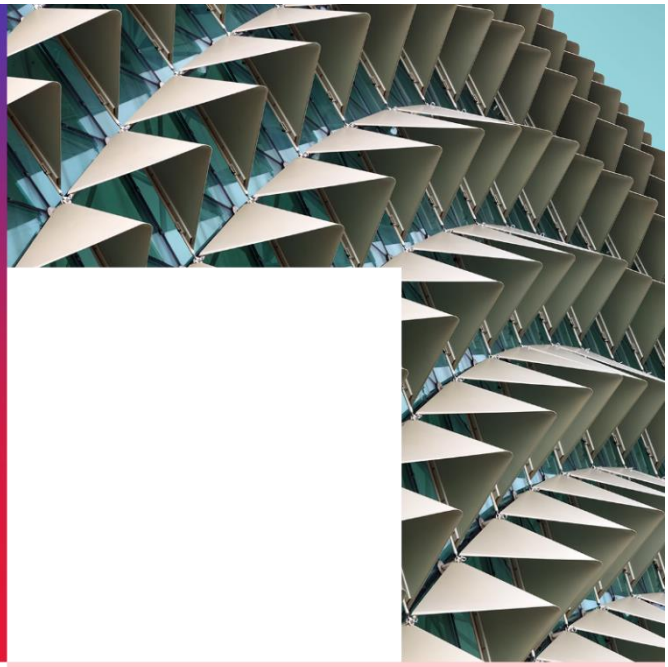


Phase Identification Report

SMITN

31/08/2023

Prepared by

Iro Psarra, Alex Gardner, Nadim Al-Hariri, Andrew Urquhart

Amendment History

Date	Issue	Status
19/12/2022	0.1	Initial version
31/08/2023	0.2	Updated following review
15/02/2023	1.0	Final version (track changes removed)
11/05/2023	2.0	Update to authors
31/08/2023	2.1	Update to clarify clock offsets

Contents

1 Introduction

1.1 Background

Smart Meter (SM) data opens opportunities for Distribution Network Operators (DNOs) to improve their database records, develop a low-voltage (LV) model which will be useful for network planning, fault detection and phase balancing.

Any LV network that supplies more than a few customers is composed of three live phases and a neutral phase. A normal domestic customer is connected between one of the live phases and the neutral phase. Phase unbalance widely exists in the UK's low voltage distribution networks. The unbalances not only lead to insufficient use of LV network assets but also cause increased energy losses.

Having reliable data is very important to understand the phase unbalance which affects losses and the power quality supplied to any three-phase loads which may be present in the distribution network. Phase unbalance will be very useful where new LCT connections requested. The SMITN project is an innovation project that will investigate how the use of SM data can be used to support network operations. In the first use case, "Phase Identification", we will assess how SM's voltage time-series data can be used to effectively identify the phase to which each customer is connected in the network. The algorithms used were established by holding a collaborative workshop with other DNOs to discuss their previous and planned work in this area.

Our analysis is focused on 46 secondary substations (SS) in the Milton Keynes area with installed GridKey monitoring devices.

In some areas the phase connections for single-phase customers are already annotated on the network diagrams but this data is incomplete and the phases of most customers are unknown.

Phases identified by the SMITN algorithms will therefore be validated against results of a detailed survey of customer phase connections using the HAYSYS Phase Finder Unit. This process begins by setting a phase reference at each secondary substation. Each customer connection is then visited, allowing polyphase meters to be identified and single-phase connections to be identified and assigned as either L1, L2 or L3 relative to the phases at the substation busbar.

The phase survey used here for validation has been found to be highly accurate, but the survey method is necessarily time-consuming as each customer connections needs to be visited. The SMITN project therefore aims to demonstrate the viability of techniques to allow phase connections to be determined from smart meter data, using the phase survey in the test network area for validation.

1.2 Document Purpose

This document presents a high-level description of the algorithms selected for the Phase Identification use case and will present and discuss the results, using validation data from HAYSYS.

The purpose of this analysis is to examine the:

- sensitivity to duration of voltage time series (1 month, 3 month 5-6 months),
- sensitivity to the resolution of the input data (1 minute vs half hourly),
- differences reflecting the availability of local monitoring data,
- differences in the clustering algorithms used,
- differences in the feature being used for clustering,
- accuracy of Electric Office (EO) phase information and network data,

- sensitivity to smart meter coverage.

This will provide useful information to enable National Grid Electricity Distribution (NGED) and other DNOs to implement Business As Usual (BAU) processes to validate existing phase data and backfilling missing data.

1.3 Overview

Section 2 of this document presents an overview of the SMITN selected algorithms and a breakdown into its different variations.

