verified by answering the following:

1. Does the data preparation organise the data such that cyclical changes are not present?
2. Are the chosen algorithms computationally efficient?
3. Were consistent testing conditions in the computer enforced?

To validate that the CPD algorithms successfully identify change points in mining-energy data:

1. The algorithms should detect a previously known change point.
2. The algorithms should be capable of identifying more than one change point.

Lastly, the objective to find algorithms for rapid NRE detection in mining-energy data should be addressed. To validate whether or not an algorithmic approach will satisfy this objective, the approach will be compared to a prescribed IPMVP method for the purposes of this study. The verification and validation of a methodology and algorithms for rapid NRE detection concludes the literature review. Important concepts, the need for the study, and objectives are all discussed in the following section.

1.5 Problem statement, need and objectives

Section 1.2 explained that NREs interfere with the success of an EEM [5]. It is important to monitor a mine's data so as to rapidly detect anomalous behaviour. Currently, this monitoring is mostly done through communication with personnel and some manual data analysis [7]. The following problems often arise:

1. Manual monitoring is subject to both missing anomalies and the subjective nature of the individual's opinion when classifying an NRE [6].
2. Due to the complexity present in mining operations, there are multiple systems and measurements in place. This makes it infeasible to monitor every data stream.

The increased availability of data, as brought about by AMI, has created the need for advanced analysis [4]. In the case of NRE detection in mining energy data, the need is to apply automated anomaly detection. M&V 2.0 is an umbrella term for the latest advances which enable analysts to harness the potential of AMI [5].

To initialise the scope of this investigation, critical measurement points of mining energy data were derived from literature. In Blaettler, significant systems for the case study mine that account for 80% of the total energy consumption were identified [62]. This narrowed the monitoring focus to systems that have the largest impact and may become costly if not properly managed. Therefore, it would be beneficial to apply automated anomaly detection to such systems.

One such focus area is data-driven change-point detection. This study will thus investigate the potential of applying change point algorithms for rapid NRE detection in the mining industry. An IPMVP-prescribed technique and two CPD algorithms will be investigated. The IPMVP method suited to time series change detection is that of cumulative sums (CUSUM) [27]. The two CPD algorithms that will be tested are Binary Segmentation and Pruned Exact Linear Time. The aim of this study is to improve the current delayed NRE detection, and the objectives are summarised below:

1. Apply the CUSUM, BinSeg, and PELT methods to the mining-energy data identified in Blaettler [62].
2. Evaluate the ability of each method to rapidly identify NREs.
3. Evaluate each method's ability to correctly identify NREs with applications on historically confirmed NREs.
4. It is also important to criticise each method's ease of implementation and computational complexity.

1.6 Outline of document

- Chapter 1: *Background and literature study*: This chapter provides a description of important elements and procedures in mines pertaining to non-routine events. It also outlines that M&V 2.0 emerged with data-driven analyses to utilise the enormous availability of energy data. Data preparation techniques and the importance of data quality is also discussed. From this discussion, change-point detection is described, followed by the description of the applicable algorithms that will be investigated in this study. To compare the relevant performances, this chapter discusses the metrics that will be used to evaluate each algorithm. In conclusion, the main objective is to implement data-driven, non-routine event detection that will improve the current slow, inconsistent, and manual monitoring process.
- Chapter 2 *Methodology*: The literature review is used to develop steps that ensure the fulfilment of each objective outlined in Chapter 1. These steps are demonstrated with a mine-compressor's power data as an example and as a proof of concept. Initially, the raw data are categorised according to season, weekday, and Eskom time of use periods. Holidays must be reassigned as either a Saturday or Sunday, as per Eskom's calendar. For IPMVP NRE detection, a baseline period and manual data inspection with outlier removal is discussed. To implement BinSeg and PELT, the method to select appropriate penalties is explained. These methods also require parameter-tuning for each data set, such as the minimum duration of a valid change point. This chapter concludes with concise scoring metrics that aim to summarise the characteristics of each algorithm.
- Chapter 3 *Results and discussion*: The case studies and raw data are shown and discussed. Data preparation is applied to each case study and the final testing data are displayed visually. The change-point detection methods are then performed on the prepared test data.
- Chapter 4 *Conclusion*: The purpose of the study and problem statement is reiterated. The following discussion highlights how each of the research objectives were achieved. Recommendations for further studies are also presented. The recommendations are followed by a critical analysis of the study as a whole. Lastly, the impact of the study and relevance to the mining industry is discussed.

Chapter 2 Methodology for rapid non-routine event detection

2.1 Preamble

The problem statement presented in Chapter 1 emphasises the need to closely monitor energy data in order to rapidly identify anomalous events. In Chapter 2, the procedure is discussed from the point of raw data collection and preparation to change-point detection. The important steps identified through the literature review are:

1. Data preparation.
2. Change-point algorithm parameter-tuning based on the properties of each data set.
3. Consistent operating conditions in the computer when comparing algorithms.
4. Application of change point detection algorithms.
5. Verification of each algorithm according to accuracy, ease of implementation, and computational complexity.

Each of these steps will be described in this chapter with respect to mining-energy data. To begin, the following section aims to identify what categorical variables and measurement resolution will produce the most sensible data for successful change point detection.

2.2 Preparation of mining-energy data

The preparation of mining-energy data for anomaly detection can be summarised via three steps: 1) label weekday holidays as either “Saturday” or “Sunday” according to Eskom’s calendar; 2) assign “Weekday”, “Saturday”, or “Sunday” to all other dates; and 3) separate the data set into summer and winter periods and label each hour as “peak”, “standard”, or “off-peak” according to Eskom’s schedule.

It was speculated in Section 1.3.1 (Page 4) that Eskom’s tariffs may govern the mine’s energy consumption [23]. This study will apply the Eskom time-of-use tariffs as categories on a case study mine. The raw data will be reorganised according to weekday/holiday, off-peak, standard, and peak time-of-use periods. This may provide more consistency in the data for a weekday, for example, and remove “dips” from holidays that fall on weekdays. The holiday calendar for 2020-2021 is shown in Table 2.1.

Table 2.1: Eskom assignment of vacation and holidays in 2020-2021 [23].

Date	Day	Actual day of week	Time of use treated as
10 April 2020	Good Friday	Friday	Sunday
13 April 2020	Family Day	Monday	Sunday
27 April 2020	Freedom Day	Monday	Saturday
01 May 2020	Workers Day	Friday	Saturday
16 June 2020	Youth Day	Tuesday	Saturday
09 August 2020	National Women's Day	Sunday	Sunday
10 August 2020	Public Holiday	Monday	Saturday
24 September 2020	Heritage Day	Thursday	Saturday
16 December 2020	Day of Reconciliation	Wednesday	Saturday
25 December 2020	Christmas Day	Friday	Sunday
26 December 2020	Day of Goodwill	Saturday	Sunday
01 January 2021	New Year's Day	Friday	Sunday
21 March 2021	Human Rights Day	Sunday	Sunday
22 March 2021	Public Holiday Day	Monday	Saturday
02 April 2021	Good Friday	Friday	Sunday
05 April 2021	Family Day	Monday	Sunday
27 April 2021	Freedom Day	Tuesday	Saturday
01 May 2021	Worker's Day	Saturday	Saturday
16 June 2021	Youth Day	Wednesday	Saturday

To demonstrate the impact of re-assigning holidays according to the Eskom calendar, an example data set is shown in Figure 2.1. This is power data logged at half-hour-intervals for every day in August 2020. The compressor power output is reduced on weekends and corresponds to the dips seen in Figure 2.1. However, there is an extended period of low power from day 7 to 10. According to the Eskom calendar, National Women's day was observed on Monday 10 August. This Monday was treated as a Saturday, with less production taking place, explaining the extended low demand seen in Figure 2.1.

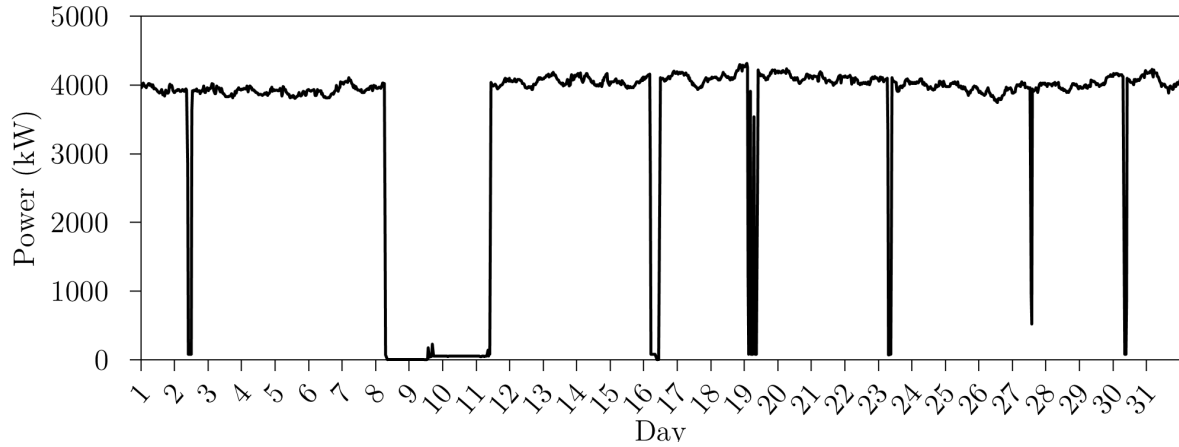


Figure 2.1: Power data from a mining compressor for August 2020.

The second step involved adding the day of week labels to the remaining dates in the data set. When the weekday holiday is relabeled as a Saturday, it is not included with the weekday data. This prevents the holidays from obscuring the profiles associated with weekdays. Figure 2.2 shows some weekdays in August with the holiday excluded. This demonstrates that the compressor can be expected to operate in the vicinity of 4000 kW on a weekday. By removing the public holiday, the long weekend from 8 to 10 August is excluded in Figure 2.2; this distinction then clearly highlights an abnormality on Friday, 7 August. It is possible that personnel received more time off and ended the workday earlier. Nevertheless, this division would reduce false change point identification that can be attributed to a holiday. The reader is reminded that outlier and data loss analysis will not be performed at this stage because it falls within the scope of change-point detection.

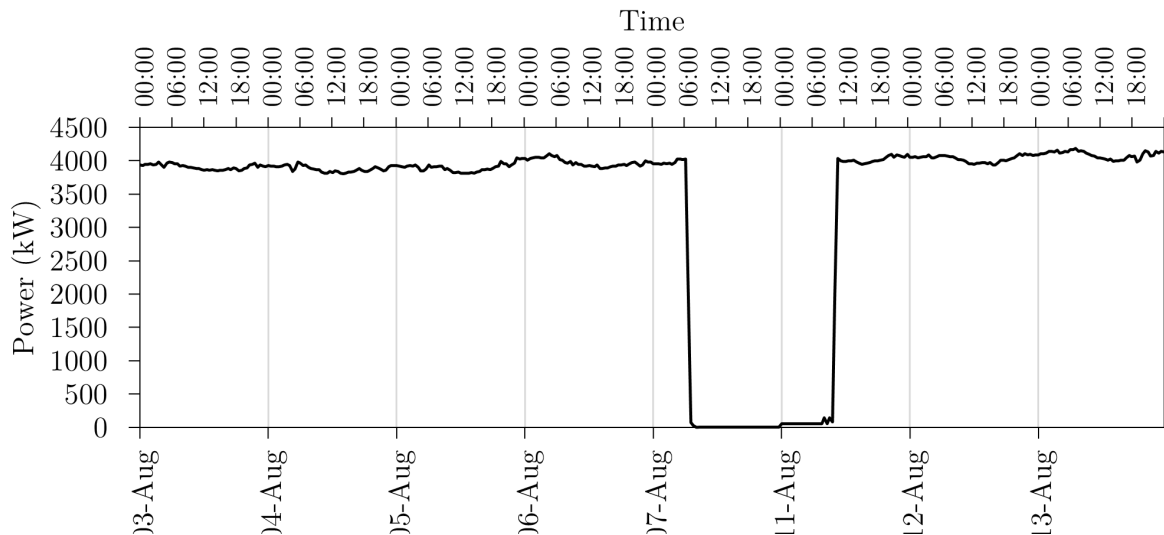


Figure 2.2: Power data from a mining compressor for a selection of weekdays in August 2020.

Figure 2.3 shows the average hourly compressor power for a typical weekday in August 2020. The grey bars indicate Eskom's peak tariff hours. This shows that the power was lowered during the morning peak period and indicates a possible trend.

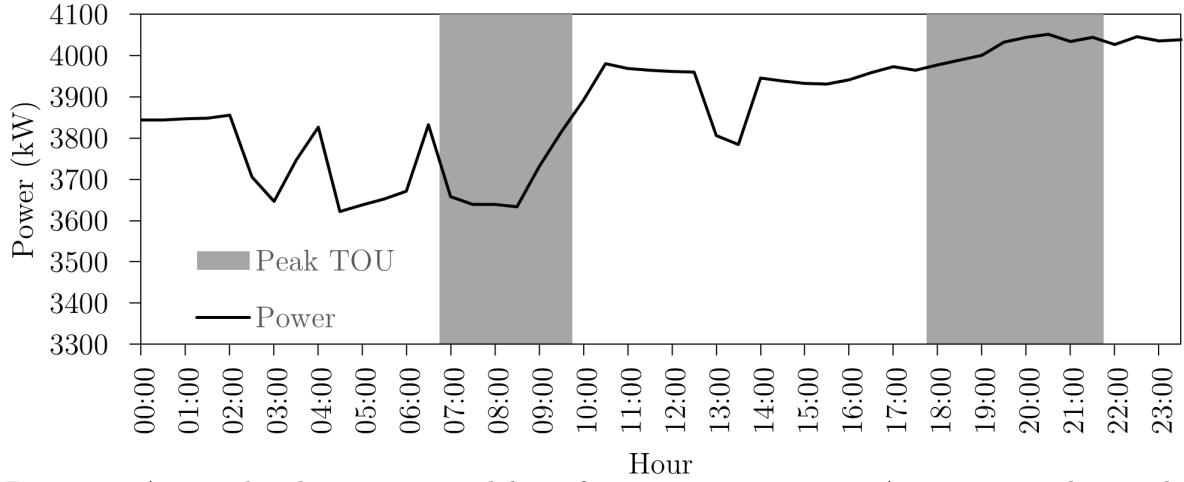


Figure 2.3: Average hourly power on weekdays of a mining compressor in August 2020. The grey bars indicate Eskom peak tariff hours.

The Eskom tariff book also stipulates the summer and winter time-of-use periods [23]. This is summarised in Table 2.2; the period starts on the hour; thus, in summer from 6am – 7am “standard” tariffs apply. The third data preparation step is to divide the data into summer and winter months. Thereafter, each hour will be labeled as “peak”, “standard”, or “off-peak” according to Table 2.2.

Table 2.2: Eskom’s assignment of time of use periods between peak, standard and off-peak times.

Time	Summer			Winter		
	Weekday	Saturday	Sunday	Weekday	Saturday	Sunday
00:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
01:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
02:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
03:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
04:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
05:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
06:00	Standard	Off-peak	Off-peak	Peak	Off-peak	Off-peak
07:00	Peak	Standard	Off-peak	Peak	Standard	Off-peak
08:00	Peak	Standard	Off-peak	Peak	Standard	Off-peak
09:00	Peak	Standard	Off-peak	Standard	Standard	Off-peak

10:00	Standard	Standard	Off-peak	Standard	Standard	Off-peak
11:00	Standard	Standard	Off-peak	Standard	Standard	Off-peak
12:00	Standard	Off-peak	Off-peak	Standard	Standard	Off-peak
13:00	Standard	Off-peak	Off-peak	Standard	Standard	Off-peak
14:00	Standard	Off-peak	Off-peak	Standard	Standard	Off-peak
15:00	Standard	Off-peak	Off-peak	Standard	Standard	Off-peak
16:00	Standard	Off-peak	Off-peak	Standard	Standard	Off-peak
17:00	Standard	Off-peak	Off-peak	Peak	Standard	Off-peak
18:00	Peak	Standard	Off-peak	Peak	Standard	Off-peak
19:00	Peak	Standard	Off-peak	Peak	Standard	Off-peak
20:00	Peak	Off-peak	Off-peak	Peak	Off-peak	Off-peak
21:00	Peak	Off-peak	Off-peak	Peak	Off-peak	Off-peak
22:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak
23:00	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak	Off-peak

At this point, the data preparation is complete for anomaly detection algorithms. In this study, an IPMVP method for anomaly detection will be applied as a comparison to demonstrate advantages of algorithmic methods. In Chapter 1, the cumulative sums (CUSUM) method was identified as suitable for high-volume mining-energy data, given that it does not require visual monitoring through plots — plots which can only provide a limited view (Page 8). This method identifies abnormal behaviour by comparing assessment-period data to a baseline and which requires an additional data preparation step.

As discussed on Page 6, a baseline must represent consistent operating conditions and should be chosen such that it does not contain anomalies. If that is unavoidable, the outliers should be removed from the baseline data. Section 1.3.1 mentioned that the measured mining energy is only compared to monthly Eskom bills. The data should also be inspected for data-loss or outliers and the corresponding dates and times must be discarded. It seems counter-intuitive to perform an outlier analysis since the focus of this study is rapid anomaly detection. However, standard M&V practice is compared to an algorithmic M&V 2.0 approach to address the objective to apply rapid anomaly detection. It demonstrates that the M&V approach relies on the time-consuming process of setting up baselines.

The standard deviation of the baseline period can be calculated with Equation 1.3. Equation 1.2 computes the distance between each data point and the mean. Any data points more

than two standard deviations from the mean may be outliers [24]. If any such points are found, it must be determined whether or not their deviation was caused by a non-routine event (NRE). For demonstration purposes, let August 2020 be the baseline for the mining compressor. Figure 2.4 shows a standard deviation plot of the Eskom peak time-of-use period of this baseline data. The distance from each point to the mean of this baseline period is plotted in terms of standard deviations $\frac{|(x_i - \mu)|}{\sigma}$.

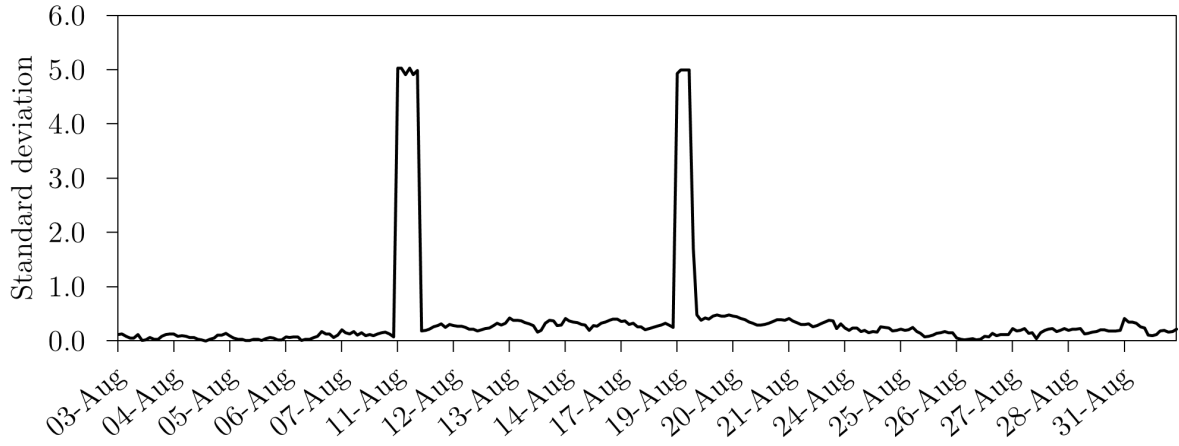


Figure 2.4: Standard deviation plot for a mining compressor's weekday peak time-of-use power output in August 2020.

Figure 2.4 highlights outliers in the peak time period on 11 and 19 August. These data points should be discarded to complete the preparation of a baseline for the IPMVP anomaly detection approach. Figure 2.5 shows the peak hours for every day in August 2020 of the mining compressor with the anomalous data removed to form the baseline for this example. This baseline outlier analysis should also be performed on the off-peak and the standard time-of-use data.

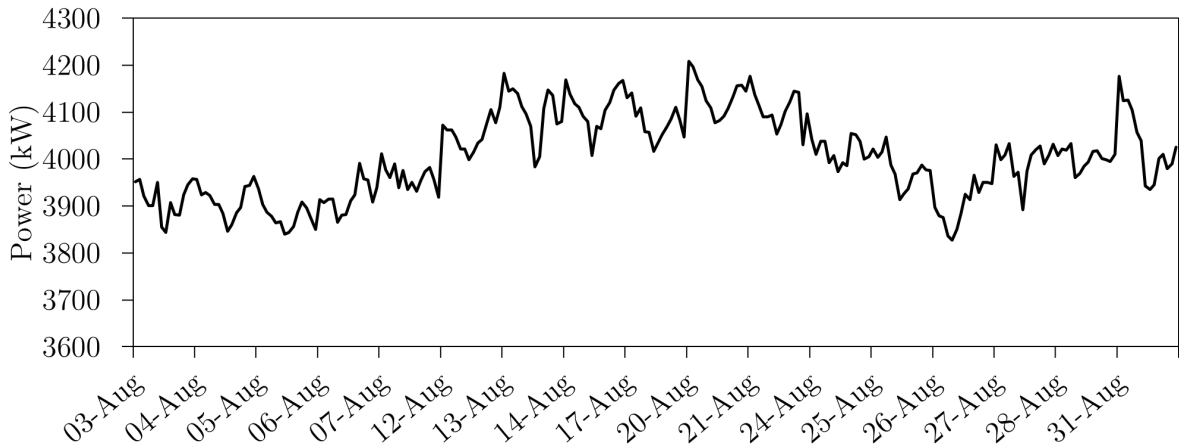


Figure 2.5: Baseline data for a mining compressor's weekday power output during Eskom's peak periods in August 2020.

According to the existing body of literature [27][41], the data should vary with approximately one standard deviation to apply anomaly-detection algorithms. Continuing with the mining-

compressor example, the data in Figure 2.5 should be scaled with respect to the median absolute deviation (Equation 1.13). The result of the scaling is shown in Figure 2.6. This concludes the baseline data preparation for the IPMVP anomaly detection technique. For this data, the average is 0.04 and the standard deviation is 1.25. These parameters can be used in the CUSUM method in order to detect non-routine events.

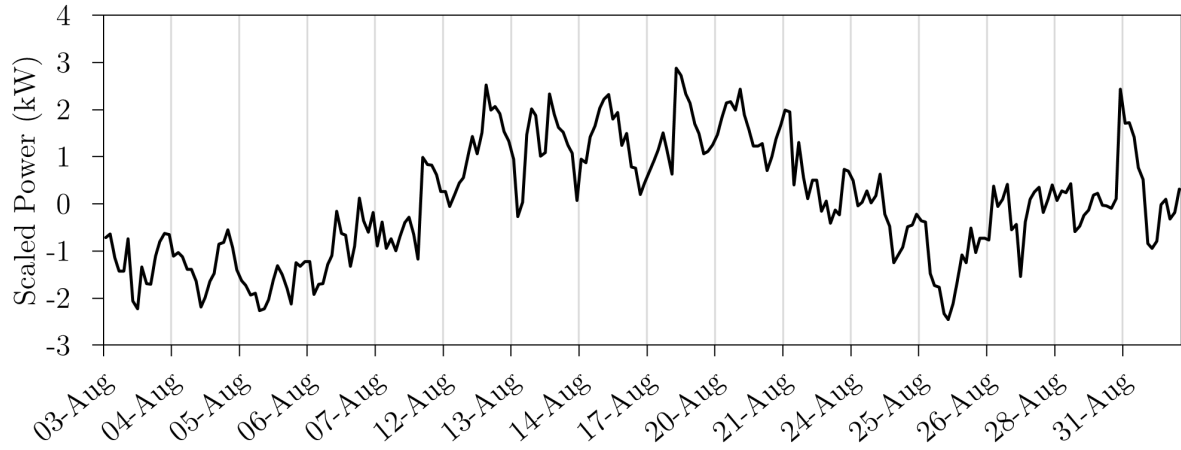


Figure 2.6: Scaled baseline data for a mining compressor’s weekday power output during Eskom peak periods in August 2020.

In summary, the raw data was categorised into season, weekday, and time-of-use periods. Thereafter, the raw data were scaled such that the mean is roughly zero, with approximately one standard deviation. The IPMVP method requires an extra step which consists of choosing a baseline with little to no outliers present. If outliers are unavoidable, these must be removed from the baseline data. The following sections explain the methodology in implementing the NRE detection methods.

2.3 IPMVP anomaly detection

Section 1.3.2 discussed the CUSUM method that is prescribed by the IPMVP. This method aims to provide an early warning if the system is approaching some predefined boundary. From the literature review, a boundary of two standard deviations ($\pm 2\sigma$) of the baseline period is applicable to mining data. Division into time-of-use periods should then create data sets that are normally-distributed. The baseline period must be chosen such that it describes the expected normal operation of the system, without any EEMs. Then the mean and standard deviation should be determined and substituted as constants into Equation 2.1 (repetition of Equation 1.2).

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (2.1)$$

Then Z_i should be calculated for every data point for the CUSUM method, as shown below.

$$S_{H(i+1)} = \max(0, S_{Hi} + Z_i - \omega) \text{ where } \omega = 2\sigma \quad (2.2)$$

$$S_{L(i+1)} = \min(0, S_{Li} + Z_i - \omega) \text{ where } \omega = -2\sigma \quad (2.3)$$

To indicate that the system is out of bounds, the limits are reached when $S_H > 2\sigma$ and $S_L < -2\sigma$. If the limit statements are true, the index in the date and time list should be recorded. This can be done in `Python` with the following pseudo code:

Algorithm 1 CUSUM

```

1:  $b$  = list of  $n$  energy data points from the baseline period
2: Calculate constant  $\mu$  = average of  $b$ 
3: Calculate constant  $\sigma^2 = \frac{\sum_{i=1}^n (b_i - \mu)^2}{n-1}$ 
4: Let  $\omega = 2 * \sigma$ 
5:  $x$  = list of energy data points from the assessment period
6:  $DateTime$  = list of date and time from the assessment data
7: Calculate  $Z_i = \frac{x_i - \mu}{\sigma}$ 
8: Calculate  $S_{H_{i+1}} = \max(0, S_{Hi} + Z_i - \omega)$ 
9: Calculate  $S_{L_{i+1}} = \min(0, S_{Li} + Z_i - \omega)$ 
10: if  $S_{H_{i+1}} > \omega$  then
11:     Add the index  $i$  from  $DateTime$  to a list
12: end if
13: if  $S_{L_{i+1}} < -\omega$  then
14:     Add the index  $i$  from  $DateTime$  to a list
15: end if

```

The list of recorded dates and times from Algorithm 1 will then indicate where the system may have approached its thresholds. An example is shown in Figure 2.7. The CUSUM method was tested on the example mining-compressor data for the period 23 November 2020 to 18 January 2021. The solid line shows the power output of the example data during the Eskom-delineated peak periods on weekdays. The dotted line is the negative cumulative sum, calculated with Equation 2.3.

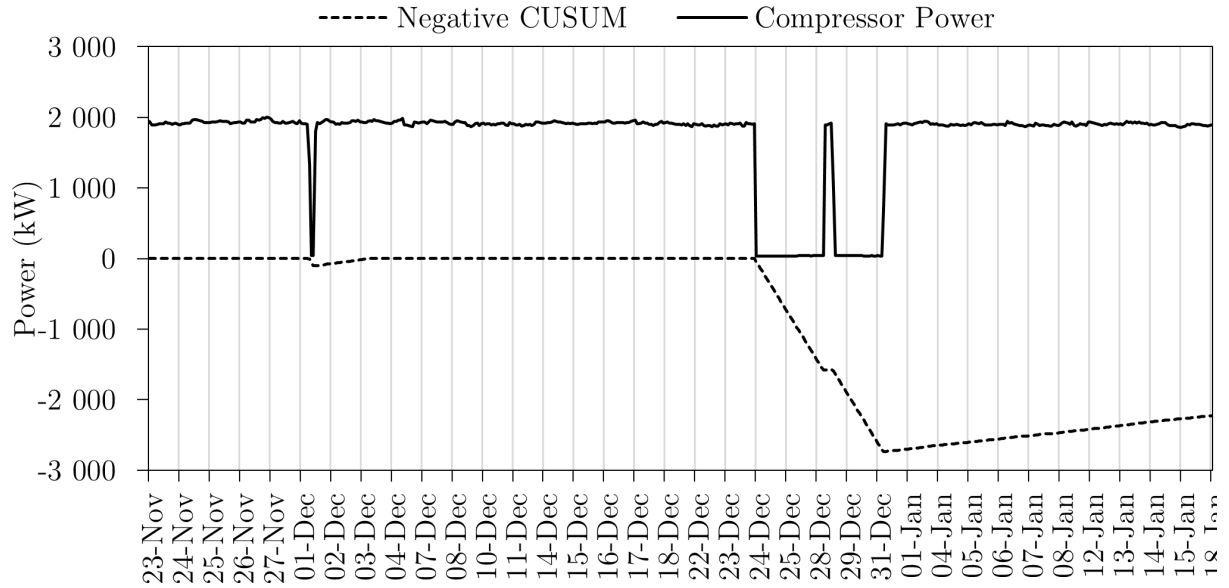


Figure 2.7: Negative CUSUM applied to a mining compressor’s weekday power output during Eskom peak periods in August 2020.

No positive accumulation was present and is confirmed by the absence of positive shifts in Figure 2.7. The CUSUM method correctly identifies a power decrease on 1 December. The dotted line shows a minor dip below zero, which recovers to null because the data became available again. A much larger decrease is present during the period of 25–31 December. This is expected during the Christmas break when no compressed air is required for mining. The magnitude of the CUSUM results give an indication of how long an NRE has occurred and will aid in classifying significant events. However, the CUSUM method shows “slow recovery” to null and this may obscure NREs that occurred shortly after. A more exact approach will also be investigated in course of this study, with the following section describing the method of implementing Binary Segmentation (BinSeg) and Pruned Exact Linear Time (PELT).

2.4 Change point detection algorithms

In Section 1.3.3, it was stated that the R programming language contained useful packages for anomaly detection. One such package is called `changepoint` [41]. This package contains R code that implements the BinSeg or PELT algorithms for identifying change points in time-series data. The package description and inputs that are required are the instruments used to answer the research question. The code example below shows the implementation to identify changes in the magnitude of time series data. Calling `cpt.mean()` essentially “fetches” the corresponding algorithms in the `changepoint` package.

```
ChangePoints = cpt.mean(data, minseqlen, penalty, pen.value, method, Q)
```

The input `data` is a list of time series data points, such as the compressor-power data in the

previous example. The input `minseglen` is the minimum segment length. This is the number of data points where no change is expected. For example, one would not choose a minimum segment length of one for the compressor example because a change point that lasts only 30 minutes is not deemed to be significant. As shown in Figure 2.4, and in concurrence with Table 1.4, it is considered significant when the compressor’s power output is out of the ordinary for a number of hours. Therefore, one would set `minseglen` = 8 (eight 30-min intervals) to ignore changes that last less than four hours. The minimum segment length influences the penalty and helps to identify the correct number of change points. If `minseglen` is too small, it may lead to many falsely-identified change points.

The `penalty` input offers a selection of penalty methods. Such methods remain an open and unresolved research question. However, `changepoint` contains functionality for some methods that have been developed thus far. In this study the penalty will be “Manual” for the PELT algorithm and selected with the elbow plot method discussed on Page 27. The apex of the elbow, at the sharpest point, shows at which penalty the algorithm identifies the most reasonable number of change points. For each data set, the number of change points will vary based on the penalty. This was applied to the example using compressor data in Figure 2.6 by selecting a wide range of penalties and iteratively adapting this range until an elbow shape was achieved. The number of change points plotted as a function of the penalty value creates an elbow plot, as shown in Figure 2.8.

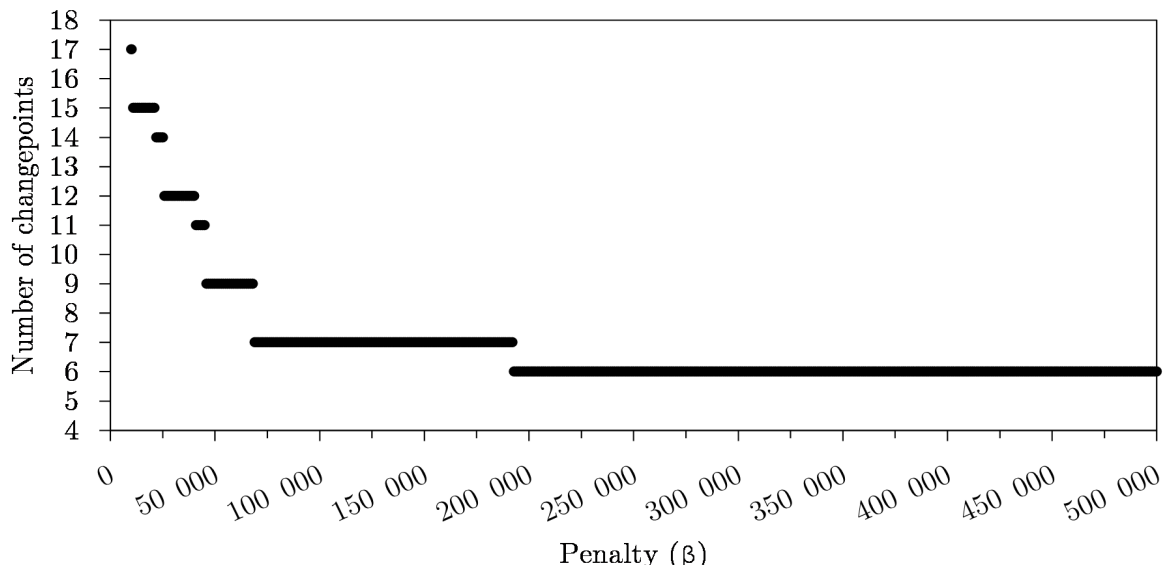


Figure 2.8: The effect of a penalty in the PELT algorithm showing how the number of change points identified changes with the given penalty range.

The reader/viewer of this figure knows that there are three change points in the weekday data based on the visual inspection of Figure 2.4. However, Figure 2.8 above suggests that there are more than five change points. It is possible that the manual penalty method is not strict enough, or that the algorithm has identified more than three true change points. Based on Figure 2.8, a penalty of 50000 corresponds to the apex and may be reasonable for the `pen.value`

input.

The next input, `method`, specifies which change point algorithm to use. If the method is set to “BinSeg”, then `Q` must be specified. This is the maximum number of expected change points and should be judged based on what is expected from the system under analysis. For the compressor example, one may use the elbow plot and set `Q = 20`. In summary, for the compressor example, the `changepoint` package inputs for BinSeg and PELT would be as follows:

```
ChangePointsBinSeg = cpt.mean(data, method="BinSeg", Q=10, minseglen=8)
```

```
ChangePointsPELT = cpt.mean(data, penalty="Manual", pen.value=50000,
method="PELT", minseglen=8)
```

The package returns the location of each change point found and the mean of that segment. These can be used to plot a change point line over the test data. Figure 2.9 shows the results from the “BinSeg” method applied to the compressor example data.

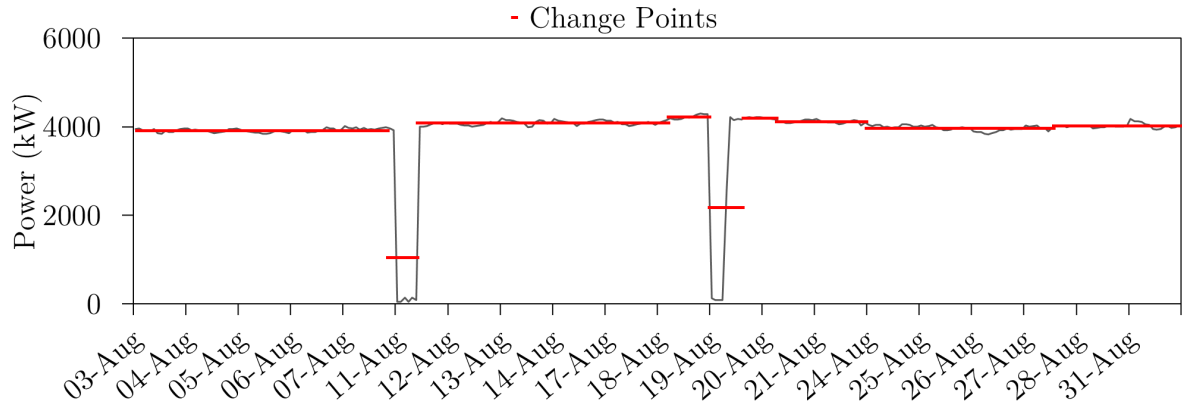


Figure 2.9: Change points discovered by the BinSeg algorithm for Eskom-delineated peak tariff data of a mining compressor’s weekday power profile.

Figure 2.10 shows the results from the PELT method applied to the example. It is clear that this algorithm identifies anomalous data points.

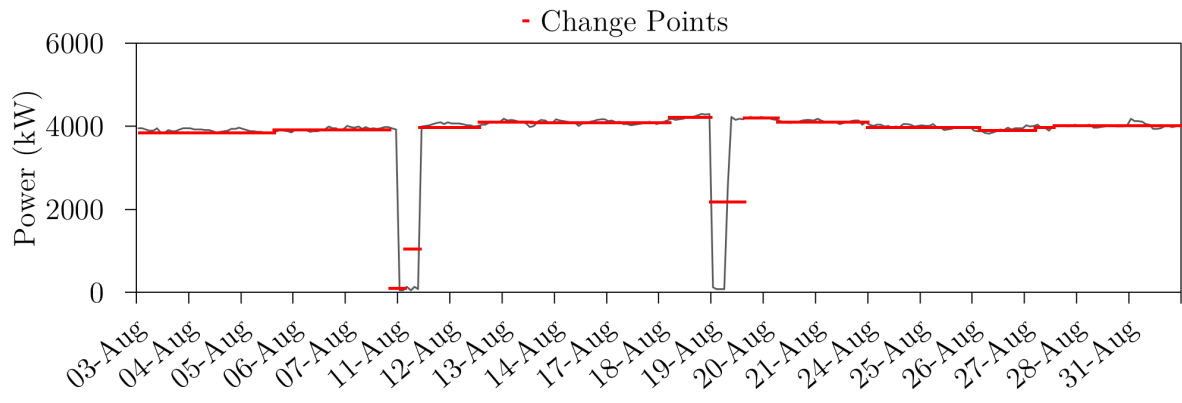


Figure 2.10: Change points discovered by the PELT algorithm for the Eskom-delineated peak tariff data of a mining compressor’s weekday power profile.

The method described above can be summarised as R pseudo code algorithms. Algorithm 2 below describes the penalty selection method. Comments are indicated with a “#” symbol; it is important to notice that the notation is slightly different from Python, such as “←” instead of “=”. The first step is to load the data into R using `import`. To start guessing, one may let *beta* range from 1.0 to 10.0. For each value in *beta*, compute the number of change points with the PELT method and store it in a list. This list is then plotted with *beta* on the x-axis to create an elbow plot. On the y-axis, the number of change points should be a realistic range. Increase the penalty — by, for example, ranging *beta* from 10 to 100 — so as to improve accuracy.

Algorithm 2 Penalty Determination

```

1: data ← import("path to directory and filename of the data")
2: beta ← list of guess values for the penalty.
3: ChangePointsPELT ← cpt.mean(data, penalty="Manual",
   pen.value=beta[i], # Where [i] fetches the value in beta at a certain index.
   method="PELT", minseglen=8)
   # Loop through each possible penalty in beta by changing the index i.
4: NrChangePoints = ChangePointsPELT.cpts # this extracts the list of changepoint locations
   from the results of cpt.mean(). The length of this list is the number of change
   points.
5: Store length(NrChangePoints) in a list and create the elbow plot.
6: Adjust the range of beta to achieve a realistic number of change points.
```

Once the penalty has been set, the PELT and BinSeg algorithms can be applied. Algorithm 3 outlines the steps for BinSeg. This method requires a guess for the maximum possible number of change points, *Q*. A reasonable value can be based on the y-axis of the elbow plot from Algorithm 2, the number of change points associated with the chosen penalty. Based on the example compressor data, a minimum segment length of 8.0 may be reasonable for mining data — this should be confirmed for each data set. The output of this algorithm, labeled *TauList* in Algorithm 3, is a vector of change point locations and the mean associated with each segment. These locations are the date and time where a change point was found.

Algorithm 3 Binary Segmentation

```

1: data ← import("path to directory and filename of the data")
2: QValue = The number of change points that correspond to the penalty derived in Algorithm 2.
3: ChangePointsBinSeg = cpt.mean(data, method="BinSeg", Q=QValue, minseglen=8)
4: TauList = ChangePointsBinSeg.cpts
```

Implementing PELT is similar to that of BinSeg and summarised in Algorithm 4. The penalty (*BetaValue*) was selected from the elbow plot as described in Algorithm 2. The output is also a vector of change point locations and the respective segment means.

Algorithm 4 Pruned Exact Linear Time

```

1:  $data \leftarrow \text{import}(\text{"path to directory and filename of the data"})$ 
2:  $BetaValue = \text{Predetermined penalty from Algorithm 2.}$ 
3:  $ChangePointsPELT \leftarrow \text{cpt.mean}(data, \text{penalty}=\text{"Manual"}, \text{pen.value}=BetaValue,$ 
    $\text{method}=\text{"PELT"}, \text{minseglen}=8)$ 
4:  $TauList = ChangePointsBinSeg.cpts$ 

```

In summary, this section described how to implement three change-point detection methods: CUSUM, BinSeg, and PELT. The various methods of implementation was demonstrated through an example of energy data concerning a mining compressor. Data preparation precedes any change point detection application. In this step, the data are reorganised according to season, weekday, and then Eskom's time-of-use periods. Doing so agglomerates periods where consistent power output can be expected, which improves change point detection. The important step for CUSUM is to select an appropriate baseline period such that it represents normal operation. The baseline data must be manually checked for data loss and outliers, and anomalous data should be removed. To implement BinSeg and PELT, a penalty is first determined. Elbow plots are created by varying guessed penalties and by plotting the number of change points. This also gives an indication of the estimated maximum number of change points required in BinSeg.

It is important to remember to terminate any background processes on the machine before running an algorithm. As explained in Section 1.4, any processes may interfere with the machine's ability to run an algorithm. If this is not done, the computation-time test metric will not be consistent. To keep track of the background processes, this study will monitor the task manager as shown in Figure 2.11.

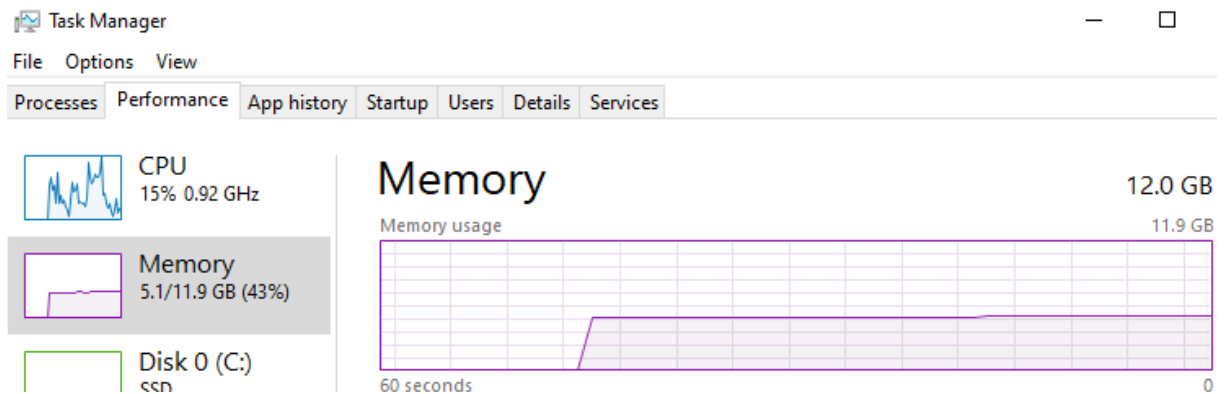


Figure 2.11: Task manager showing the memory usage on the computer to be used for change point detection implementation.

Once the data preparation is complete and all apps/processes are terminated, change-point detection algorithms can be implemented. The outputs of BinSeg and PELT are vectors that contain the date and time of change points, along with the mean of each segment. What remains are steps to verify the success, accuracy and usability of each change-point detection method.

2.5 Verification

Verification of the methodology is to prove that the steps explained in this chapter will enable one to run a change point detection algorithm that successfully and efficiently locates meaningful change points. The verification of data preparation and algorithms for NRE detection in mining energy data was gathered from literature reviewed in Chapter 1. First, the data must be reorganised to account for expected and cyclical changes [27]. In Figure 2.3, it is clear that lower demand can be expected in the example data during the first Eskom peak time of use. Therefore, this categorisation accounted for the energy reduction that occurred daily, as shown in Figure 2.3.

Second, the chosen algorithms must maintain low computational complexity with regards to the size of the data that they are applied to [55]. This does not apply to the CUSUM method, given that it performs a single operation on no more than two data points at a time [59]. According to Killick et al. [53], the BinSeg and PELT algorithms are inexpensive when working with large data sets. Therefore, the CUSUM, BinSeg, and PELT methodologies are verified to be suitable for NRE detection in mining-energy data. To conclude methodology verification, consistent testing conditions were enforced by terminating all unnecessary processes on the computer [57].

It is necessary to ensure that the chosen algorithms and methodology address the research objectives. In this study, the objective is to apply algorithms that rapidly identify non-routine events in mining energy data. An algorithmic approach addresses the goal to rapidly identify NREs because it does not rely on feedback from any site personnel. An algorithm can be run continuously and will thus identify anomalous data very near in time to the commencement of the NRE. According to the literature review, a changepoint detection algorithm is validated if it detects a historically known changepoint. The CUSUM method was proven to be suitable, as shown in Figure 2.7, the reduced power during the Christmas period was highlighted. Similarly, Figure 2.9 shows that the BinSeg and PELT algorithms successfully identified the anomalous data points on 11 and 19 August. Literature states that a CPD algorithm must be capable of identifying more than one changepoint in a data set [63]. The CUSUM method indicated reduced power on two occasions: 01 December and 25 - 31 December. The BinSeg and PELT algorithms also identified two anomalous periods in Figures 2.9 and 2.10.

The final validation will involve comparing the new algorithms to the IPMVP method. The comparison metrics include ease of implementation, runtime, and accuracy. This addresses the main study objective, which is to show that an algorithmic approach is faster and easier to implement than the prescribed IPMVP method. This can improve the current reactive and slow NRE-identification process utilised in South African mines.

2.6 Conclusion

In this chapter, the methodology to implement change point detection was developed. This was done with aid of a test case to validate the methodology. In the next chapter, the case study data will be explained. Thereafter follows the presentation and discussion of the results.

Chapter 3 Rapid change point-detection implementation and results

3.1 Preamble

The problem statement in Chapter 1 emphasises the need to closely monitor energy data in order to rapidly identify anomalous events. The body of literature reviewed has revealed important data preparation steps and data monitoring algorithms that can identify change points. In Chapter 2, the procedure was outlined from the raw data stage to the point of change-point detection. With the aid of the example compressor data, it was determined what resolution is best and which level of categorising is useful for mining-energy data. Thereafter, the steps to implement CUSUM, BinSeg, and PELT were explained. In this chapter, the methodology outlined in Chapter 2 will be applied to four case studies.

3.2 Case study explanation and data preparation

The case study in this section was chosen from the study mentioned on Page 33 in Chapter 1. The case study mine receives electricity from Eskom which is distributed into the mine via a cable system known as “feeders”. In this case, there are 300 feeders which log half-hourly power data — of these 300, the 57 most significant were identified [62]. This accounts for 80 % of the mine’s total energy consumption in the case study. In this section, the data preparation and change-point detection results will be discussed for four of these feeders.