Section 3 presents the data pre-processing and the summary results of the selected algorithms.

Section 4 discusses the key learning points of this analysis.

1.4 Abbreviations

Term	Definition
BAU	Business As Usual
CB	Circuit Breaker
CIM	(IEC 61970/61968/62325) Common Information Model for the electricity industry.
CSV	Comma-Separated Value
DCC	Data Collection Company (for SM data)
DD	Data Dictionary
DER	Distributed Energy Resource
DMS	Distribution Management System (such as GE PowerOn)
DNO	Distribution Network Operator
EAC	Estimated Annual Consumption
EAM	Enterprise Asset Management (such as ARM, SAP or Oracle)
EO	Electric Office (NGED's GIS)
ERD	Entity-Relationship Diagram
ETL	Extract, Transform and Load
IEC	International Electrotechnical Commission
GIS	Geographic Information System (such as ESRI or GE Electric Office/Smallworld)
GUID	Globally Unique Identifier
HH	Half-Hourly
HV	High Voltage
LV	Low Voltage
INM	Integrated Network Model
JSON	JavaScript Object Notation
LCT	Low-Carbon Technology e.g. heat pumps, electric vehicles or photovoltaic generation
MC	Measurements Calculator
MDM	Master Data Management
MPAN	Meter Point Administration Number (core – the 13-digit format)
NGED	National Grid Electricity Distribution
NHH	Non-Half-Hourly
NOP	Normally Open Point
ODS	Operational Data Store
OGC	Open Geospatial Consortium
RDBMS	Relational Database Management System

Term	Definition
RMS	Root Mean Square (= $\sqrt{\frac{1}{n}\sum x^2}$)
RMU	Ring Main Unit
SCADA	Supervisory Control and Data Acquisition
SM	Smart Meter
SMITN	Smart Meter Innovations and Test Network
SS	Secondary Substation
SQL	Structured Query Language
TK	Unique INM record identifier
Tx	Transformer
URI	Uniform Resource Identifier
UUID	Universally Unique Identifier
VIPQ	Voltage, Current, Active and Reactive Power
WKT	Well-Known Text (an OGC spatial geometry representation format)
XML	Extensible Markup Language

2 Phase Identification Data Usage

2.1 Trial area

The core trial area comprises 46 SS with installed GridKey monitoring devices from the Milton Keynes area. Customers with SM's and fed from the 46 substations were grouped into three phase groups per SS ('L1', 'L2' and 'L3').

2.2 Data Requirements

The following datasets were used for the phase identification use case:

1. Smart meter voltage data, with 1-minute time resolution.
2. Smart meter voltage data, with 30-minute time resolution.
3. Distribution substation voltage monitoring data, from GridKey loggers (1-minute resolution).
4. Premises to distribution substations connectivity data from CROWN.
5. Electric Office mapping data to provide connection phases of customers where this is already known.
6. Electric Office network data.
7. Phase Validation data from HAYSYS.

2.3 Pre-Calculation Validity checks

The following validation checks were performed and where possible data was corrected.

1. Data Anomalies

Low SM voltage readings (e.g. 0V) may exist in the time-series voltage data due to reading failures. The presence of very low values may have a negative effect in the model, for this reason records with unreasonable voltage readings were assessed and removed or corrected.

2. Clock Offsets

Another issue that may impact the correlation results is the clock offsets between GridKey voltage data and the SM voltage data, i.e. a discrepancy in voltage data time logging between the SMs and the GridKey data. This issue is particularly present in voltage data with lower resolution (i.e., average 1-minute voltage data). An algorithm that identifies the offset difference between the two sources was created and applied. The aim is to synchronise the two voltage curves.

3. Measurement Interval Synchronisation

The HH voltage readings from SM's use intervals with an arbitrary starting time within clock half hours. Similarly,, voltage data with higher time resolutions (i.e. 1 minute voltage data) may start in the middle of minutes. Time synchronisation is important as there is a direct comparison of the voltages from SMs and GridKey for each time period.

As a result, higher resolution voltage data provides not only an improved visibility of the short-term voltage variations, but also a reduction in the time offsets between the measurement intervals used by each SM.

Two approaches have been developed to deal with this issue:

- Timestamp rounding

For the 1-minute voltage data, the timestamp seconds have been rounded to the closest minute, while for the HH data, the minutes have been rounded to the closest 30-minute period.

- Linear Interpolation

Linear interpolation is helpful while searching for a value between a given set of points. It is a method to construct new data points within the range of a discrete set of known data points. In other words, Linear Interpolation allows us to estimate voltage values at times that we do not have them at, given that we have readings soon before and soon after that time.

2.4 Phase Identification Approaches

2.4.1 Approach A: Voltage Correlation between SMs and Secondary Substations

This approach requires measured voltage data to be available for the secondary substations and this data was used as a reference.

The correlation used time-series voltage data from a single-phase SM paired with SS monitoring data (Appendix A.2). For each single-phase SM, three correlation results are obtained, one for each phase measurement at the substation. The assigned phase was the phase with the highest correlation.

The calculation is more complicated for three-phase meters as there are 9 possible single-phase correlations that could be calculated. Selecting the three highest of three correlations for each meter smart phase may result in an invalid phase mapping, for example if more than one smart meter phase has the highest correlation with the same substation phase. This is resolved by calculating the correlation of the six possible phase mappings. The mapping with the highest correlation is selected.

We initially assumed that the connectivity between premises and distribution SS's obtained by CROWN was correct. However, during the SMITN analysis it became clear that there are errors in the mapping between customers and LV feeders and between customers and SS recorded in CROWN. It was known from previous work by Scottish and Southern Electricity Networks that meters with very low correlation results may be fed from a different substation and for this reason outliers were identified and analysed.

Moreover, this approach used both the time-series voltage data and the magnitude of differential step changes between consecutive voltage readings, and their performance was assessed.

Additionally, it is important to understand how the resolution of the voltage data may impact the phase results. National Grid have set the default voltage averaging period for smart meters to 30-minutes. For this project, National Grid have changed their default average voltage measurement period to be 1-minute, hence both 30-minute and 1-minute voltage data are available for the trial area and have been used in correlation and compared.

It is necessary to understand the volume of the sample voltage data that is needed in order to receive reliable results. For this reason, different time ranges were used for both HH and 1-minute data. This will help to understand the volume of the data that is needed to be stored and used in BAU for phase identification.

2.4.2 Approach B: Clustering SMs into 3 groups

In the absence of any SS monitoring, which is the case for the vast majority of secondary substations, the phase identification algorithm attempts to cluster the single phase SMs into three phase groups. The limitation of this approach is the absence of the actual phase reference (L1, L2, L3). However, to deal with the issue of

unbalancing in LV feeder capacity calculations, the real phase labelling is not necessary. The phase information can be stored in a sequence of numbers (1,2,3) and the label can be assigned afterwards. In some single-phase meters, the phases are already known, and this information could be used to assign the real phase in the cluster. Similarly, data may be provided at a later stage if one set of customers are affected by a single phase fault on a known phase.

The algorithm is outlined as follows:

1. Calculation of voltage correlation between each pair of SMs. This is a matrix of dimensions NxN, where N is the number of SMs in the analysis.
2. Clustering the SMs into 3 groups using unsupervised machine learning – Clustering techniques.
3. Repeat step 1,2 using multiple variations of the input data and assess their performance.

Two clustering approaches were applied, K-means and hierarchical clustering and their performances were analysed (see Appendix A3,A4). Fisher-Z transform (Appendix A9) was also applied. This transformation accentuates the differences between correlations that are close to unity and gives transformed values that extend to infinity. When the sample correlation coefficient r is near 1 or -1, its distribution is highly skewed, which makes it difficult for the clustering algorithm to create well-separated clusters. The performance of Fisher-Z transformation has been analysed.

In this approach, we have tested again the algorithm's performance of using the time-series voltage data and the step changes of the voltage. Moreover, the impact of the resolution of the data (1 or 30 minute) has been assessed as well as the performance of the algorithm by using different time intervals.

2.4.3 Validation Data

2.4.3.1 HAYSYS

The HAYSYS phase identification unit, also known as the "phase finder" provides a means of identifying the electrical connected phase of single-phase properties. The data from the phase surveys for our core area was assumed to be correct and used as a standard against which the performance of our algorithms could be assessed.

2.4.3.2 Phase Information from EO

Around 25% of the properties in the core area has their phase populated in EO. The phase information in EO, however, is not always accurately populated. In this analysis, we compared the phase information in EO with the HAYSYS validation data and the results from our selected algorithms.

2.4.3.3 Looped Services

Network data from EO can be used as a validation method for the phase identification use case. The phases of all the connections of a looped service can be assumed to be the same. This is especially important when the other connections of a looped service do not have a SM, hence our selected algorithms cannot be applied.

An algorithm was created to group together the properties that are part of the same looped service. The algorithm traversed the network using a depth first search and assigned the same label for the properties that belong to the looped service.

The "looped services algorithm" was used to help in:

- assessing the performance of the selected algorithms. (two or more SMs that belong to the same looped service should be clustered in the same phase).

- assessing the looped services accuracy in EO, to understand if looped services can be used for properties that have not installed a SM.