The first case study concerns a feeder that supplies electricity to a rock mill. Figure 3.1 shows the raw data covering activity between January and June 2020 (*inclusive*). These data points represent the power output as they were logged every half-hour. This data set displays a high degree of variance while also indicating the presence of a significant period of decreased power in April of 2020.

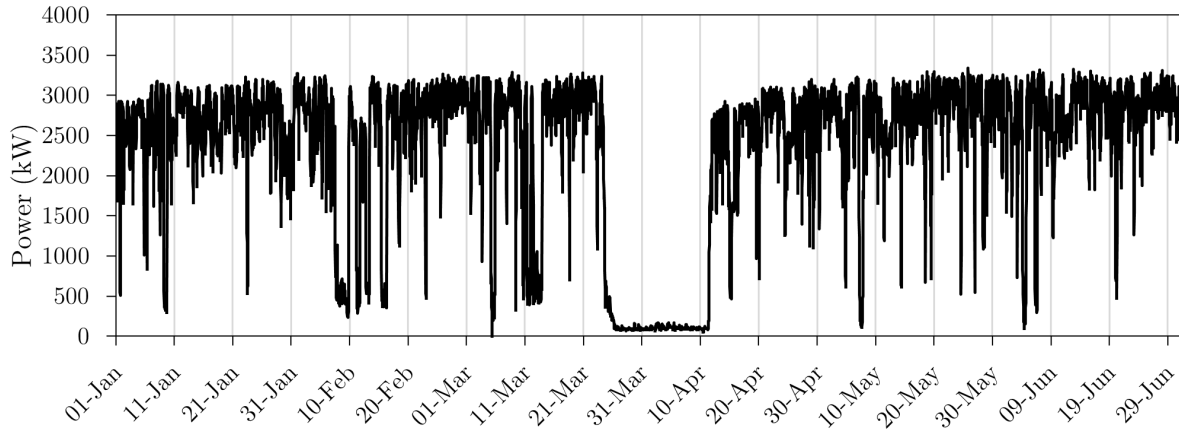


Figure 3.1: Half-hourly power output of a rock mill.

The second case study concerns a compressor which services a different shaft than that of the example compressor in Chapter 2. The raw data have been recorded at half-hourly power measurement intervals, as shown in Figure 3.2. The recorded sharp drops in power make it difficult to identify trends in the raw data, with a prolonged period of inactivity also having been recorded for the period spanning from the end of March to the beginning of May 2020.

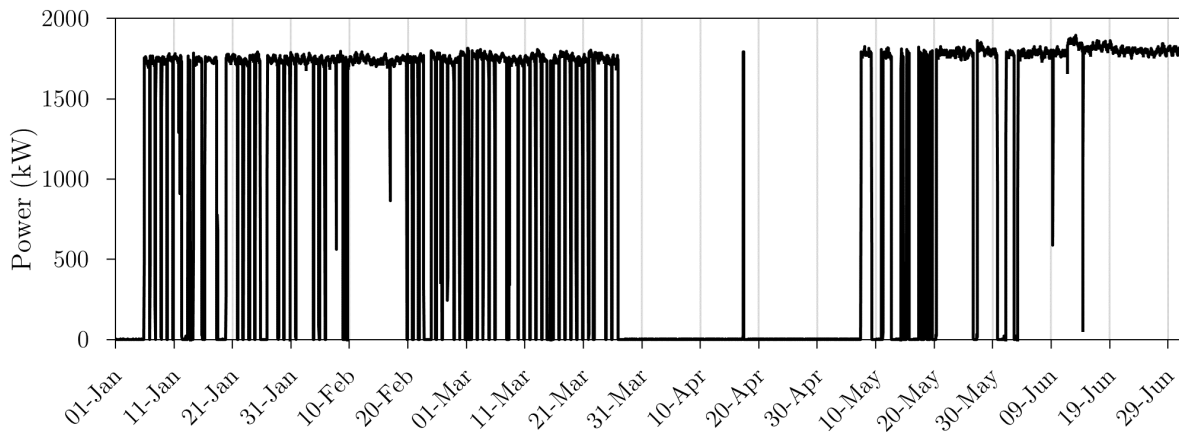


Figure 3.2: Half-hourly power output of a compressor.

The third case study is of a feeder connected to a gold plant, which extracts gold from raw materials. Various machines and processes present at the plant are supplied with electricity through this feeder. The gold plant power demand data are shown in Figure 3.3. These power readings were recorded every half-hour. At this resolution, the data appear sporadic which makes it difficult to identify trends.

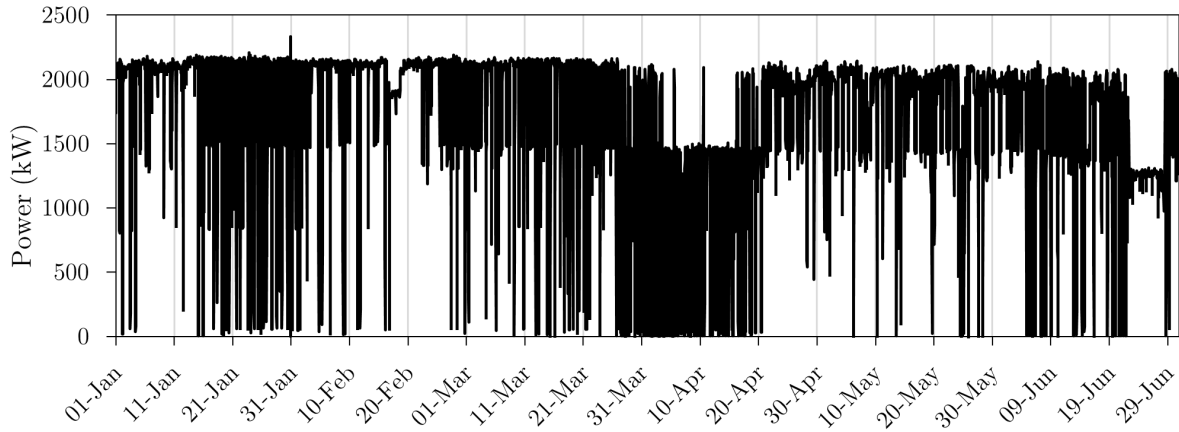


Figure 3.3: Half-hourly power output of a gold plant.

The fourth data set covers the power output from a fridge plant located in the mine. These data were also collected at a half-hourly-interval basis and have been presented in Figure 3.4. Although these data appear less sporadic than that of the gold plant above, there is a change in the variance on 10 May. It is unclear if any trend can be identified at this resolution.

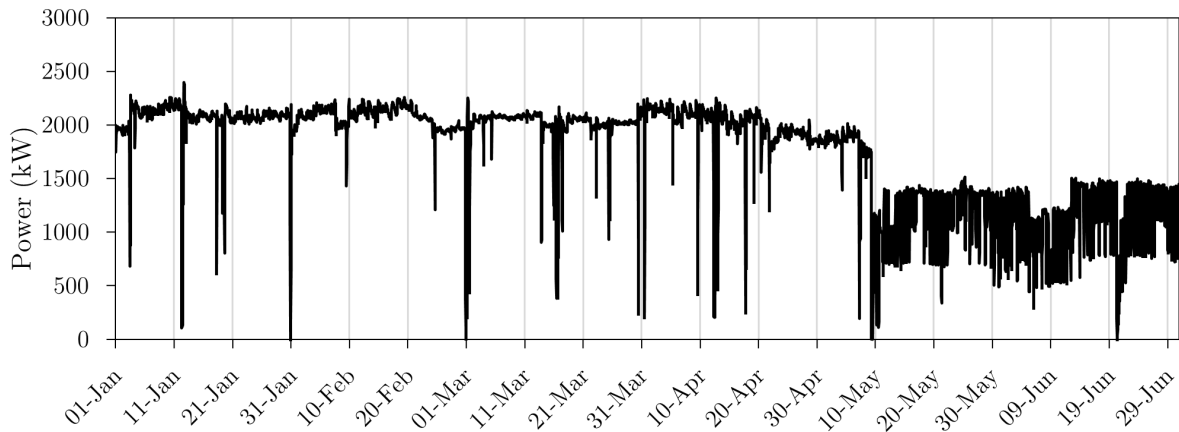


Figure 3.4: Half-hourly power output of a fridge plant.

Section 1.3.2 explained that the CUSUM method requires a baseline period [27]. For this study, February 2020 was chosen as the baseline for all four case studies, with its selection having been derived from the example data in the Chapter 2 that showed decreased consumption during the Christmas break. This continues into January 2020 given that the 1st of the month is a public holiday and is followed by a start up period [24]. Therefore January does not represent normal operating conditions, as shown in Figure 3.5.

Figure 3.5 shows that the fridge plant initially ran at reduced level of power and that the compressor was off until 6 January. As explained by Schutte, Kleingeld, and van der Zee [11], equipment and systems in a mine may affect one another. Therefore, it cannot be assumed that the mill and plant's operations are not influenced by the reduced power from the fridge plant and the compressor. Furthermore, February is a reasonable baseline period because little-to-no anomalous data is present during this period in Figures 3.1 through 3.4.

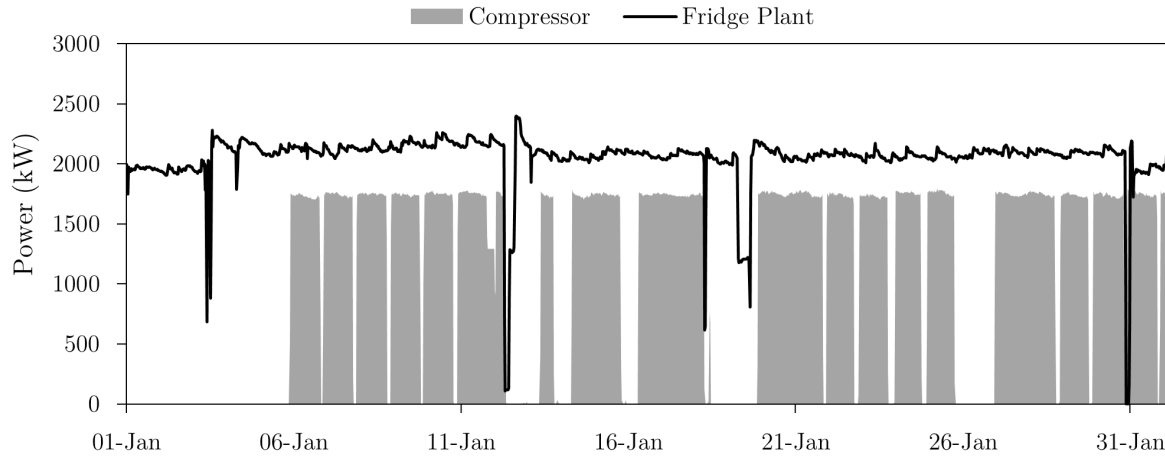


Figure 3.5: End of Christmas break start-up period in January.

In all of the case studies above, it was made apparent that the raw data do not show any patterns. The aim of data preparation and categorisation is to expose trends, with the methodology laid out in Section 2.2 being applied to the four case studies for this reason. The first step was to reassign holidays as “Saturday/Sunday”, as per Eskom’s calendar in Table 2.1.

The data were then split into season, weekdays, and Saturdays and Sundays. This was done so as to label each time according to the Eskom time-of-use periods found in Table 2.2. The next step requires the scaling of the data such that the standard deviation is approximately 1.0, as achieved via the median absolute deviation method defined in Section 1.3.4. The following discussion presents the data preparation results, and specifically for weekday Eskom off-peak periods in the summer — in short, S_WKD_OFF.

Figure 3.6 shows the formatted power output from January 2020 to March 2021. Winter months are excluded and only the off-peak time-of-use periods are plotted. Similar plots for the standard and peak Eskom time-of-use periods are in Appendix Figures C.1 and C.2.

This configuration clearly highlights the dip in consumption in April 2020 of plots (a) and (b). In plot (b), the scaled parameters are a mean = -27 and a standard deviation = 39 . This process did not achieve a standard deviation of 1.0, which is necessary for change-point detection as stated on Page 18. However, in this study, no further data formatting will be applied. In plots (c), the sporadic data points seen in Figure 3.3 are not present. This makes it possible to see the period of lower demand in April 2020. Plot (d) shows increased demand near the end of November 2020 which may not be visible in the raw data. It remains to be seen if these statistical parameters will hinder the performance of the change-point detection algorithms

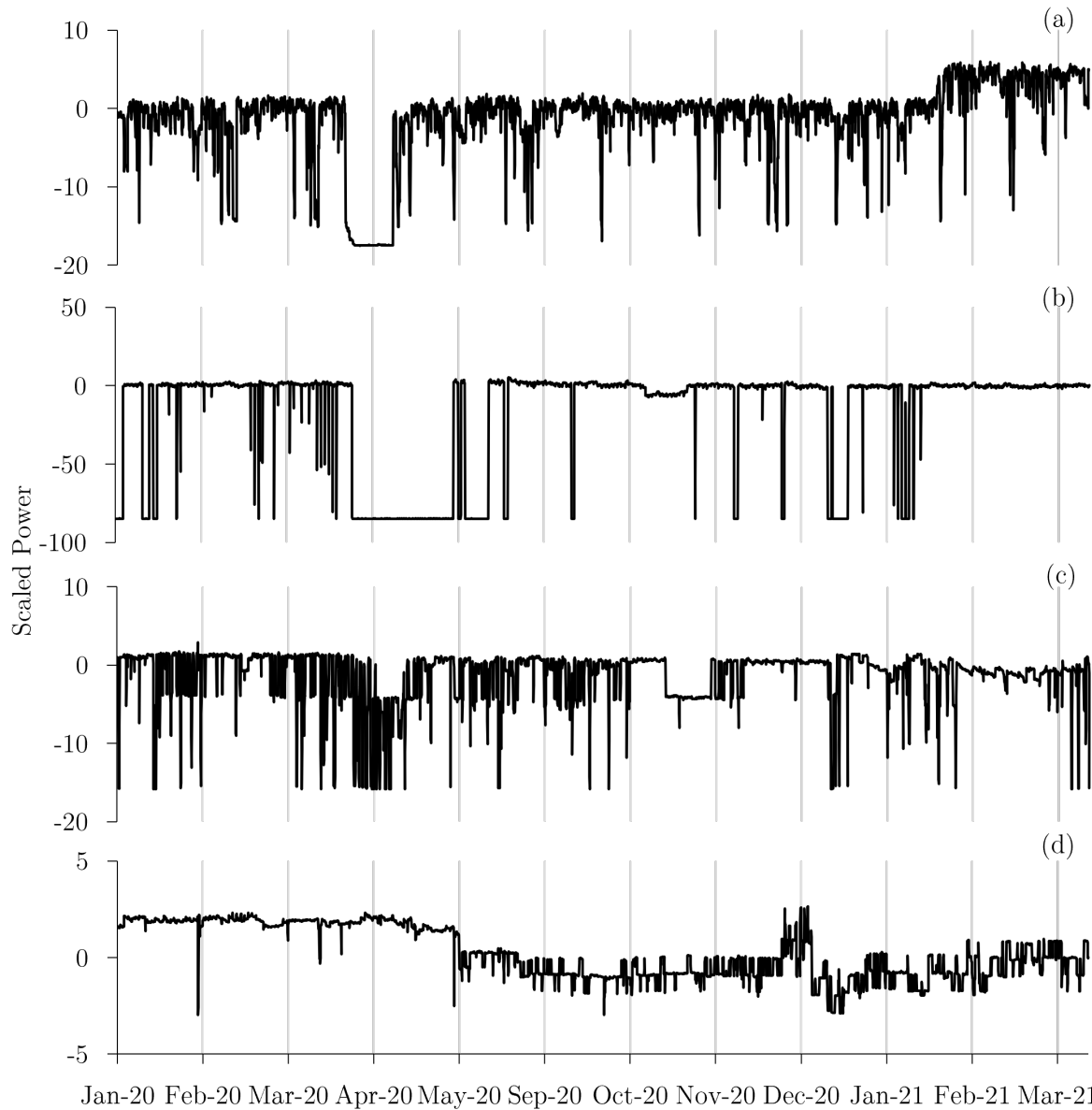


Figure 3.6: Scaled Summer weekday off-peak time-of-use plot of (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.

To apply the CUSUM method, the baseline data must be cleaned — as described on Page 39. The standard deviation for February 2020 was calculated with Equation 1.3. Each point in Figure 3.7 is $x_i - \mu$ divided by the standard deviation. These plots are of the S_WKD_OFF periods in February 2020 and show the outliers on the following days:

- (a) The 7th and 11-13th.
- (b) The 20-25th and 27th.
- (c) The 3rd, 6th, 10th, 13-14th, and 26-28th.
- (d) The 12-14th, 18-19th, and 24th.

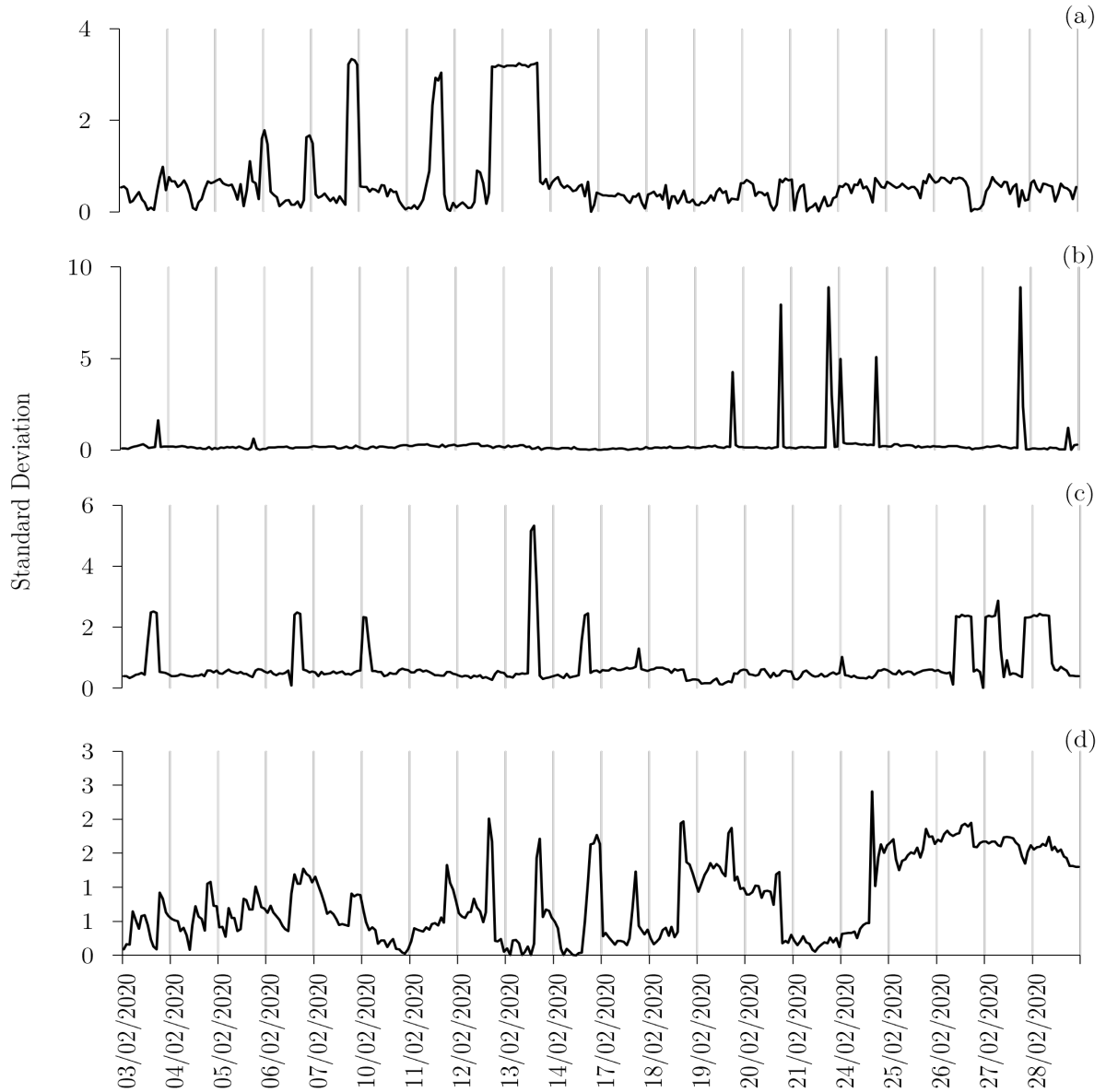


Figure 3.7: S_WKD_OFF standard deviation plots of February 2020 for (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.

The data points that fall outside of the 2σ limit were removed. Thereafter the data was scaled with the MAD method described on Page 18 and plotted in Figure 3.8. The resulting data provide the statistical parameters for the CUSUM method. These parameters are summarised in Table 3.1.

Table 3.1: Baseline statistical parameters for CUSUM-anomaly detection.

Parameter	Rock mill	Compressor	Gold plant	Fridge plant
μ	-1	0	-2	0
σ	2	3	5	2

The scaling process ensured that most data points are within one standard deviation of the mean and, therefore, satisfies the condition set in Section 1.3.1. With this condition satisfied, the data in Figure 3.8 present suitable baselines for CUSUM-anomaly detection. On Page 8, the literature reviewed [5][27] explained that the standard deviation of a suitable baseline period can be used to identify outliers in the assessment period. The baseline preparation was also applied to Eskom standard and peak time-of-use periods. This can be seen in Appendix Figures C.3 and C.4. The following section describes the results of the CUSUM method.

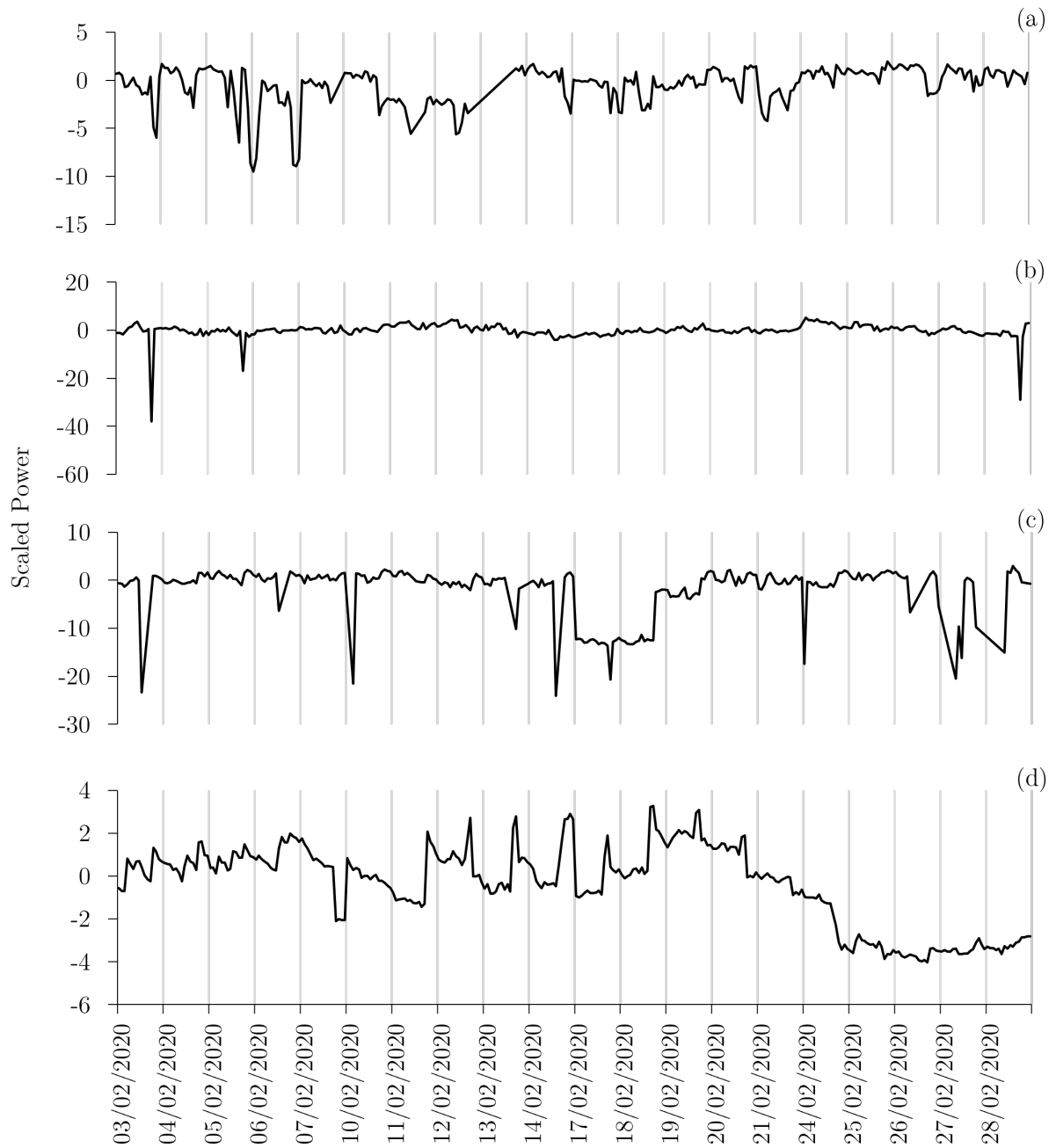


Figure 3.8: S_WKD-OFF cleaned and scaled baseline plots of February 2020 for (a) a rock mill, (b) a compressor, (c) a gold plant, and (d) a fridge plant.