3 Results

3.1 Data Preparation

3.1.1 Core Area

In the core area, there are 8809 properties, from which 47% of them have installed a SM up to the end of November 2022. This is slightly higher than the overall average for NGED, which is nearer 40%, as substations with a higher proportion of smart meters were preferentially selected for inclusion in the SMITN test network.

Initially, HH voltage data was requested from the devices for the period 1st May 2022 until 31st August. Around 2200 devices gave a response, which gives a success rate of 57%. A request was then issued to all devices to amend the Average Voltage Measurement Period to 1 minute from the default of 30 minutes. 1810 devices gave a response. 1 minute voltage data was requested and from the beginning of October until 30th November 2022.

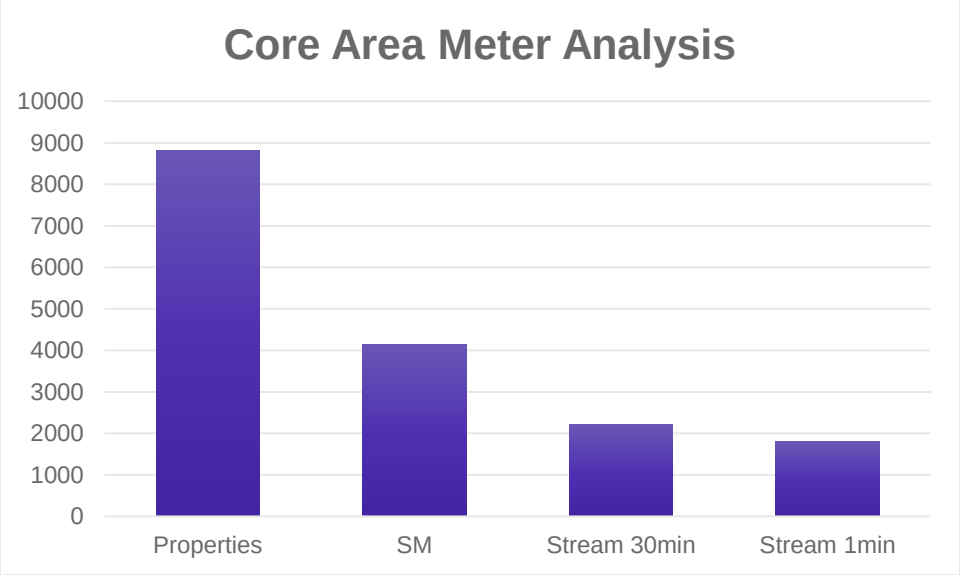


Figure 1: Core Area Device Analysis

3.1.2 Sample Voltage Data

Figures 2 and 3 show an example of SM time-series voltage data and the time-series SS voltage data with 1 minute and HH resolution. The SS monitoring data with one minute time resolution is post-processed to derive HH samples that are aligned with the timing of the 30-minute voltages of SM's.