3.3 IPMVP anomaly detection

In short, this IPMVP method utilises the cumulative sum of each data set's Z-score. The constants summarised in Table 3.1 were substituted into the CUSUM Equations 2.1, 2.2, and 2.3 for each, with the case study power data plotted above the respective CUSUM plot. On the x-axis, the most significant dates are those leading up to the COVID-19 shut down on 26 March 2020. Although the scaled data were used when applying the change-detection method, they would not add comprehensive value upon visual inspection of the results. For this reason, the original power output for each section was used to visually display the results.

In Figure 3.9, the negative (S_L) CUSUM of the rock mill is plotted. No positive change points were identified during this time period and the S_H CUSUM plot remained at null. However, the S_L plot in Figure 3.9 indicates negative accumulation. Dips in power output were detected on 05 March and 10-12 March. This correlates with the reduction seen in the power plot. As such, the national lockdown was identified successfully, as indicated by the decreasing slope of the CUSUM plot.

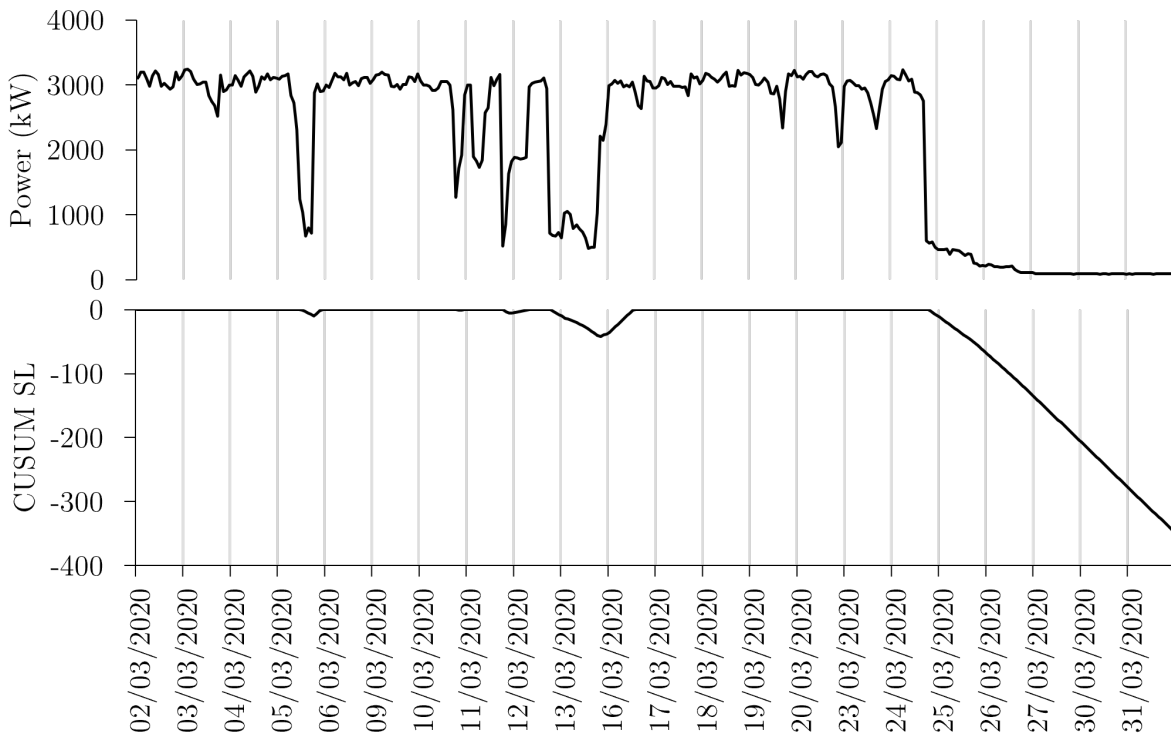


Figure 3.9: Top: S_WKD-OFF power output of a rock mill in March 2020. Bottom: S_L CUSUM result showing detection of anomalies.

Similar to the previous case study, no positive accumulation was identified for the compressor. Figure 3.10 shows the power output of the compressor and the corresponding S_L plot. The national lockdown was correctly identified; however, the extent of this negative accumulation obscures the sharp dips present on previous days.

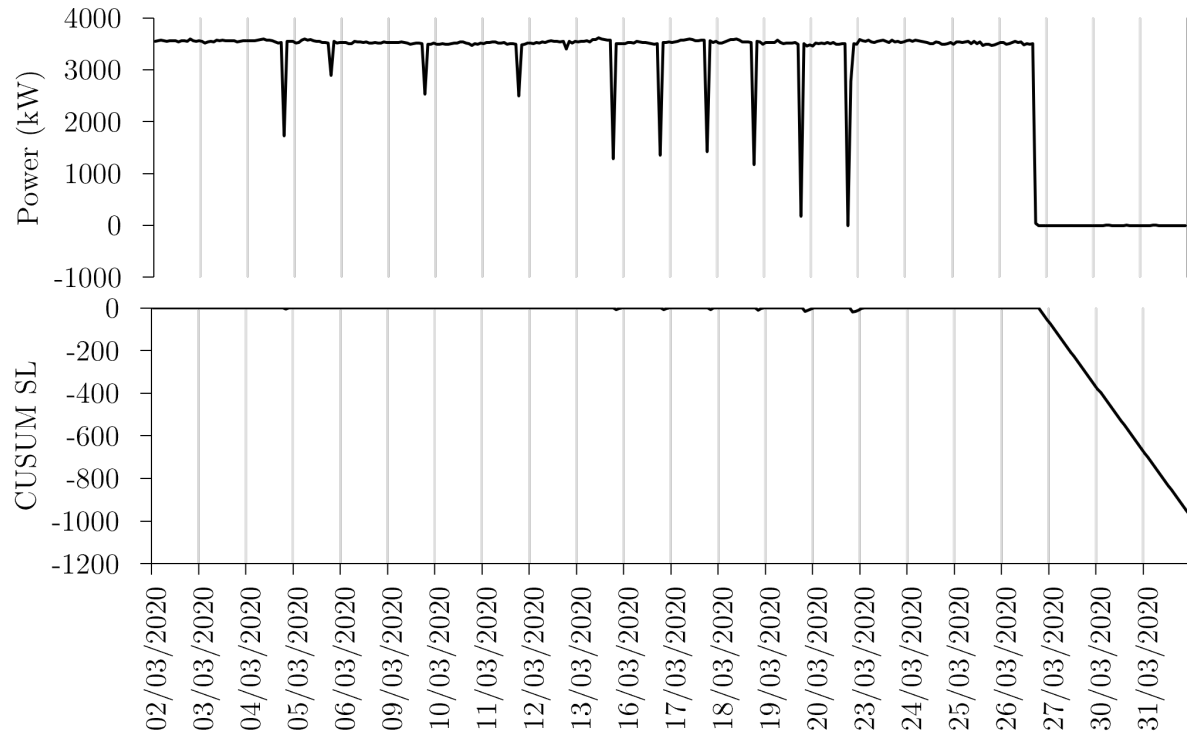


Figure 3.10: Top: S_WKD_OFF power output of a compressor in March 2020. Bottom: S_L CUSUM result showing detection of anomalies.

The IPMVP method did not detect any positive or negative accumulation for either the gold plant or the fridge plant in March 2020. In Figure 3.11, the power output dips to nearly zero on multiple occasions. However, no negative accumulation developed. The reason is due to the variability of the raw data. The standard deviation of the baseline period is large enough such that the data points in Figure 3.11 do not exceed 2σ .

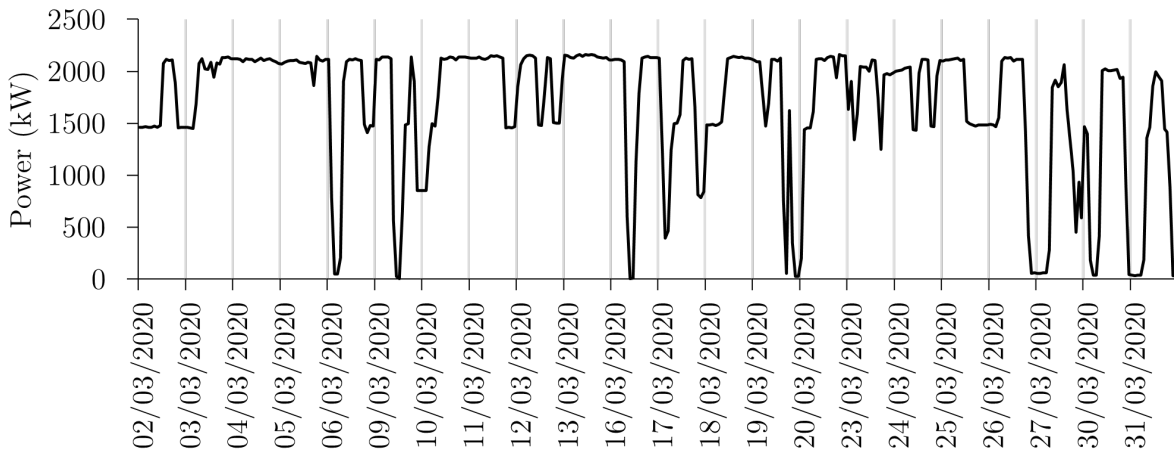


Figure 3.11: S_WKD_OFF power output of a gold plant in March 2020.

Figure 3.12 shows that the power output of the fridge plant did not increase or decrease out of bounds during March 2020. However, the dip in power output on 30 March was not identified. The threshold for the CUSUM method is 2σ , where σ is the standard deviation of the baseline period. It is evident that the baseline data had a high degree of variability and that 2σ was, thus, not exceeded during the assessment period.

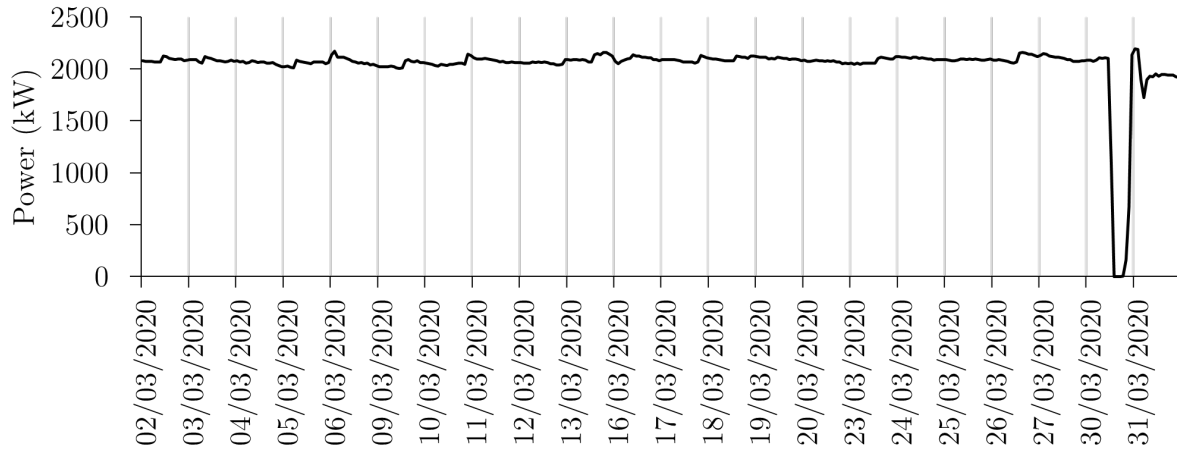


Figure 3.12: S_WKD-OFF power output of a fridge plant in March 2020.

Overall, the CUSUM method is capable of detecting change points. However, this method becomes less accurate when there is high variability present in the baseline data. In the following section, the results of the BinSeg and PELT algorithms are analysed.

3.4 Change-point detection algorithms

In this section, the results of the BinSeg and PELT algorithms are discussed, with these algorithms having been applied to each case study — the relevant S_WKD-OFF periods will, thus, be shown in this section. The results of the S_WKD-STD and S_WKD-OFF periods can be found in Appendix D. These methods require parameter-tuning to limit the number of change points to that of a realistic number. The three parameters that must be tailored to each data set are the minimum segment length, penalty, and maximum number of change points (Q).

From the literature reviewed, a significant change point should last a duration of four hours. The case study data are measured in half-hour increments and, therefore, the minimum segment length was chosen to be that of 8.0. With this parameter in place, a penalty was determined for each data set via the use of the elbow plot method which was applied to the March 2020 portion of the scaled-power data sets that were developed in Section 3.2. The plots in Figure 3.13 were created by varying the penalty in the PELT method. The number of change points identified by the PELT algorithm were plotted as a function of penalty values. The penalty values were created by guessing a range of penalties and adjusting this range until an elbow-shaped curve was achieved.

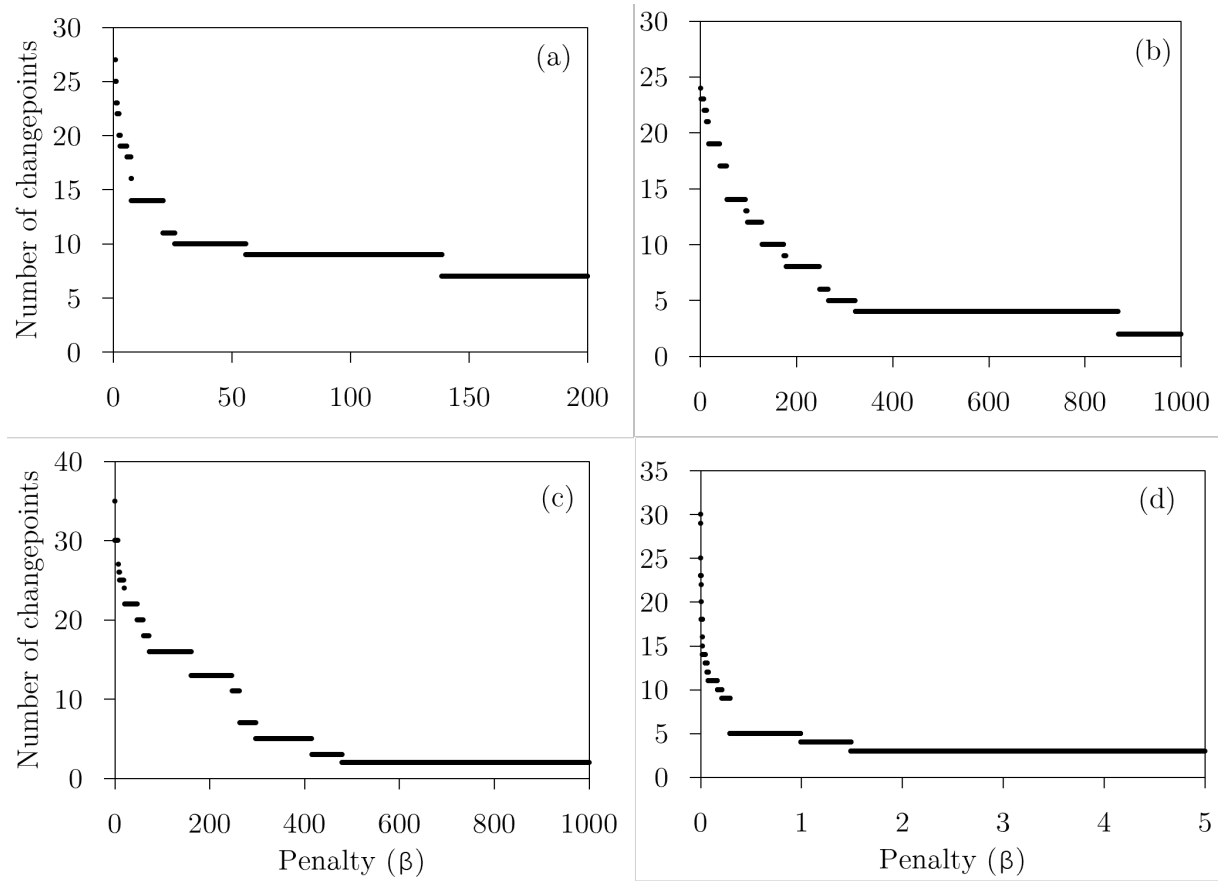


Figure 3.13: Elbow plots showing the number of change points along varying penalties in the PELT method for the March 2020 S.WKD.OFF power data for: (a) a rock mill, (b) a mining compressor, (c) a gold plant, and (d) a fridge plant.

Figure 3.13 (a) shows that the apex of the elbow plot is at the coordinate location (30,10). This means that the PELT algorithm identifies 10.0 change points when the penalty is set to 30.0. The apex of plot (b) lies somewhere between points (200,8) and (300,4). Upon referring to Figure 3.6 (b), little variation can be seen for the period ranging between February and March 2020. Therefore, the more strict penalty of 300 was chosen for the mining compressor data set. For the gold plant data, a penalty of 260 was chosen based on Figure 3.13 (c). Figure 3.6 (c) shows variation during February and March; therefore, an estimation of 10.0 change points may be reasonable. There is significant discontinuity along the y-axis in Figure 3.13 (d) where the number of change points reduce from 9.0 to 5.0 as the penalty increases from 0.293 to 0.294. It is not clear which penalty should be used — therefore, both 0.293 and 0.294 will be discussed for the fridge plant data.

The BinSeg method requires that a maximum number of change points be set. According to Section 2.4, this limit can be based on the elbow plot method. The maximum number of change points for each data set was selected from Figure 3.13 and summarised in Table 3.2.

As derived from the review of the extant body of literature, it is crucial to maintain consistent testing conditions when comparing algorithms. This means that any background process must

Table 3.2: Summary of maximum change point limits for the BinSeg algorithm.

	Rock Mill	Compressor	Gold Plant	Fridge Plant
Max CPT	10	4	10	9

be turned off, if possible. Section 2.4, showed that the task manager on a system using Windows can be used to achieve this outcome. Figure 3.14 shows what portion of background processes were active while implementing the CPD algorithms. The little memory that was used was to maintain critical processes, without which the computer could not otherwise function.

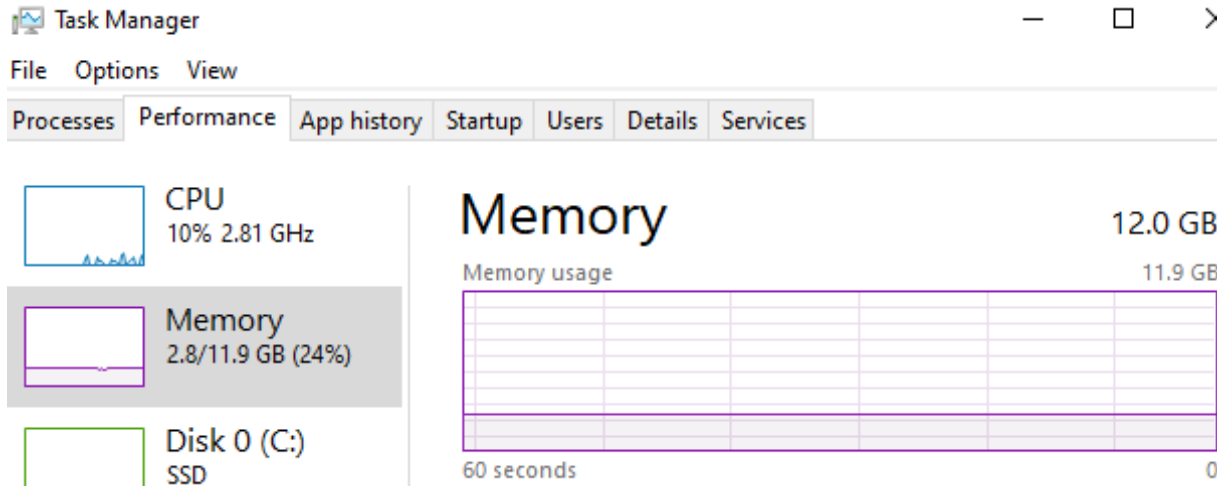


Figure 3.14: A snapshot of the memory usage on the machine that was used to implement change-point detection algorithms.

With consistent testing conditions ensured, the change-point detection algorithms were implemented. In the following section, the results are presented visually for the four case studies. The convention in this section is that the original assessment power data for March 2020 are plotted first. Below that, the scaled power data from Section 3.2 are presented as an area graph — this is the data set used in the CPD algorithms. The CPD results for BinSeg and PELT are plotted over the scaled area graph. These lines represent the estimated mean of each segment between change points. The following discussion will focus on the summer weekday off-peak Eskom time-of use-periods — refer to Appendix D to see the results of the standard and peak periods.

Figure 3.15 shows the BinSeg and PELT CPD results for the rock mill case study. It is clear that both algorithms detected the national lockdown from 25 March 2020. The BinSeg algorithm located five change points, and the PELT algorithm identified eight. The CPD results for the mining compressor are shown in Figure 3.16. Keeping the scaled power data as an area graph did not simplify the visual representation and was, therefore, changed to a line. Both algorithms successfully identify the periods when the compressor turned off for the national lockdown from 27 March 2020.

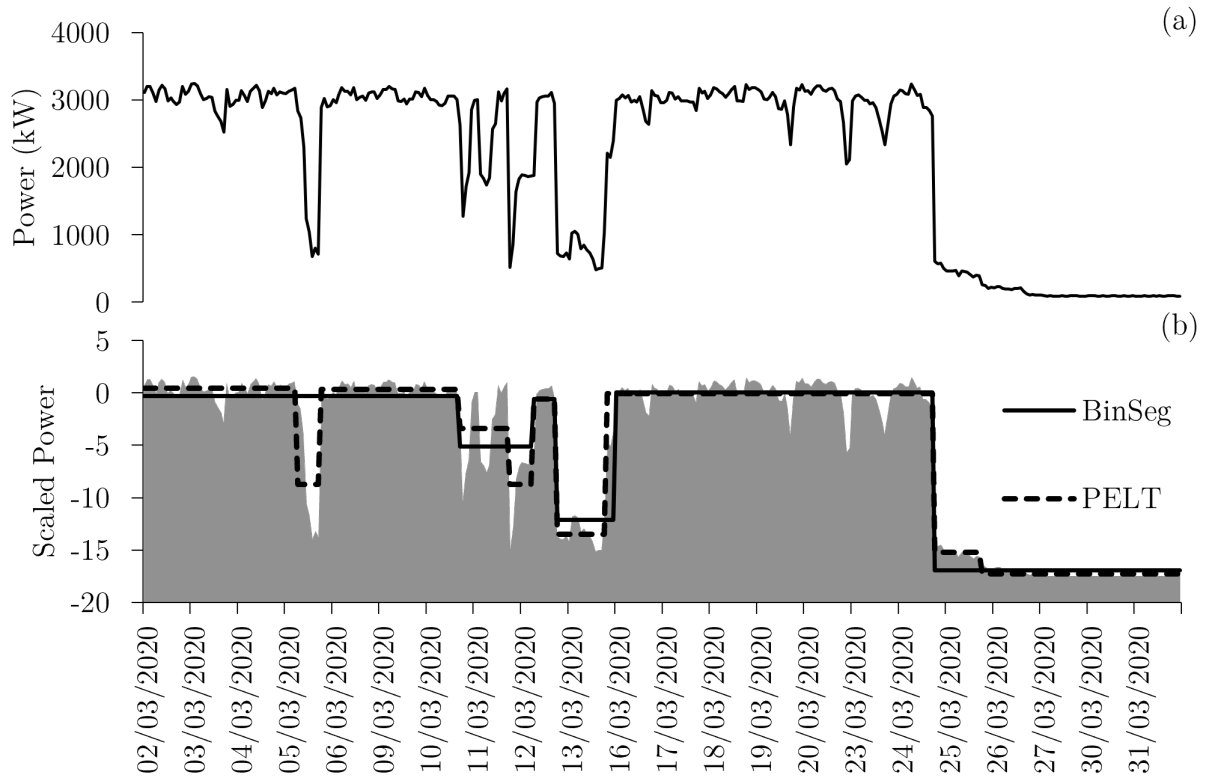


Figure 3.15: (a) S_WKD_OFF power output in March 2020 of a mining rock mill; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.

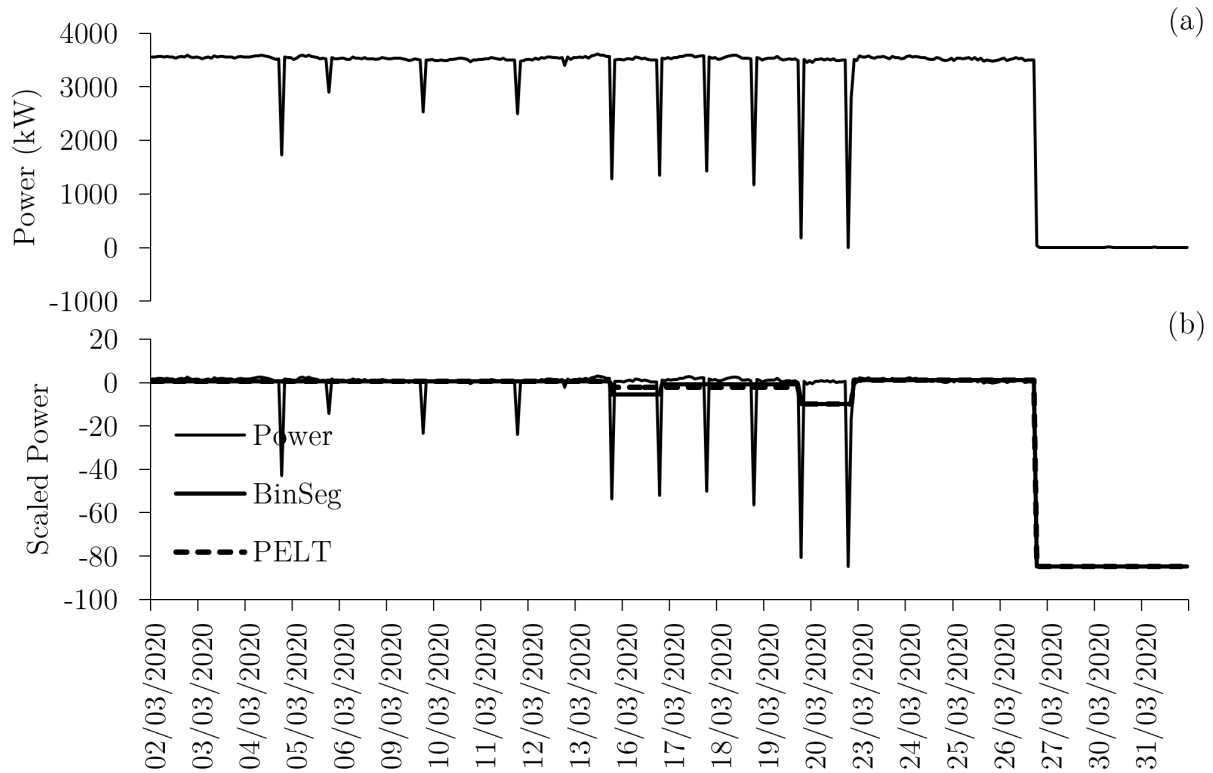


Figure 3.16: (a) S_WKD_OFF power output in March 2020 of a mining compressor; (b) change-point detection results along with the scaled power data from plot (a).

Figure 3.17 shows a closer view of Figure 3.16. The short duration dips in power were not identified as change points. This shows that the minimum duration setting of four hours prevents

the identification of insignificant change points. However, the lower power values bring down the mean of that segment and make the CPD results seem less accurate when compared to that of the power data. Furthermore, the sharp drops in power obscure the overall trend and renders it difficult to discern the segments identified via the CPD algorithms.

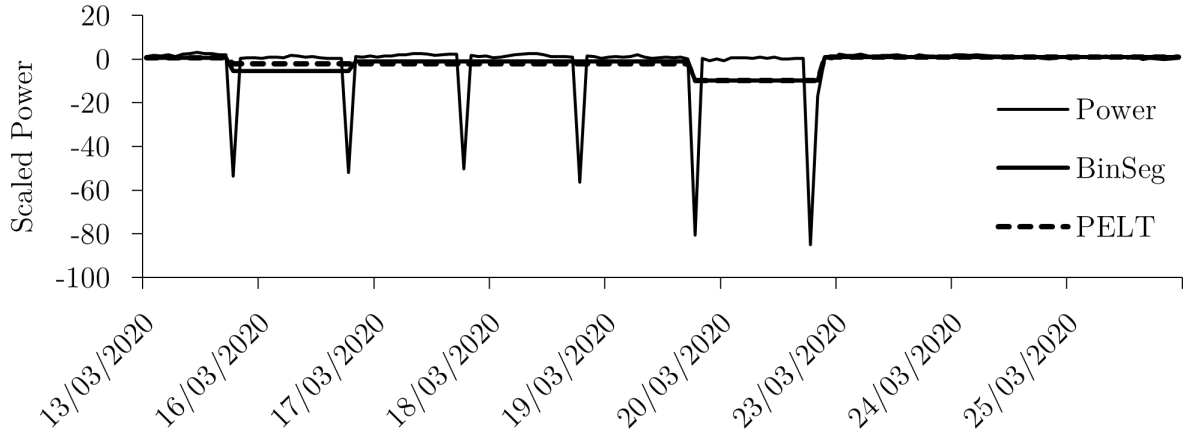


Figure 3.17: (a) S.WKD.OFF power output of a mining compressor from 13-25 March 2020; (b) change-point detection results along with the scaled power data.

The gold plant power data for March 2020 are plotted in Figure 3.18 (a). Although the elbow plot estimation showed that 10.0 change points can be expected for this data, the BinSeg algorithm converged to a value of 5.0. However, the PELT algorithm located 10.0 and is, thus, more robust when there is variation within a segment.

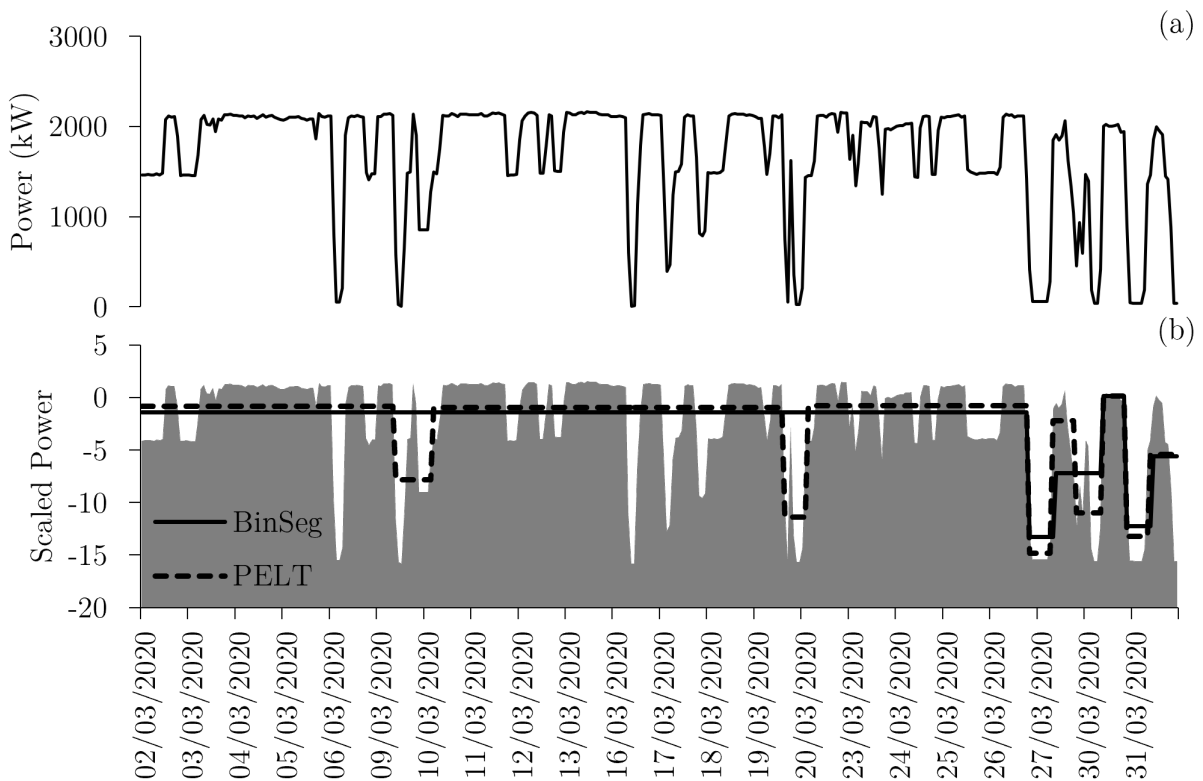


Figure 3.18: (a) S.WKD.OFF power output in March 2020 of a gold plant; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.

For example, refer to 9 March in Figure 3.18 (b), the PELT algorithm shows a segment of lower power output. There is a spike in the centre of this segment which would cause the segments on either side to be less-than the minimum valid change-point duration. For this reason, the BinSeg algorithm did not classify either segment as a change point. Neither algorithm identified a change point around 6 March. This was unexpected given that there was a reduction in power at this time — as shown in Figure 3.18 (b). In the case of PELT, the constraints limited the number of change points to 10.0 — therefore, the 10.0 most critical changes were highlighted.

The fridge plant results are shown in Figure 3.19. The BinSeg algorithm did not identify any change points. In this case, the BinSeg algorithm has failed. The PELT algorithm classified nine segments and, thus, identified inconsistent operation on this system. Furthermore, a decrease in production was expected during the COVID-19 national shutdown. The PELT algorithm would have alerted mining personnel of the increased power consumption seen on 31 March.

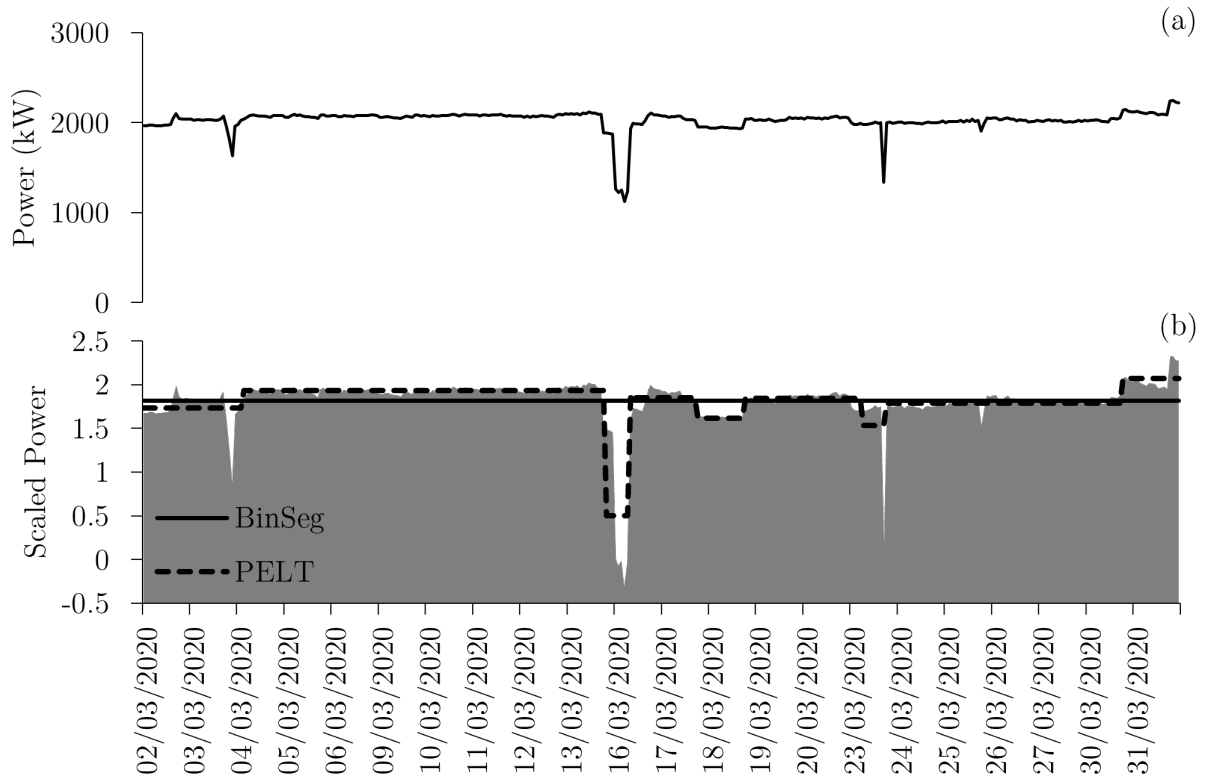


Figure 3.19: (a) S_WKD_OFF power output in March 2020 of a fridge plant; (b) the change-point detection results overlaid on the scaled version of plot (a) as an area graph.

This concludes the results discussion. What remains is a discussion of the validation of the chosen change-point detection algorithms. This will be covered in the following section.

3.5 Validation

In Section 2.5, the validation metrics for change-point detection algorithms were outlined: Namely to detect a historically-known change point and to rapidly detect NREs in mining-

energy data. It was also necessary to confirm that the algorithms show low time complexity.

From the results discussed, it was made clear that all three algorithms successfully identified a historically-known change point — the national lockdown. The main objective of this study was to test the implementation of algorithms so as to rapidly identify non-routine events (NRE) in mining-energy data. The literature review indicated that NRE detection in industry and mines can take multiple days. The methods tested in this study would reduce this time-frame as it is capable of locating anomalies within a few hours of the event occurring. This is as a result of the segment lengths of change points in mining-energy data, which are recorded in hourly intervals.

To test time complexity, the runtime for the BinSeg and PELT algorithms were measured. These are summarised in Appendix Table D.1. For all of the data sets, both the BinSeg and PELT algorithm runtimes remained less than 25 ms. Therefore, the time complexity is sufficiently low for mining-energy data at a half-hourly resolution.

Another objective of this study is to develop M&V 2.0 techniques which aim to improve the utilisation of large scale data. This was addressed through application of a prescribed IPMVP technique and change-point detection algorithms. The following discussion will focus on the comparison of the CUSUM method with the BinSeg and PELT algorithms. The comparison metrics are “ease of implementation” and “accuracy”.

The CUSUM formulas are much simpler than that of the BinSeg and PELT algorithms. To understand the CUSUM method, basic mathematical knowledge is required. In comparison, an understanding of the BinSeg and PELT fundamentals requires background knowledge of algebra and statistical principles, namely: likelihood and probability.

In terms of implementation, the CUSUM method requires an extra step in data-preparation. A baseline must be chosen for every data set, with this data undergoing an analysis to remove any outliers. The next step is to calculate two constants for each baseline: mean and standard deviation. Thereafter, Equation 2.1 must be applied to calculate the Z -score of each data point. Lastly, there are two outlier tests in the CUSUM method, used to identify positive and negative accumulation. Therefore, CUSUM results for each data set consist of two graphs showing the positive and negative accumulation of the Z -score if it was larger than two standard deviations.

One criticism of this method is that there is no “recovery” term. In reference to Figure 3.9, and as per the definition, the CUSUM values will keep decreasing as long as the data points continue to lie beyond the threshold. However, this is inconvenient when interpreting the results. The y-axis scale can become very large and may obscure change points that lasted for a shorter period. Suggestions to improve this will be addressed in chapter 4.

The implementation of BinSeg and PELT was done in a programming language called **R** with no prior knowledge. The code to implement BinSeg and PELT already existed in a package called **changepoints**; as such, it was not necessary to re-code the algorithms. However, these

algorithms required a penalty selection step for each data set. This involved running the PELT algorithm with varying penalties to create the elbow plots. The BinSeg and PELT results are a list of locations where a change point was found, along with the respective mean of each segment. The visual application of this process enables the reader to see defined regions within the data.

In terms of accuracy, the PELT and BinSeg algorithms are superior to the CUSUM method. For example, in Figure 3.11, the CUSUM method did not indicate any change points. This was related to the large variance in the baseline data and, thus, none of the assessment-period data points exceeded two standard deviations. Therefore, the CUSUM method is not suitable for data sets that show high degrees of variability.

The PELT algorithm is more accurate than BinSeg. Overall, it was able to identify more change points than the BinSeg algorithm, with the exception of figure 3.17. In this case, BinSeg located five change points, whereas PELT identified four.

3.6 Conclusion

In this chapter, the change point detection methodology was applied to a case study mine. Specifically, the energy data of four significant systems were presented. The methodology steps were to: categorise the data into Eskom time of use periods, select a baseline period for the CUSUM method and to apply the chosen algorithms. The Concluding chapter hereafter will revisit the problem statement and objectives and present a critical analysis of this study.

Chapter 4 Conclusion and recommendations

4.1 Preamble

This study implemented change-point detection algorithms on mining-energy data. The aim was to determine if the investigated methods reduce the identification time of non-routine events. In this chapter, the problem statement and objectives are summarised. Furthermore, some points of criticism, along with points concerning successes, are raised. This is followed by recommendations for further study. Lastly, the potential impact of the study is discussed.

4.2 Study overview and problem statement

Chapter 1 discussed that a number of initiatives that aim to reduce energy consumption and waste are currently implemented in South African mines. The impact of such projects has typically been quantified via standardised M&V procedures. One prevalent roadblock is that of an unforeseen or non-routine event (NRE), which formed the topic of this study and the subsequent formulation of the problem statement: in industry and mines, NREs are often discovered days or weeks after they occur. Therefore, this study aims to improve upon the slow detection time of NREs by investigating the application of algorithms on South African mining-energy data. Advanced-metering infrastructure has been widely implemented throughout South African mines. Such infrastructure includes energy meters on equipment that log data every 30 minutes, resulting in large data sets along with a need to analyse and utilise the information within. Thus, the objectives of this study are:

1. Apply M&V 2.0 techniques to mining-energy data so as to justify that algorithm-driven methods can rapidly extract information from big data sets.
2. Investigate whether change-point detection algorithms can rapidly identify NREs in mining-energy data.
3. Determine if change-point detection algorithms can identify a historically-confirmed NRE present in mining-energy data.
4. Scrutinise the ease of implementation and the suitability to large data sets in terms of computational time complexity.

4.3 Objectives addressed

The first objective is expansion of the advanced M&V, or M&V 2.0, field of study. This is addressed via a comparison of algorithms along a prescribed IPMVP method. The second objective concerns an investigation into, and subsequent application of, algorithms capable of anomaly-detection. In this study, three methods are performed: cumulative-sums (CUSUM), binary-segmentation (BinSeg), and pruned exact linear time (PELT). The algorithms were validated by ensuring that a historically-confirmed NRE was successfully identified.

The last objective is to compare the algorithms according to their ease of implementation and computational time complexity. The chosen algorithms have predefined time complexity, as discussed in Section 1.4. Additionally, the time complexity was tested by measuring the runtime of each algorithm. This investigation confirmed that each algorithm will not run too slowly when applied to mining-energy data. The ease of implementation was addressed by scrutinising the methodology and the prior knowledge required to understand and implement each algorithm. This critical analysis is extended to the success of the study in the following section.

4.4 Critical analysis of the study

The literature review provided insights into effective data preparation. One critical step was the need for dividing the raw data into time-of-use periods, enabling clear divisions of consistent operational periods. Therefore, each data set was organised to clearly determine what power output is typically expected in each Eskom-delineated tariff period. Doing so reduced the change in identifying false change points that could otherwise be attributed to expected differences in power output. These expected changes were correlated to more- or less-expensive Eskom tariffs.

The literature focus concerning IPMVP techniques did not include an implementation of the CUSUM method on a real world study. Such a study should have been included so as to attempt implementing a tried-and-tested CUSUM methodology. This would possibly improve the CUSUM results and lead to more change points being identified. The second point of criticism involves the parameters of the algorithms.

The CUSUM algorithms requires thresholds: in this study, the threshold was derived from the reviewed body of literature and remained constant across each case study. The CUSUM algorithm failed to identify any change points in gold plant and fridge plant data sets. This could be improved by researching and applying a method to tailor the thresholds needed for each data set. More specific limits may have increased the sensitivity of the CUSUM technique as based upon the relevant data set.

The BinSeg and PELT algorithms require a minimum segment length parameter. This prevents insignificant, momentary spikes or dips from being identified as a change point. In this

study, the minimum segment length remained constant for all of the case studies. This could have been improved upon through an investigation into the equipment that generated the energy data and by tailoring a minimum segment length accordingly. For example, it is possible that a momentary spike could indicate a surge in demand on the mining compressor. In this case, a surge could damage the equipment, requiring that a minimum segment length to be set to a smaller number. Alternatively, the minimum segment length could be a function of the power output of the equipment.

It is suggested that an investigation be done to identify realistic segment lengths for change points present in mining data. The minimum duration of time for a change point to be considered significant should be related to the characteristics of the mining equipment under measurement. This knowledge can then also be applied as a data preparation step. For example, the momentary dips in Figure 3.17 may have hindered the detection of significant change points. It is possible that no non-routine events occurred in the mining compressor between 13 and 25 March 2020. If it were known that such dips are not of concern, they could have been removed from the assessment data.

4.5 Study impact

The impact of rapid NRE detection can be quantified. It is possible to determine the cost implication of a historical NRE. The Eskom tariff book contains the R/kWh cost for every time-of-use period. Therefore, if an NRE caused an increase in energy consumption, the data from that period can be used to calculate the increased cost. The national lockdown prevented any mining from taking place. Therefore, it was not possible for the case study mine to generate revenue during this period. As a result, it was crucial that the energy consumption remained at the lowest reasonable level.

In Figure 3.18, the power output was reduced to nearly 0 kW. However, the power increased three times between 27 and 31 March 2020. Similar spikes occurred in the standard and peak time-of-use periods shown in Appendix figures D.7 and D.11. The BinSeg and PELT algorithms identified all three spikes. To quantify this impact, the total energy consumed from the 29th - 31st was calculated (27th and 28th were weekend days). With the assumption that the system should have been turned off completely, the electricity cost of this period amounted to R27.000/day. The case study mine has 300 feeders that log power demands similar to that of the gold plant. It is possible to apply a change-point detection algorithm to every energy data source, allowing for the a more rapid flagging of anomalous data.