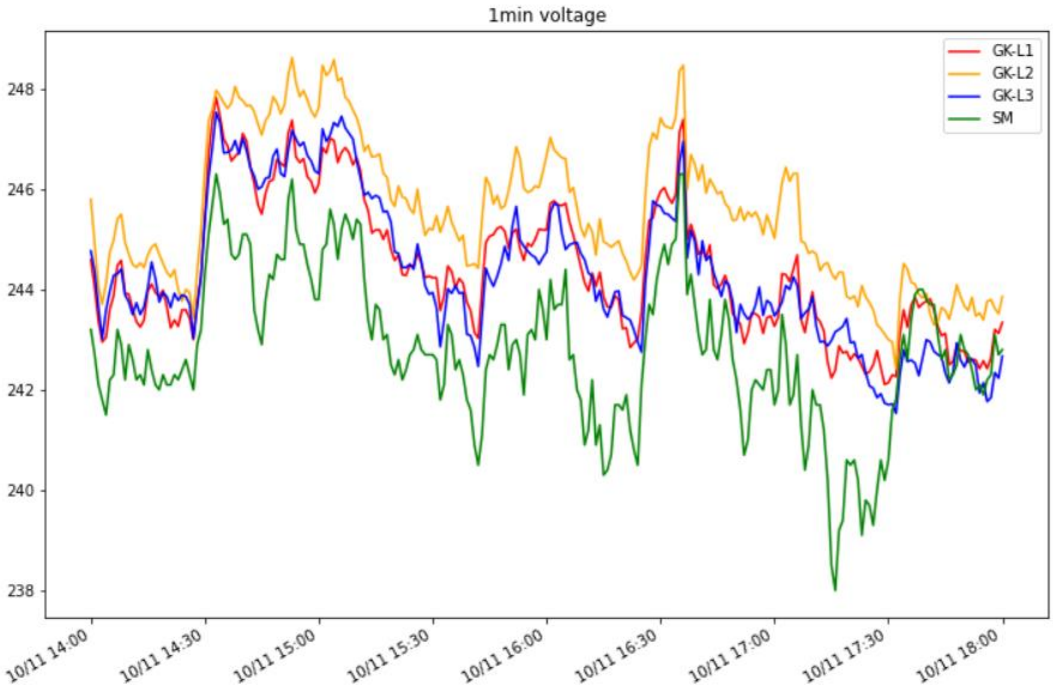


Figure 2:Sample 1-minute voltage data from a SM and its Distribution Substation

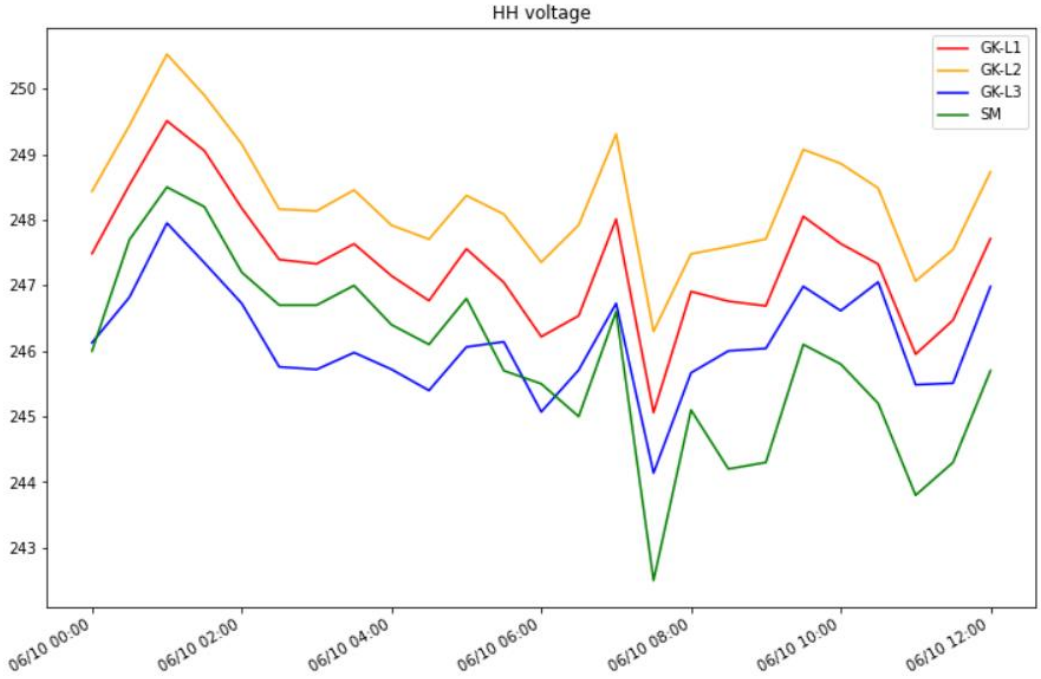


Figure 3: Sample HH voltage data from a SM and its Distribution Substation

3.1.3 Data Cleaning and Meter Delays

Analysis has been undertaken to identify anomalies in the time-series voltage data. In particular, unusual low voltage readings in both 1-minute and 30-minute data. These data quality issues could degrade the performance of the analysis and to the model, so data cleaning was required. Voltage readings lower than 200 V were identified and removed.