Bibliography

- [1] University of Utah. “Functions.” Accessed on: Jul 11, 2021. (2021), [Online]. Available: <https://www.cs.utah.edu/~germain/PPS/Topics/functions.html>.
- [2] ORACLE. “What is a package?” Accessed on: Jul 11, 2021. (2021), [Online]. Available: <https://docs.oracle.com/javase/tutorial/java/concepts/package.html>.
- [3] R. Ihaka and R. Gentleman. “R.” Accessed on: Jul 11, 2021. (2021), [Online]. Available: <https://www.r-project.org/about.html>.
- [4] J. Granderson, S. Fernandes, E. Crowe, and S. Earni, “Ipmvp’s snapshot on advanced measurement & verification,” Lawrence Berkeley National Laboratory, Tech. Rep., 2020, Accessed on: Jul 04, 2021. [Online]. Available: https://evo-world.org/images/corporate_documents/NRE-NRA_White_Paper_Final_2701.pdf.
- [5] J. Granderson, E. Crowe, S. Earni, and S. Fernandes, “An introduction to the international performance measurement and verification protocol (ipmvp),” *International performance measurement and verification protocol*, vol. 104000, pp. 1–7, 2020.
- [6] S. Touzani, B. Ravache, E. Crowe, and J. Granderson, “Statistical change detection of building energy consumption: Applications to savings estimation,” *Energy and Buildings*, vol. 185, pp. 123–136, 2019.
- [7] S. Earni and P. Therkelsen, “Non-routine adjustments – towards standardizing m&v approach for quantifying the effects of static factors,” *Energy Technonolgies Area*, 2020.
- [8] Gold Fields. “South deep gold mine.” Accessed on: Jul 28, 2021. (2021), [Online]. Available: <https://www.sibanyestillwater.com/business/southern-africa/gold-operations/kloof/>.
- [9] Sibanye Stillwater. “Sa gold operations - kloof.” Accessed on: Jul 28, 2021. (2021), [Online]. Available: <https://www.goldfields.com/reports/annual-report-2018/mrr/sa-south-deep-gold-mine.php>.
- [10] D. Stephenson, “Distribution of water in deep gold mines,” 2006.
- [11] A. Schutte, M. Kleingeld, and L. Van Der Zee, “An integrated energy efficiency strategy for deep mine ventilation and refrigeration,” in *2014 International Conference on the Industrial and Commercial Use of Energy (ICUE)*, 2014. DOI: doi:10.1109/ICUE.2014.6904168.
- [12] G. Fields, “Gold fields limited south deep gold mine,” South Africa, Tech. Rep., 2020, Accessed on: Apr 05, 2021. [Online]. Available: <https://www.goldfields.com/reports/annual-report-2018/mrr/sa-south-deep-gold-mine.php>.
- [13] Energy Insight. “Case studies.” Accessed on: Jul 28, 2021. (2021), [Online]. Available: <http://www.energyinsight.co.za/case-studies/>.

- [14] D. Landsberg, "Survey of international experience in advanced metering infrastructure and its implementation," *Global Practice Group South Asia Region (SAR)*, 2018.
- [15] S. Touzani, S. Fernandes, C. Taylor, and J. Granderson, "Application of automated measurement and verification to utility energy efficiency program data," *Energy and Buildings*, vol. 142, pp. 191–199, 2017.
- [16] R. Silvola, J. Harkonen, O. Vilppola, H. Kropsu-Vehkaperä, and H. Haapasalo, "Data quality assessment and improvement," *International Journal of Business Information Systems*, vol. 1, pp. 62–81, 2016.
- [17] J. Vaughan. "Definition: Data quality." Accessed on: Jul 29, 2021. (2021), [Online]. Available: <https://searchdatamanagement.techtarget.com/definition/data-quality>.
- [18] Lotame. "Why is data quality important?" Accessed on: Jul 29, 2021. (2021), [Online]. Available: <https://www.lotame.com/why-is-data-quality-important/>.
- [19] U. N. G. Marketplace, *Data Quality Assessment Handbook*. 2021.
- [20] S. Sheriff. "Extracting actionable insights from data: What you need to know." Accessed on: Jul 31, 2021. (2020), [Online]. Available: <https://acuvate.com/blog/extracting-actionable-insights-from-data-what-you-need-to-know/>.
- [21] S. Touzani, S. Fernandes, J. Granderson, C. Custodio, M. Sohn, and D. Jump, "Accuracy of automated measurement and verification (m&v) techniques for energy savings in commercial buildings," *Applied Energy*, vol. 173, pp. 296–308, 2016.
- [22] J. Mathieu, P. Price, S. Kiliccote, and M. Piette, "Quantifying changes in building electricity use, with application to demand response," *IEEE Transactions on Smart Grid*, vol. 2, 2011.
- [23] *Schedule of standard prices for eskom tariffs*, Accessed on: Sep 09, 2021., Eskom, South-Africa, (2021). [Online]. Available: https://www.eskom.co.za/CustomerCare/TariffsAndCharges/Pages/Tariffs_And_Charges.aspx.
- [24] C. Pelser, "Improved energy budgeting process for mines using measurement and verification principles," Ph.D. dissertation, North West University, 2020.
- [25] M. Brown, C. Barrington-Leigh, and Z. Brown, "Kernel regression for real-time building energy analysis," *Journal of Building Performance Simulations*, vol. 5, pp. 263–276, 2012.
- [26] H. Doa and K. Cetin, "Evaluation of the causes and impact of outliers on residential building energy use prediction using inverse modeling," *Building and Environment*, vol. 138, pp. 194–206, 2018.
- [27] T. Amundson, J. Granderson, E. Crowe, S. Earni, and S. Fernandes, "Ipmvp application guide on non-routine events & adjustments," *International Performance Measurements and Verification Protocol*, vol. 104000, pp. 1–105, 2020.
- [28] Mining Technology. "Mining data: Advanced capture methods help miners do more with less." Accessed on: June 17, 2021, [Online]. Available: <https://www.mining-technology.com/features/feature-mining-data-advanced-capture-methods-miners-more-with-less/>.

- [29] Industy Software Solutions and Support. “Enterprise-wide power management gives sibanye’s mines the control they need.” Accessed on: Jul 01, 2021, [Online]. Available: <https://www.wonderware.co.za/customer-stories/enterprise-wide-power-management-gives-sibanyes-mines-the-control-they-need/>.
- [30] D. Olivotti and D. Eilers, “Understanding anomalies: Visualizing sensor data for condition monitoring of manufacturing machine,” *Leibniz Universität Hannover, Institute for Information Systems*, 2018.
- [31] R. Chuahan, “Recent advances in scada alarm systems,” *International Journal of Smart Home*, vol. 4, no. 10, 2010.
- [32] J. Haberl, D. Claridge, and C. Culp, “Ashrae’s guideline 14-2002 for measurement of energy and demand savings: How to determine what was really saved by the retrofit,” Jan. 2005.
- [33] Britannica, The Editors of Encyclopaedia. “Normal distribution.” Accessed on: January 03, 2022, [Online]. Available: <https://www.britannica.com/topic/normal-distribution>.
- [34] M. Mahbobi, *Introductory Business Statistics With Interactive Spreadsheets*, 1st ed. BC-campus Open Education, 2020.
- [35] A. Hayes. “What is varaince?” Accessed on: June 19, 2021, [Online]. Available: <https://www.investopedia.com/terms/v/variance.asp>.
- [36] O. Grigg, V. Farewell, and D. Spiegelhalter, “The use of risk-adjusted cusum and rsprt charts for monitoring in medical contexts,” *Statistical Methods in Medical Research*, vol. 12, pp. 147–170, Apr. 2003.
- [37] B. McNeese. “Keeping the process on target: Cusum charts.” Accessed on: June 22, 2021, [Online]. Available: <https://www.spcforexcel.com/knowledge/variable-control-charts/keeping-process-target-cusum-charts>.
- [38] National Institute of Standard and Technology, *Engineering Statistics*. U.S. Department of Commerce, 2013, Accessed on: Jun 28, 2021. Available: <https://www.itl.nist.gov/div898/handbook/index.htm>.
- [39] M. Benoit and P. Petitgas, “Detection of changes in time-series of indicators using cusum control charts,” *Aquatic living resources (EDP Sciences)*, vol. 22, pp. 187–192, Apr. 2009.
- [40] S. Aminikhanghahi and D. J. Cook, “A survey of methods for time series change point detection,” *Knowledge and Information Systems volume*, vol. 51, pp. 339–367, 2017. DOI: doi:10.1007/s10115-016-0987-z.
- [41] R. Killick and I. Eckley, “Changepoint: An r package for changepoint analysis,” *Journal of Statistical Software*, vol. 58, no. 9, 2014.
- [42] A. Engelward and A. Mater. “Math for teaching advanced algebra and trigonometry.” Accessed on: Jul 16, 2021. (2021), [Online]. Available: <http://people.math.harvard.edu/~engelwar/MathS305/>.

- [43] S. Chong and A. Treisman, “Representation of statistical properties,” *Vision Research*, vol. 43, no. 4, pp. 393–404, 2003. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0042698902005965>.
- [44] S. Otto. “Comparison of change point detection methods.” Accessed on: Jul 04, 2021. (2019), [Online]. Available: <https://www.marinedatascience.co/blog/2019/09/28/comparison-of-change-point-detection-methods/>.
- [45] M. Basseville and I. Nikiforov, *Detection of Abrupt Change Theory and Application*. Apr. 1993, vol. 15, ISBN: 0-13-126780-9.
- [46] B. Namono, A. Starr, C. Emmanouilidis, and C. Ruiz-Cárcel, “Online change detection techniques in time series: An overview,” pp. 1–10, Jun. 2019. DOI: 10.1109/ICPHM.2019.8819394.
- [47] D. Rindskopf and M. Shiyko, “Measures of dispersion, skewness and kurtosis,” *International Encyclopedia of Education*, vol. 3, 2010.
- [48] T. Grundy, R. Killick, and G. Mihaylov, “High-dimensional changepoint detection via a geometrically inspired mapping,” *Stat Comput*, vol. 30, pp. 1155–1166, 2020.
- [49] R. Wilcox, “Applying contemporary statistical techniques,” in Burlington: Academic Press, 2003, pp. 55–91, ISBN: 978-0-12-751541-0.
- [50] J. Brooks-Bartlett. “Probability concepts explained: Maximum likelihood estimation.” Accessed on: Jul 22, 2021. (2018), [Online]. Available: <https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1>.
- [51] R. Riffenburgh, *Risks, Odds, and ROC Curves*, 3rd ed. Statistics in Medicine, 2012.
- [52] A. Edwards and L. Cavalli-Sforza, “A method for cluster analysis,” *Biometrics*, vol. 21, pp. 362–375, 1965.
- [53] R. Killick, P. Fearnhead, and I. Eckley, “Optimal detection of changepoints with a linear computational cost,” *Journal of the American Statistical Association*, vol. 107, pp. 1590–1598, 2012.
- [54] C. Truonga, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods,” *Signal Processing*, 2020.
- [55] M. Ray, M. Mongiardini, C. Plaxico, and M. Anghileri, *Procedures for Verification and Validation of Computer Simulations Used for Roadside Safety Applications*. National Cooperative Highway Research Program, 2011.
- [56] S. Shastri, V. Mansotra, and A. Sharma, “Sorting algorithms and their run-time analysis with c#,” *International Journal of Computer Science and Information Technology & Security*, vol. 5, p. 5, 2015.
- [57] S. Balakirsky and T. Kramer, “Comparing algorithms: Rules of thumb and an example,” *National Institute of Standards and Technology*,
- [58] P. Black. “Big-o notation.” Accessed on: Aug 25, 2021. (2019), [Online]. Available: <https://www.nist.gov/dads/HTML/bigOnotation.html>.

- [59] A. Coluccia and A. Fascista, “An alternative procedure to cumulative sum for cyber-physical attack detection,” *Internet Technology Letters*, vol. 1, 2018.
- [60] K. Pedersen, J. Emblemstvag, R. Bailey, J. Allen, and F. Mistree, “Validating design methods and research: The validation square,” 2000. [Online]. Available: <https://cecas.clemson.edu/cedar/wp-content/uploads/2016/07/2-Mistree2000.pdf>.
- [61] S. Consulting, *Potential analytics for non-routine adjustments*. Bonneville Power Administration (BPA), Bellevue, WA, 2017.
- [62] L. blaettler, “Identifying essential data to evaluate and monitor energy performance,” M.S. thesis, North West University, 2020.
- [63] S. Touzani, B. Ravache, E. Crowe, and J. Granderson, “Statistical change detection of building energy consumption: Applications to savings estimation,” *Journal of Chemical Thermodynamics*, vol. 185, pp. 123–136, 2019.
- [64] W. Kenton. “What is probability density function (pdf)?” Accessed on: Jul 22, 2021. (2020), [Online]. Available: <https://www.investopedia.com/terms/p/pdf.asp>.

Appendix A Statistical concepts

The well-known probability density function for normally distributed data is Equation A.1 [64] — it states what the probability is of observing data point x_i , given a known mean and standard deviation. Therefore, the mean and standard deviation of the data set is calculated and are constants in Equation A.1.

$$P(x_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(x_i - \mu)^2}{2\sigma^2} \right] \quad (\text{A.1})$$

The concept is illustrated in Figure A.1, which was created with arbitrary data that ranges between -3 and 5 , with $\sigma = 1$ and $\mu = 1.5$. Equation A.1 and Figure A.1 show the likelihood of what value a random sample from the series x will hold. The likelihood that a random sample will be more than three standard deviations from the mean is low — it is more likely that this sample will be within one standard deviation from the mean.

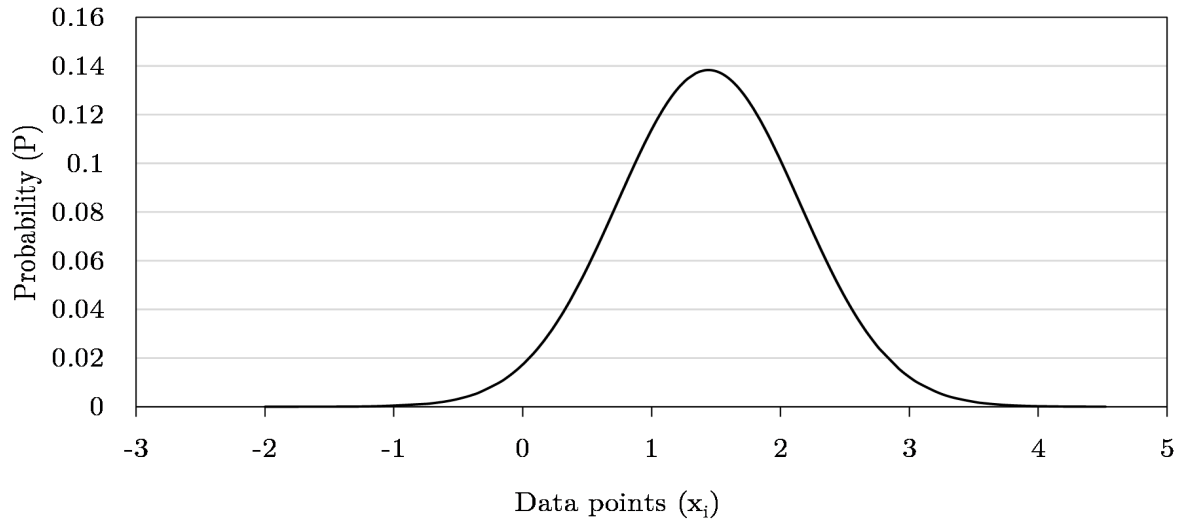


Figure A.1: Probability density curve of arbitrary time series data with $\mu = 1.5$ and standard deviation $\sigma \approx 1$.

Appendix B Further insight from literature and mining

Figure B.1 shows a typical SCADA. Such a program provides live feedback from measurement points such as power sensors in the mine. It is also possible to set up alarms that sound warnings when systems show a power failure or become out of bounds [29].

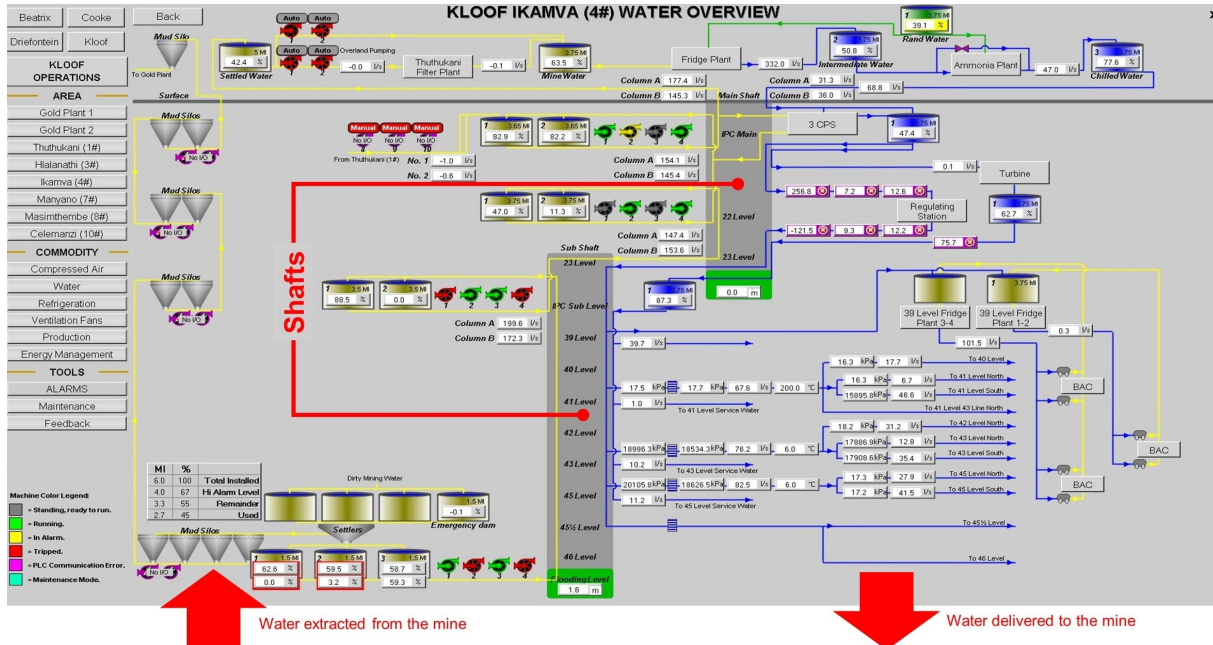


Figure B.1: Sibanye Gold Klooof SCADA of pumps and water lines [29].

Appendix C Further insight on the case study data

The data preparation results for standard and peak time of use (TOU) periods are discussed in this section. Abbreviations S_WKD_STD and S_WKD_PEAK are summer weekday standard and peak respectively. First, results of the categorisation according to Eskom time of use periods (TOU) are shown. Figure C.1 is the scaled power output of the S_WKD_STD period.

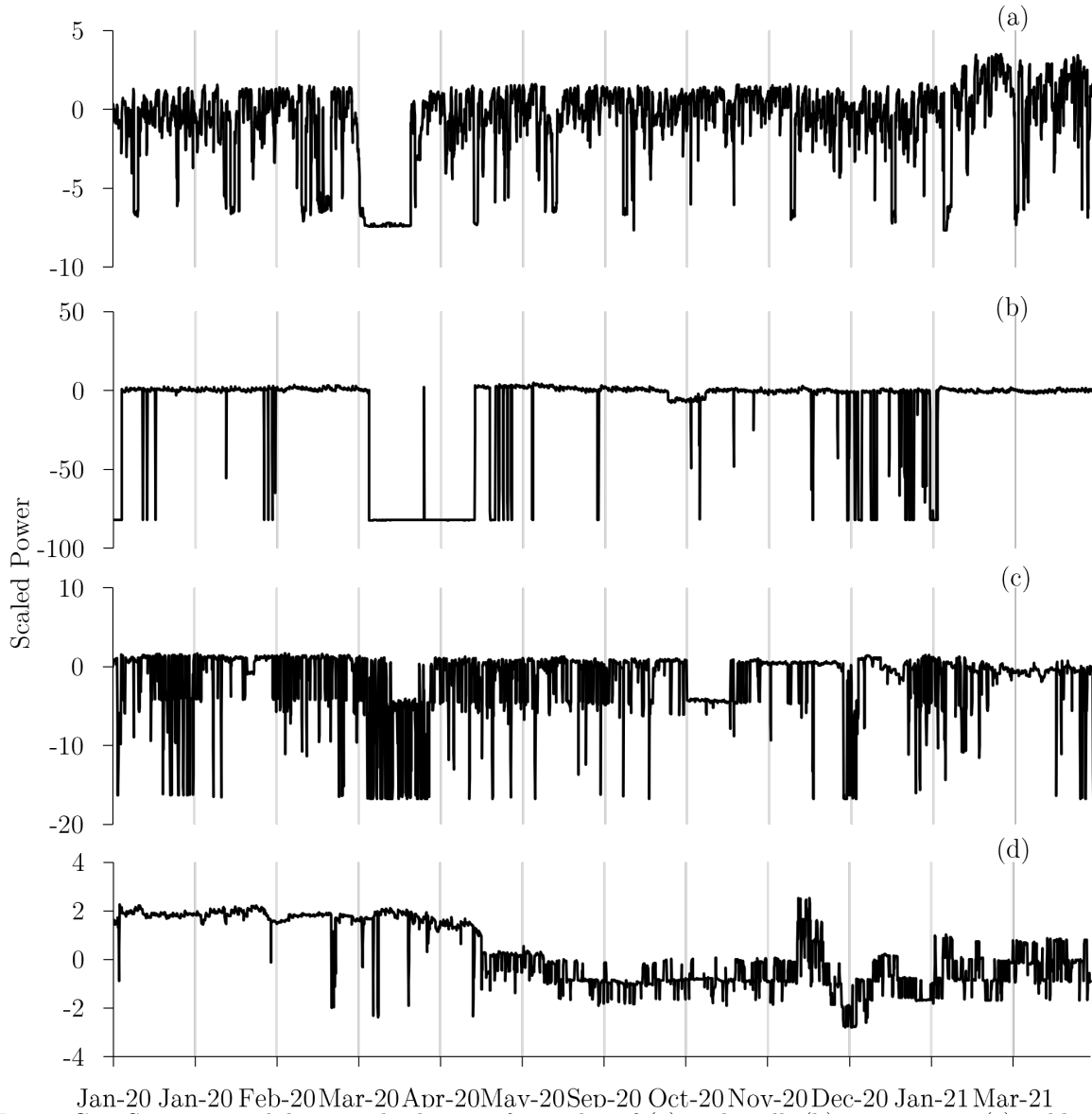


Figure C.1: Summer weekday standard time of use plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.

Figure C.2 shows the S_WKD_PEAK scaled power output for the case studies.

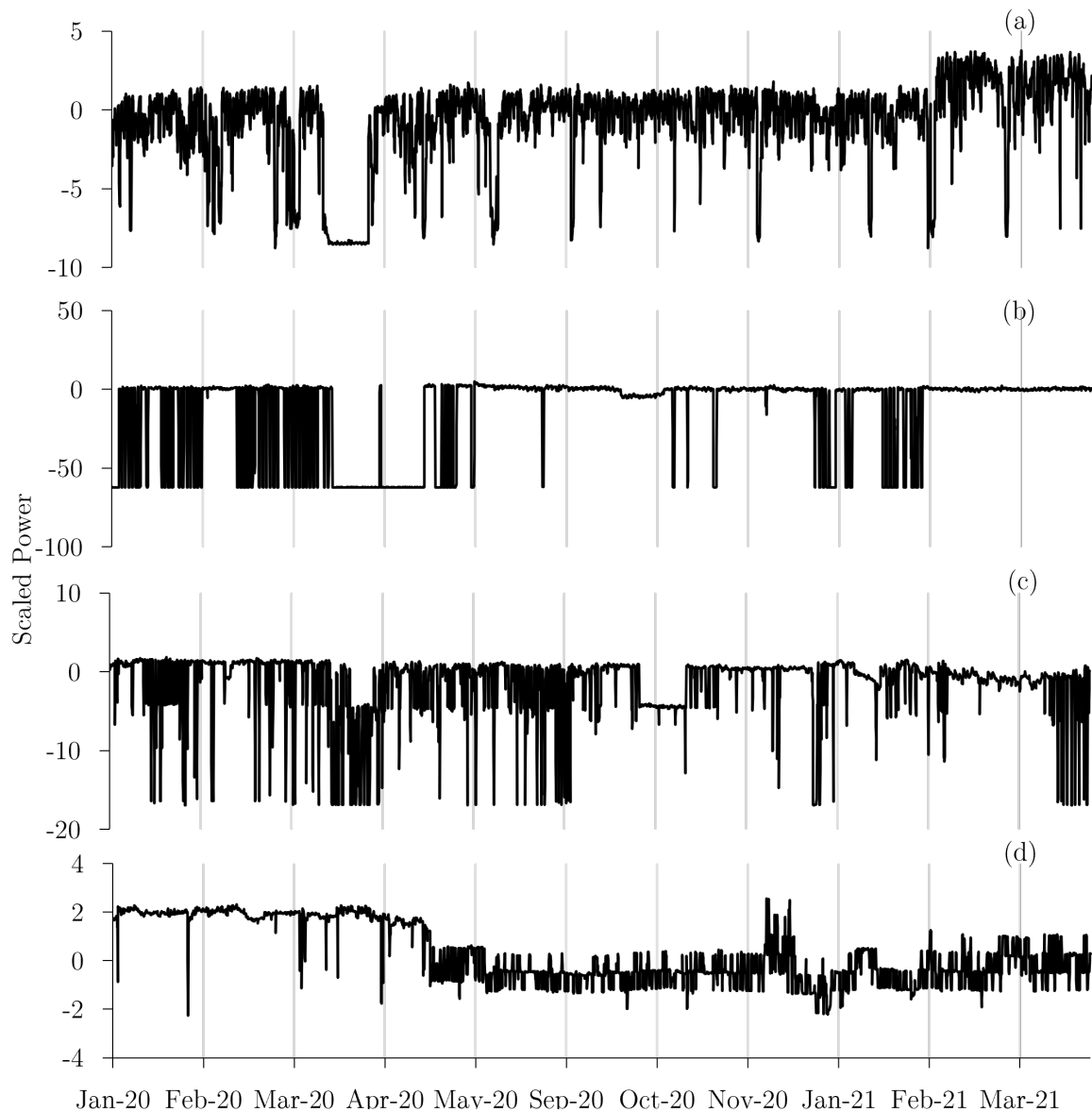


Figure C.2: Summer weekday peak time of use plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.

Figure C.3 is a standard deviation plot. The data sets are the baseline summer weekday standard (S_WKD_STD) TOU power outputs of the four the case studies. These data sets were scaled by applying the median absolute deviation equation. Thereafter, the Z-score (Equation 1.2) for each data point was calculated to create the plot below. This clearly shows which data points are more than two standard deviations away from the mean. Such data points were considered outliers and removed from the baseline data sets.

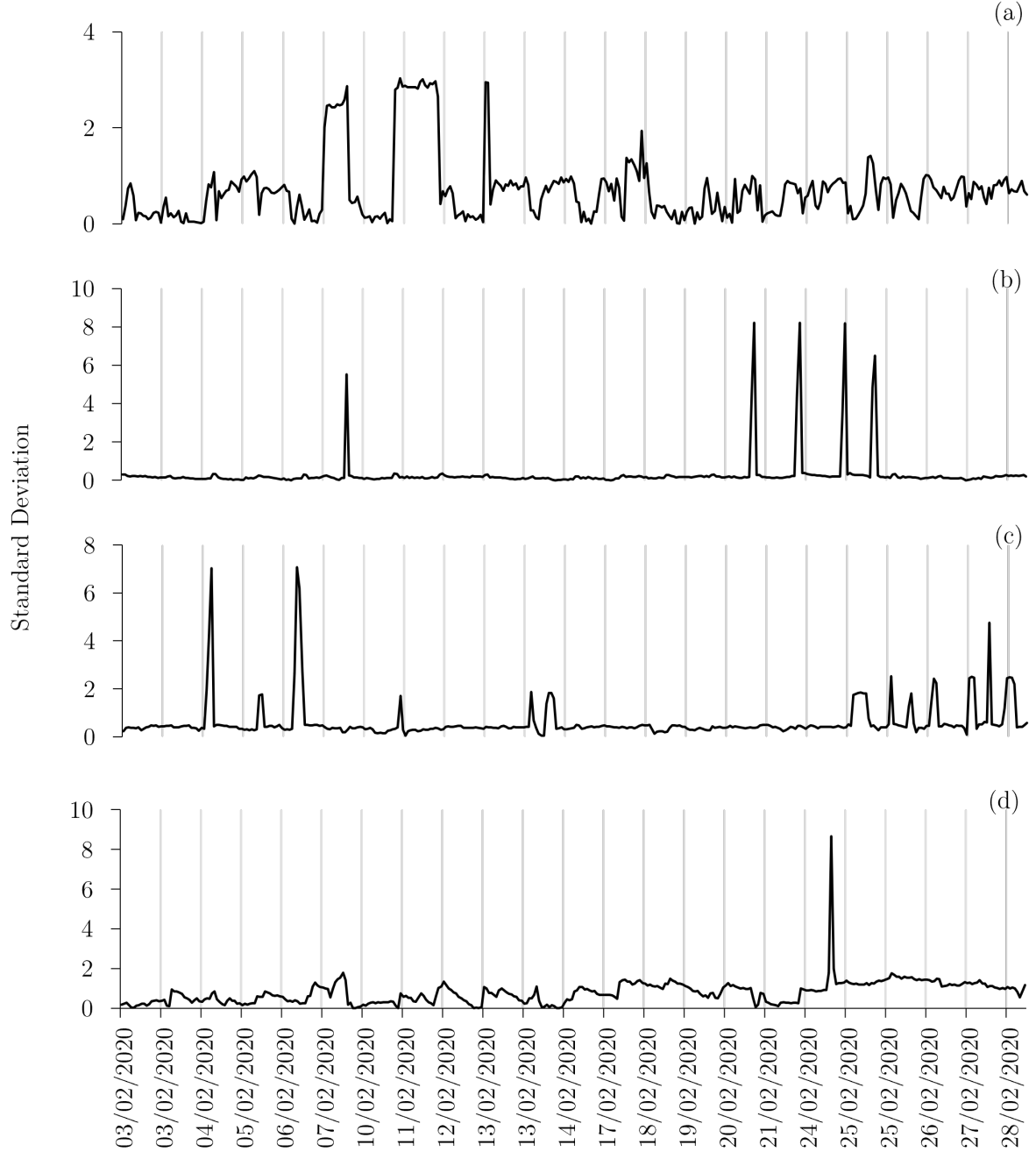


Figure C.3: S.WKD.STD standard deviation plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.

Similar to the previous figure, Figure C.4 show the S_WKD_PEAK TOU data. These plots show the deviation of each value in the baseline data. The final baseline data preparation step is to remove outliers.

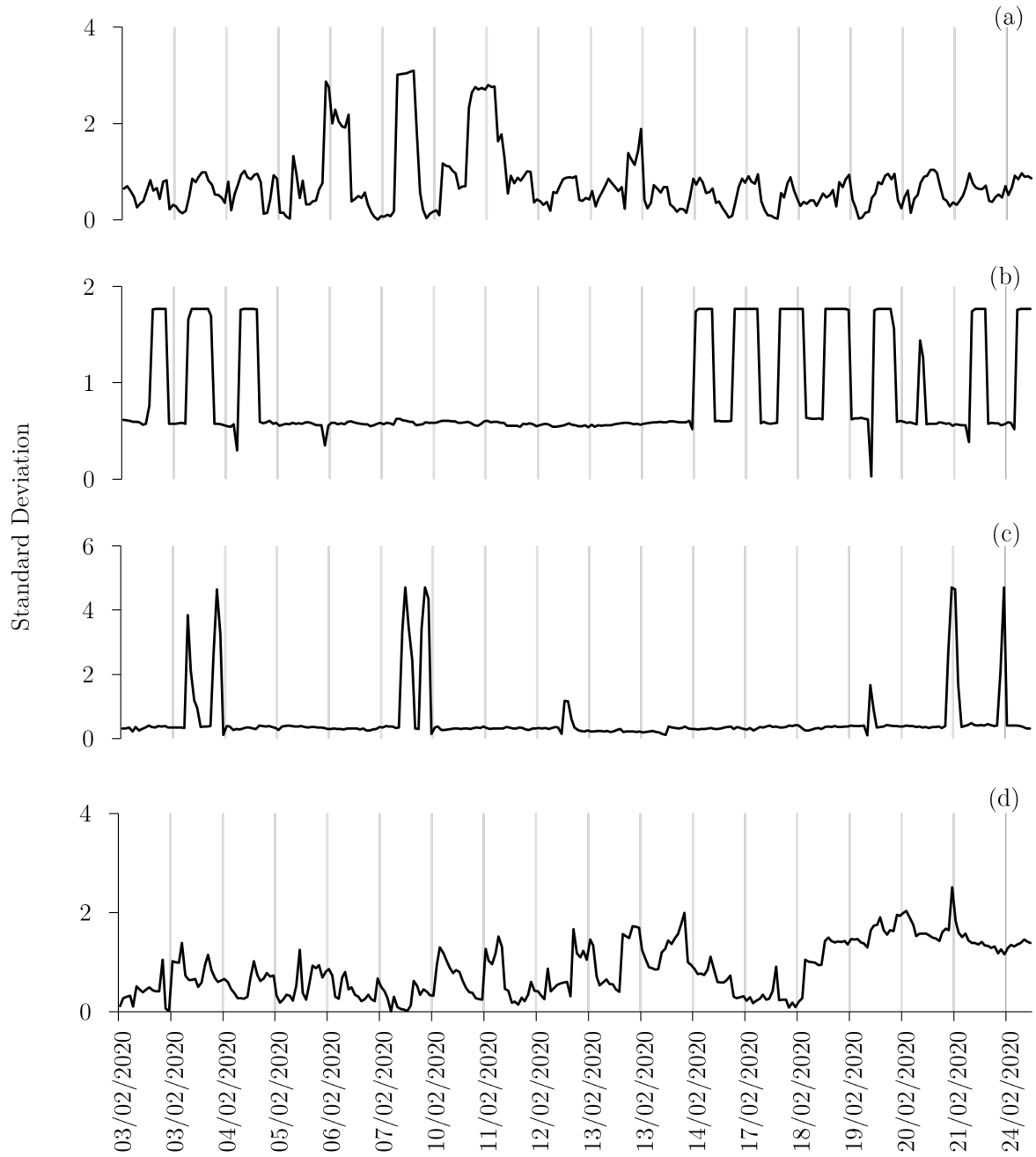


Figure C.4: S_WKD_PEAK standard deviation plot of (a) rock mill, (b) compressor, (c) gold plant and (d) fridge plant.

Appendix D Change point detection results

CUSUM results

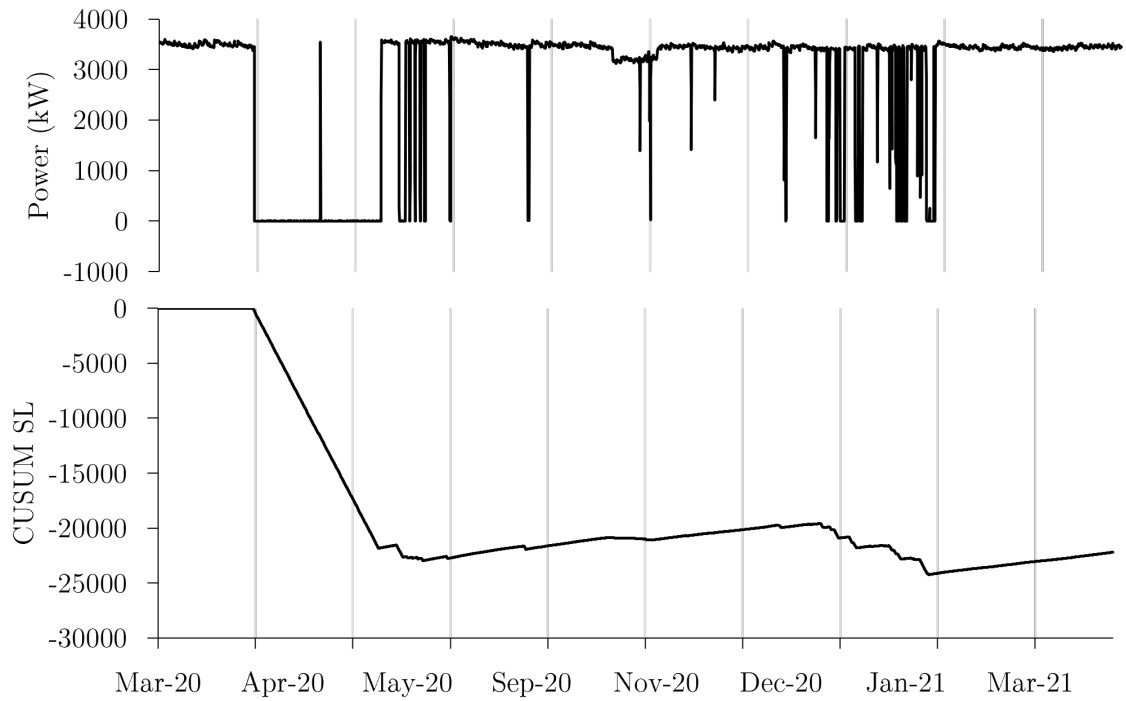


Figure D.1: Top: S_WKD_OFF power output of a compressor in 2020. Bottom: S_L CUSUM result.

The next section shows the BinSeg and PELT parameter selection. Specifically, for the summer weekday standard and peak time of use periods. The purpose of this parameter is to set a realistic number of changepoints according to the data set. Figure D.2 shows the elbow plots of the S_WKD_STD power data. Plot (a) is of the rock mill and shows that the number of possible change points varies from 26 to 3 with a penalty range of 0 to 500. The apex of this curve is at coordinate (180,7), this means that seven change points can be expected for this data set. Figure D.2 (b) is the elbow plot for the mining compressor. Inspection of the raw data shows little variation, therefore the more strict penalty in Figure D.2 (b) of 44 was chosen, to limit the number of plausible change points to three.

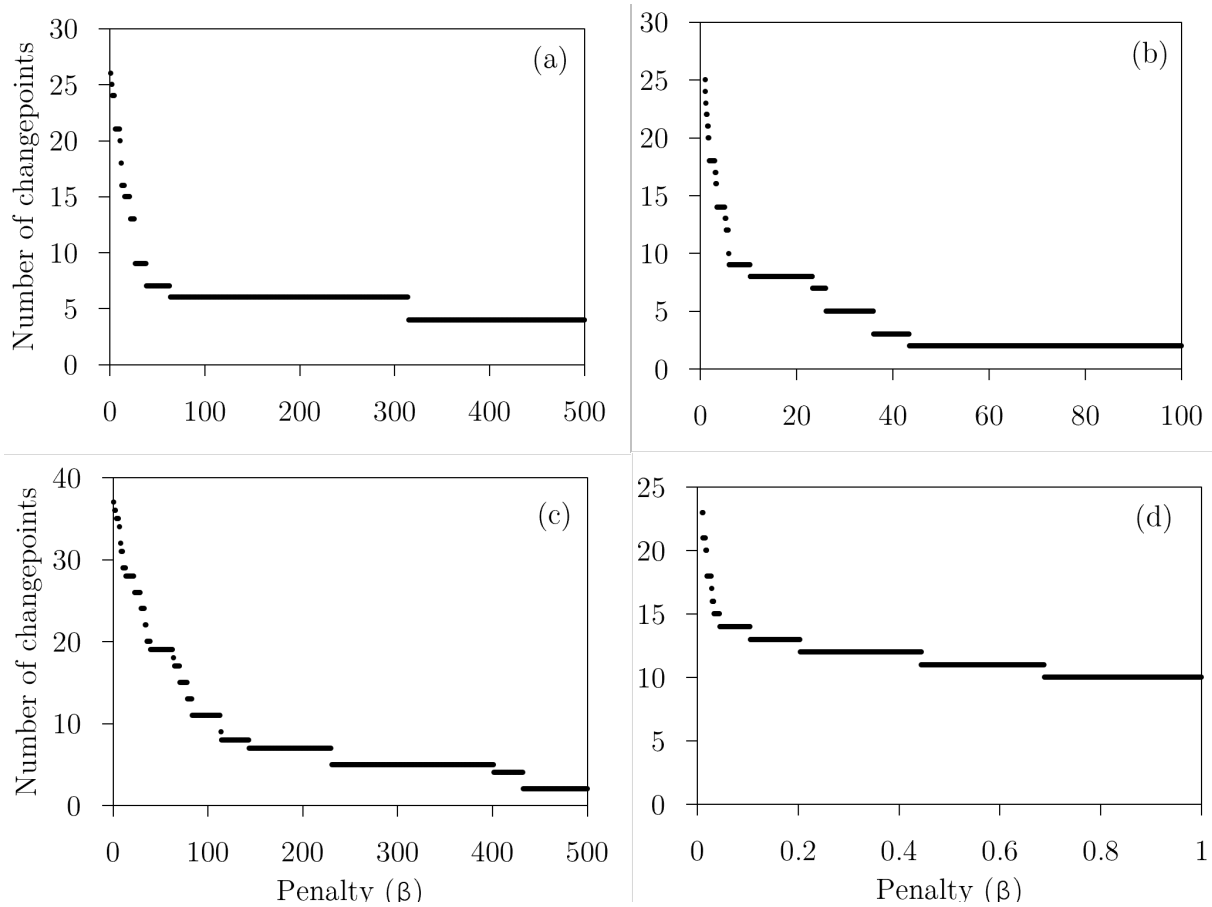


Figure D.2: Elbow plots showing the number of changepoints with varying penalties in the PELT method for the March 2020 S_WKD_STD power data of: (a) rock mill, (b) mining compressor, (c) gold plant and (d) fridge plant.

In plot (c) of Figure D.2, the elbow curve shows that 15 is a plausible number of changepoints with a penalty of 100. For the fridge plant data in plot (d), the penalty range is quite small. This elbow plot suggests that there are many changepoints in the data set. The apex of the curve is at (0.03, 13) and such a penalty suggests that 14 changepoints can be expected.

The summer weekday peak time of use elbow plots are shown in Figure D.3. The apex in plot (a) is not well defined, it shows a more rounded shape. Upon inspecting the data, it was decided that a penalty of 20 is reasonable. Plot (b) in Figure D.3 shows that no less than 3

change points are expected for this data set. The penalty range is large, from 0 to 20000. It converges to 3 at $\sigma = 6000$, therefore a more realistic change point could be closer to (5300, 5).

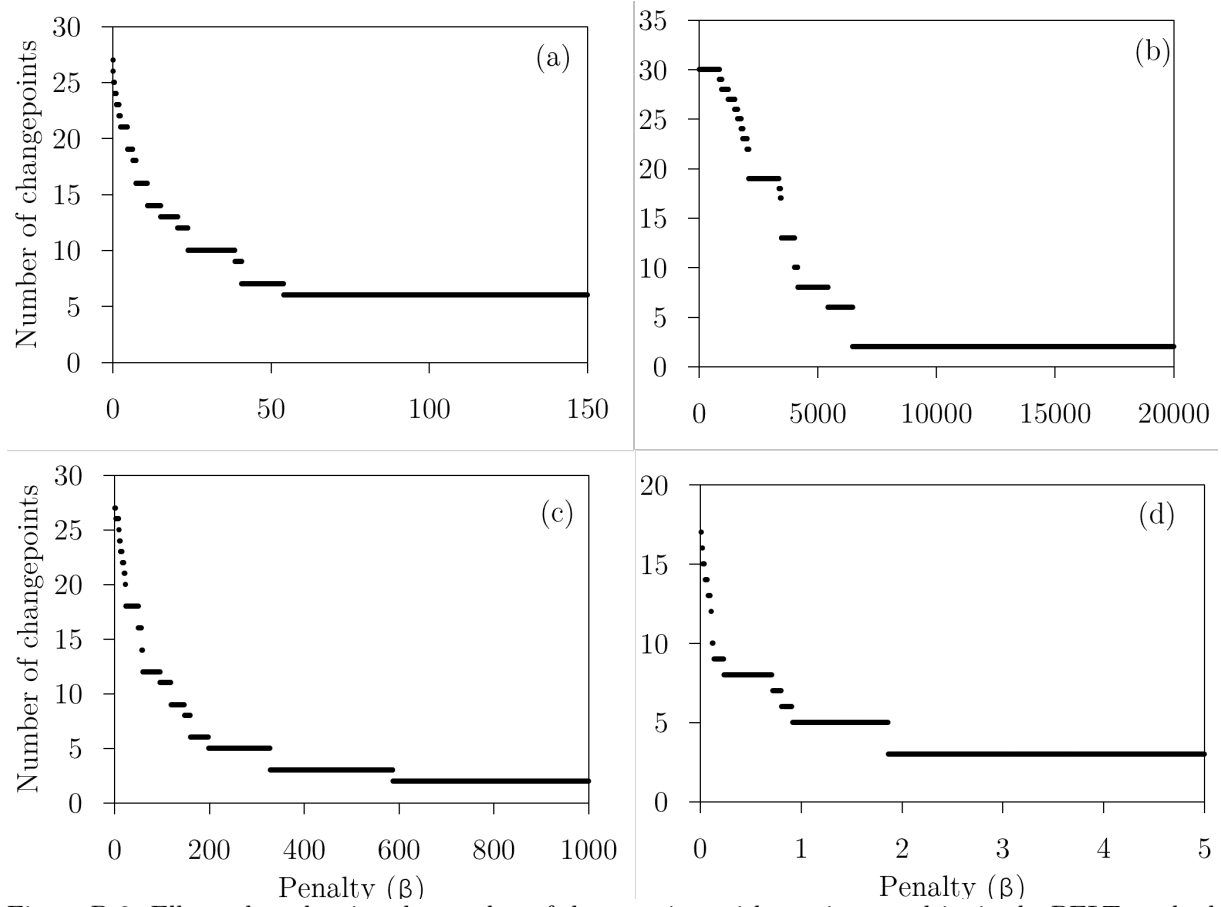


Figure D.3: Elbow plots showing the number of change points with varying penalties in the PELT method for the March 2020 S-WKD-PEAK power data of: (a) rock mill, (b) mining compressor, (c) gold plant and (d) fridge plant.

Figure D.3 (c) shows an apex at (50, 7). Therefore, a realistic number of change points for this data set is 7. In Figure D.3 (d), there are two “elbows” and it suggests that both 5 and 8 number of change points can be expected. The parameters in this discussion were applied to the BinSeg and PELT algorithms. An analysis of the standard and peak TOU change point detection results are presented next.

In Figure D.4 (a), the assessment period data of the rock mill is shown. Plot (b) shows the scaled version of this data as an area graph. The changepoint detection results are plotted as lines that show the mean of each identified segment.

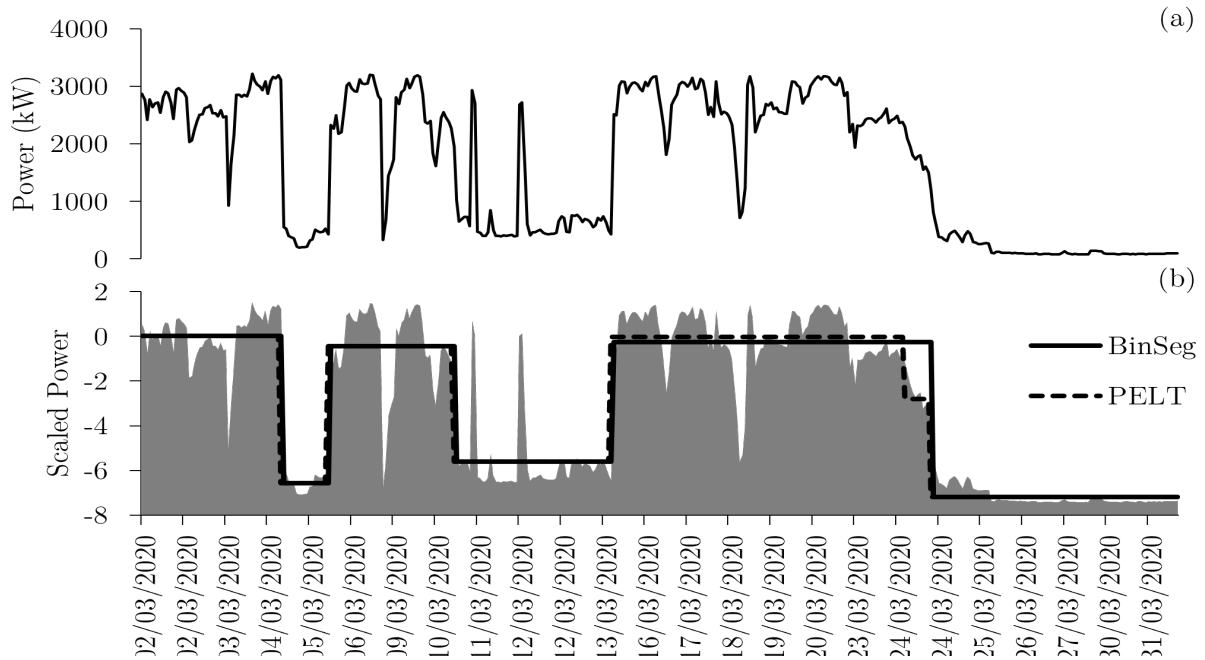


Figure D.4: (a) S_WKD_STD power output in March 2020 of a rock mill. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

Figure D.5 shows the change point detection results of the mining compressor. It is clear that both the BinSeg and PELT algorithms successfully identified decreased power output that started on the commencement of the national lockdown.