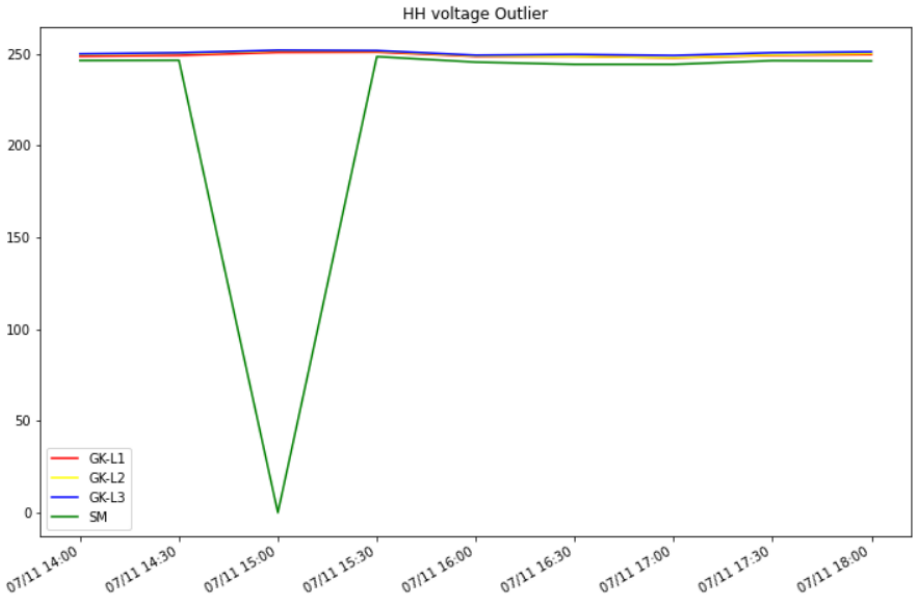


Figure 4:Sample "0" reading in SM data

A second data issue that was identified was SM voltage reading clock offsets relative to substation logger. The timing of the smart meter voltage data was generally well-aligned to the substation monitoring data. However, 56 SMs out of 1810 found to have a clock synchronisation issue relative to substation logger. It is unclear from this data if the offset is introduced from the SM or from the substation logger.

While the voltages from GridKey and the voltages from the SM's have the same timestamps, there are a number of SM's whose voltage profile curves lead the secondary SS's voltage, and others which lag significantly. The presence of this issue is more often in higher resolution data, in our case in 1-minute data. Figure 5 shows an example of a SM voltage having a 5-minute offset. The graph shows a period of 1 hour.

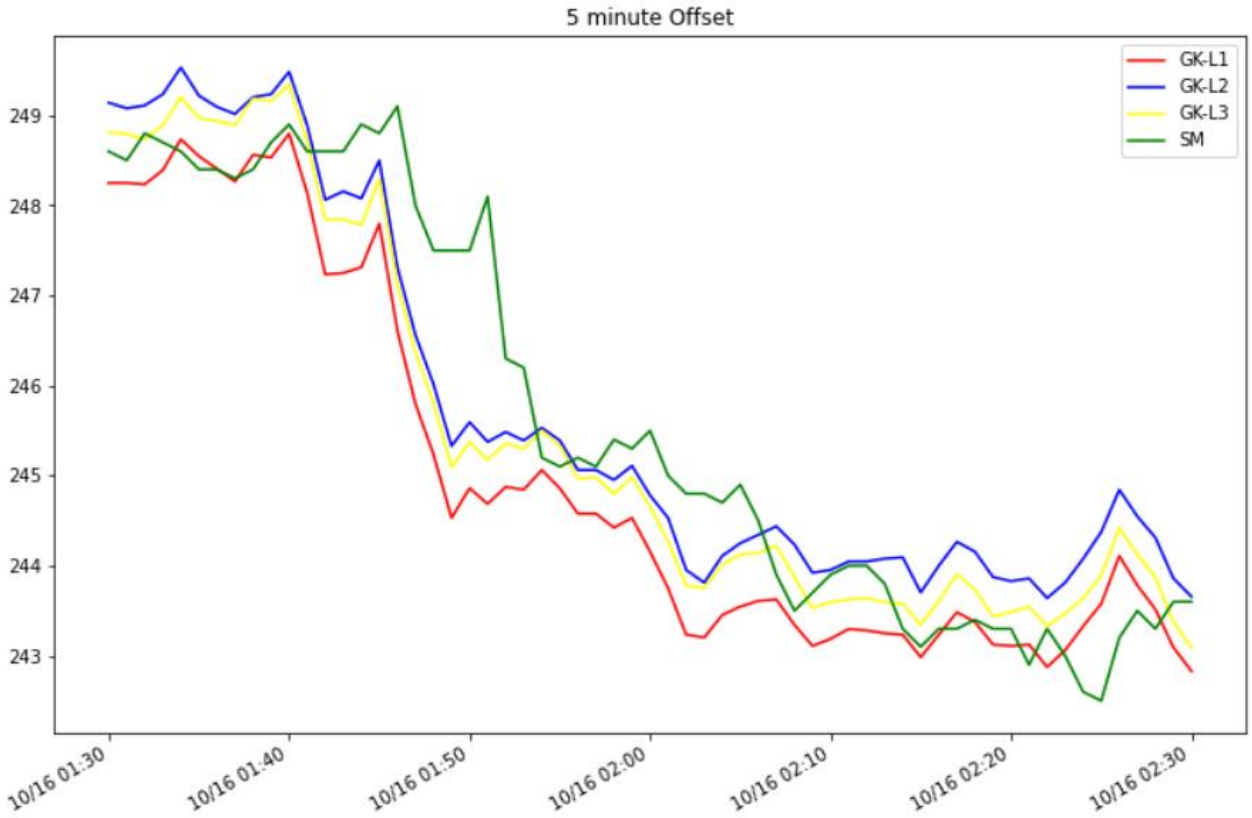


Figure 5: A SM with 5-minute offset relative to substation logger

To deal with this issue, an algorithm was created to identify potential offsets. The offset checking algorithm checks offsets between -5 and +5 minutes with a 15 second resolution. If the algorithm doesn't find an improvement in voltage correlation by a threshold in that range, then it does a broader search, checking between -60 and +60 minutes with a resolution of 60 seconds. The voltage correlation is calculated by using the magnitude of voltage step changes. If the correlation result was improved by more than a set threshold (i.e. 0.1), then the offset was applied to the voltage data. The voltage curve is then corrected. 56 SM devices were identified to have a clock offset relative to substation logger. Table 1 presents the recommended minute offset from the algorithm for sample devices and Figure 6 shows the distribution of the clock offsets for the 56 SMs.

Substation	Device	Approximate minute offset (m)
942757	Device 1	1.75
942086	Device 2	7
945237	Device 3	-0.75
942756	Device 4	1
945113	Device 5	2.25
942756	Device 6	-0.75
942774	Device 7	-2.5
942774	Device 8	-0.75

Table 1:Sample devices with the recommended minute offset

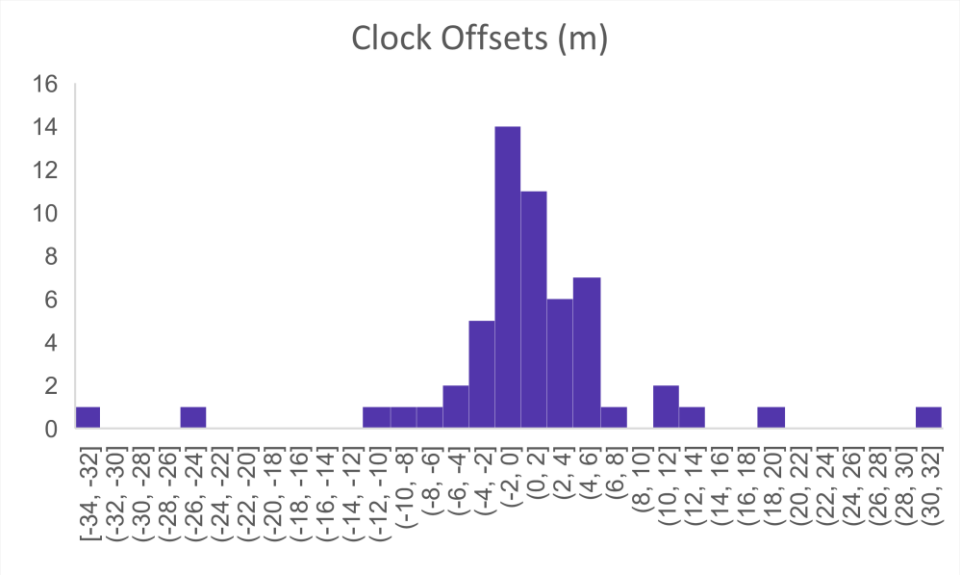


Figure 6:Distribution of clock offsets relative to substation logger (for the 56 SM)

3.1.4 Looped Services

As described in Section 2.6.3, a “looped service algorithm” has been created. First, a connectivity model was created to connect the nodes (properties, connection points etc.) with the branches (cables, OHL etc.). The algorithm then traverses the network using a depth first search, starting from the exit points (properties). Then the trace continues until the algorithm finds a node that connects the service cable with the main cable. The properties that are found in its path are grouped together and labels are assigned.

Figure 7: Sample Network Data from EO with looped Services

Figure 7 illustrates an example of looped services in EO. The properties with green circles, are properties with SMs with the assigned numbers being the result of the looped services algorithm. Both properties in group 391 have an SM installed. Groups similar to 391 were used in order to assess the results from the selected algorithms. At the same time, the looped service algorithm is useful in order to assign phases to properties that haven't installed a SM (i.e. 390,392,395 and 396).

To calculate the accuracy of the looped services algorithm, the following formula has been applied:

Accuracy = the number of looped services where all the devices are put in the same cluster / total number of looped services

3.2 Approach A: Voltage Correlation between SMs and SS - Results

3.2.1 Selection Criteria

To apply and evaluate the selected algorithms for Approach A, each MPAN needs to have:

- Available time-series voltage data,
- Available GridKey time-series voltage data,
- Available validation data from HAYSYS.

There are 1347 devices that fulfil the above criteria for 1-minute data and 1577 that fulfil the criteria for HH data.

To compare the 1-minute versus the HH voltage data, properties that have both 1-minute and HH data available were selected.