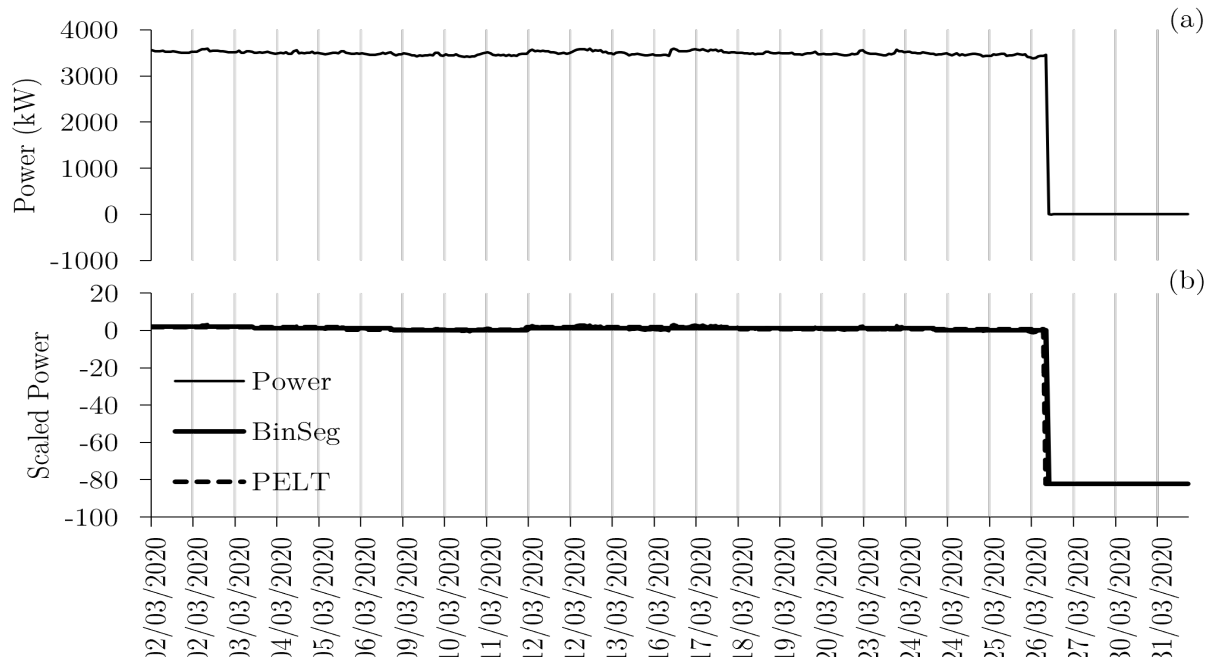


Figure D.5: (a) S_WKD_STD power output in March 2020 of the mining compressor. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

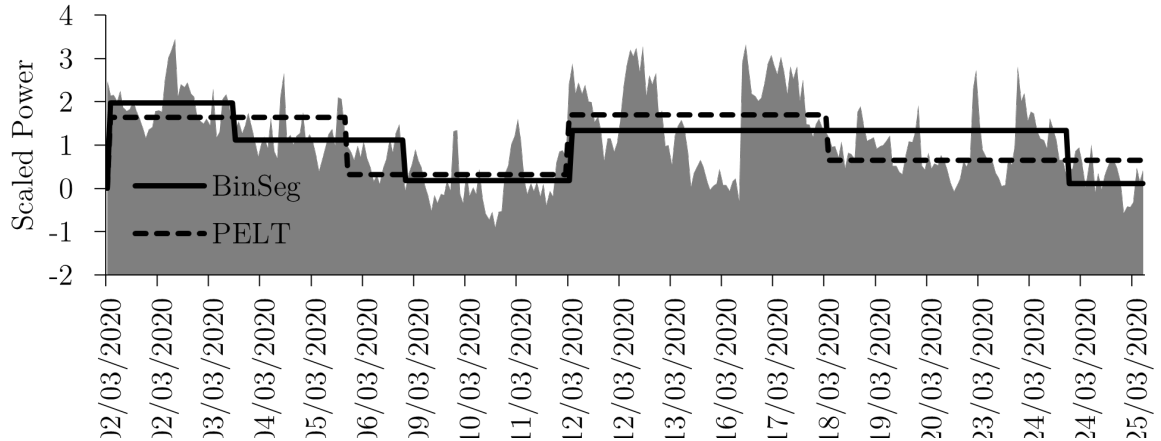


Figure D.6: Area graph of the scaled compressor data with the change point detection results overlaid.

A closer look at the compressor data is shown in Figure D.6. This is the same result and data as that for the previous figure, but it excludes the national lockdown days. When observing the case study data, one might argue that the algorithms did not identify enough changepoints. This statement is supported by the three distinct regions of increase power output from 11 to 18 March 2020. One reason for this may be the fact that the standard deviation of the test data was not exactly 1, but rather 27.

The gold plant S_WKD_STD period changepoint detection results are shown in Figure D.7. Even though this system does not show a rapid decrease, as seen with the rock mill and mining compressor, the BinSeg and PELT algorithms successfully identified the reduced power due to the national lockdown.

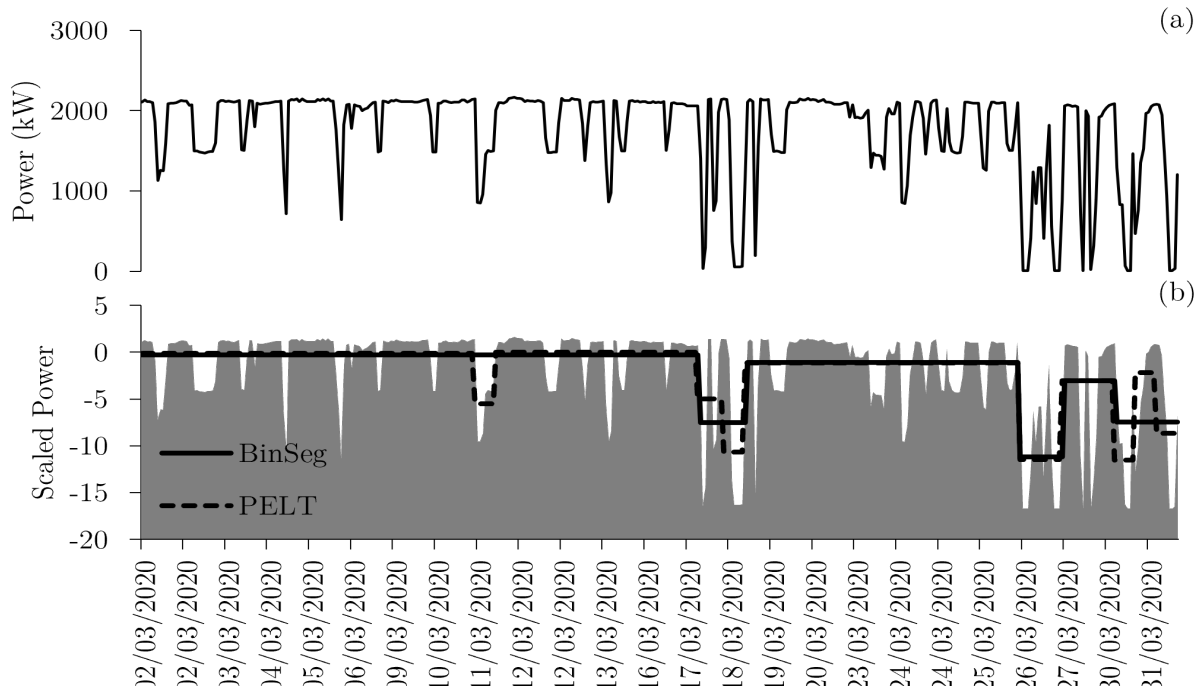


Figure D.7: (a) S.WKD_STD power output in March 2020 of the gold plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

The changepoint detection results of the fridge plant S_WKD_STD power data is plotted in Figure D.8. There is a distinct period of decreased power and increased variability from 13 to 16 March 2020. Both the BinSeg and PELT algorithms classified this as a changepoint.

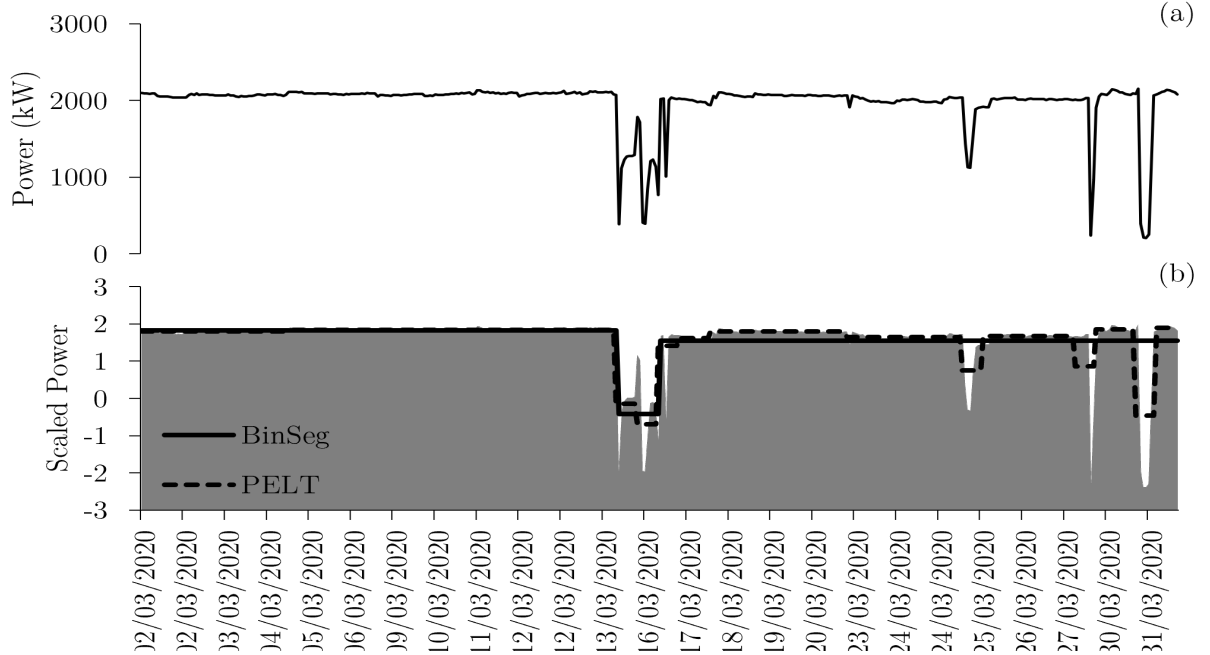


Figure D.8: (a) S_WKD_STD power output in March 2020 of the fridge plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

In Figure D.9, the changepoint detection results are shown for the S_WKD_PEAK data of the rock mill.

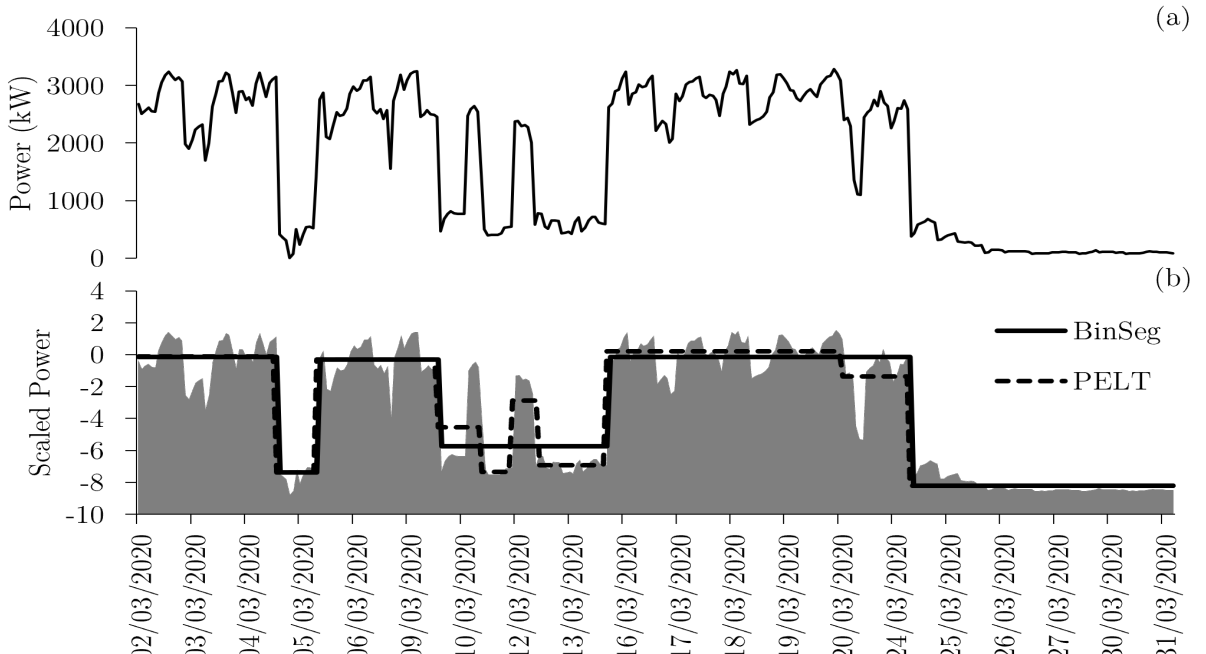


Figure D.9: (a) S_WKD_PEAK power output in March 2020 of a rock mill. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

The S_WKD_PEAK power data of the mining compressor is plotted in Figure D.10 (a). The

general profile of this power data shows higher variability than that of the off peak and standard TOU periods (Figures D.5 and 3.16). Further analysis is required to determine if this profile is normal. It may be necessary to revisit the minimum segment length setting to establish whether less than four hours is more suited to a mining compressor.

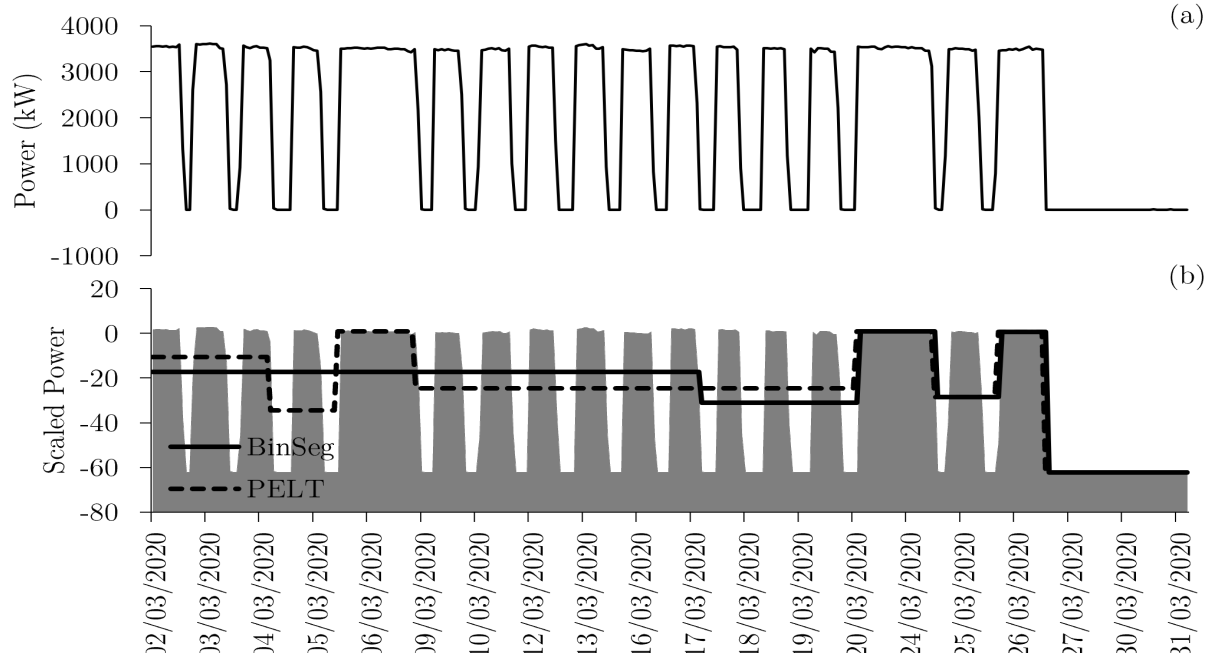


Figure D.10: (a) S_WKD_PEAK power output in March 2020 of the mining compressor. (b) The change-point detection results overlaid on the scaled version of plot (a) as an area graph.

The gold plant S_WKD_STD period change-point detection results are shown in Figure D.11. From the national lockdown commencement, the power output was reduced more in the peak Eskom TOU periods than in the standard TOU periods. Both the BinSeg and PELT algorithms identified this as a change-point.

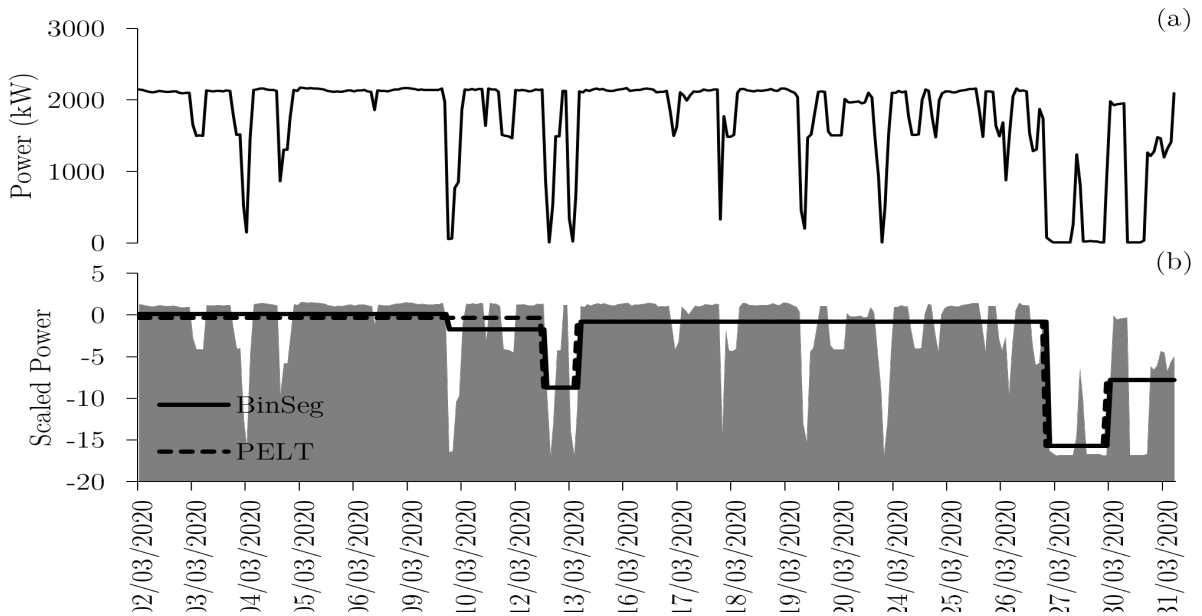


Figure D.11: (a) S_WKD_PEAK power output in March 2020 of the gold plant. (b) The change-point detection results overlaid on the scaled version of plot (a) as an area graph.

Figure D.12 (a) show the peak TOU power data for the fridge plant. The BinSeg algorithm did not identify any changepoints, whereas the PELT algorithm identified two.

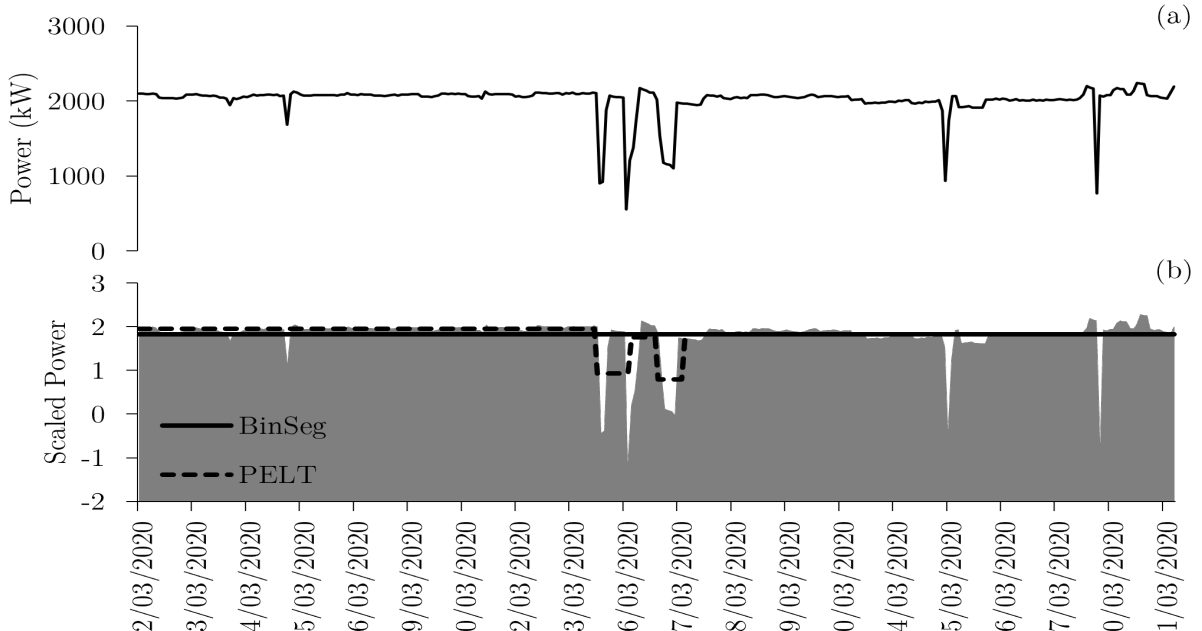


Figure D.12: (a) S_WKD_PEAK power output in March 2020 of the fridge plant. (b) The changepoint detection results overlaid on the scaled version of plot (a) as an area graph.

The average runtime is summarised in Table D.1. This was created by measuring the start and end timestamp as an algorithm was initiated and ran to completion. Each test was conducted ten times and the average runtime was calculated in milliseconds.

Table D.1: Average runtime of the BinSeg and PELT algorithms.

BinSeg	Rock Mill (ms)	Mining Compressor (ms)	Gold Plant (ms)	Fridge Plant (ms)
S_WKD_OFF	19.35	0.60	0.60	0.50
S_WKD_STD	1.50	0.90	0.90	0.70
S_WKD_PEAK	19.75	0.70	0.50	0.40
PELT	Rock Mill (ms)	Mining Compressor (ms)	Gold Plant (ms)	Fridge Plant (ms)
S_WKD_OFF	22.54	0.70	0.60	0.60
S_WKD_STD	0.90	1.09	0.90	1.00
S_WKD_PEAK	0.90	0.70	0.80	0.70