3.2.2 Issue with Phase Labelling

The HAYSYS phase finder identifies the phase by setting a reference phase and then assigning the other phases based on the phase angle between the phase and the reference phase. The phase finder assumes a positive rotation for the phases i.e. the sequence is L1, L2, L3. However, some networks have a negative rotation and his causes the phase finder to assign an incorrect label. The incorrect label is dependent on the reference phase used. A description is provided in Appendix A.1.

The incorrect labelling has a negative impact on "accuracy" as accuracy compares the predicted labels with the true labels. For this reason, another metric was used, the Rand index. The Rand index is similar to the accuracy, but it is designed for comparing similarity of clusterings. It does this by going through all pairs of points and checking if they have the same label within each of the two clusterings. While accuracy compares the true labels with the predicted labels, the Rand index looks for the relationship between two points in a dataset rather than the relationship of a point and its true label, and for this reason is invariant to renaming clusters.

Rand index is an important metric in our analysis because it can work even if the labels of the true data are changing in case of voltage reversals. For example, the accuracy for the “942687” distribution substation is very low (0.03), while the Rand index score is high (0.90) (see Table 2). This is an example of wrong labelling in the validation data. However, we need to note that with low sample sizes and where all the results belong to the same cluster, the Rand index is very sensitive to error. (i.e., “940458”). The Rand index has similarities to the accuracy index i.e. values are all between 0 and 1 and the higher the value the better.

Substation number	Frequency	Data	Correct	Total	Accuracy	Rand index
940337	1min	time-series voltage	41	41	1.00	1.00
945113	1min	time-series voltage	36	36	1.00	1.00
942661	1min	time-series voltage	42	43	0.98	0.96
942774	1min	time-series voltage	32	33	0.97	0.96
945487	1min	time-series voltage	26	26	1.00	1.00
942368	1min	time-series voltage	47	49	0.96	0.95
942071	1min	time-series voltage	93	101	0.92	0.90
940458	1min	time-series voltage	2	3	**0.67**	**0.33**
941916	1min	time-series voltage	9	75	**0.12**	**0.82**
942687	1min	time-series voltage	1	39	**0.03**	**0.90**

Table 2: Example correlation results per substation, accuracy vs Rand index

To correct the issue with the labelling and to continue to use the accuracy as a metric in our analysis, the wrong labels in the validation data have been identified and corrected. For the substations that the Rand Index and the accuracy differs more than threshold (i.e. 0.3), the accuracy is re-calculated by going through the 6 permutations of 3 phases for the validation (HAYSYS) data. The (‘L1’,’L2’,’L3’) are being replaced with ('L1', 'L3', 'L2'), ('L2', 'L1', 'L3') etc., and the permutation with the highest accuracy is selected.

The validation labels are corrected. Table 3 presents the SS’s identified with the labelling issue by the algorithm, the new accuracy results, and the suggested label change.

Substation number	Accuracy	Rand index	Accuracy after permutation	Label rotation
941985	0.35	0.73	0.78	L1->L2, L2->L1, L3->L3
940717	0.12	0.85	0.88	L1->L2, L2->L1, L3->L3
941988	0.31	0.66	0.69	L1->L2, L2->L1, L3->L3
941987	0.26	0.71	0.74	L1->L2, L2->L1, L3->L3
941916	0.12	0.82	0.85	L1->L3, L2->L1, L3->L2
942687	0.03	0.90	0.92	L1->L3, L2->L1, L3->L2

Table 3: Labelling correction after phase permutation

3.2.3 Approach A-Optimisation

This section will compare different variations and approaches discussed in 2.4.1 with the aim of finding the best variation that improves the correlation results.

3.2.3.1 Time Reference Rounding Vs. Linear Interpolation

To address the clock synchronisation issue, two approaches have been developed, rounding and linear interpolation as mentioned in section 2.4. Tables 4 and 5 present the accuracy of both approaches using the total number of MPANs that fulfil the criteria discussed in Section 3.2.1.

Frequency	Data processing	Data	Total	True	Accuracy
1min	**Linear Interpolation**	Step Changes	1347	1116	0.83
1min	**Rounding**	Step Changes	1347	1115	0.83

Table 4: Linear Interpolation vs Rounding in 1 min data

Frequency	Data processing	Data	Total	True	Accuracy
30min	**Linear Interpolation**	Step Changes	1577	1312	0.83
30min	**Rounding**	Step Changes	1577	1311	0.83

Table 5: Linear Interpolation vs Rounding in 30 min data

From Tables 4,5, we can observe that no major differences were found between the two approaches. For the rest of the analysis, linear interpolation was used.

3.2.3.2 Time-Series Voltage data Vs Step changes in Voltage Correlation

In this section, the use of time-series magnitude voltage data (time-series voltage) and the time-series of the voltage step changes (step changes) are compared, and their correlation results are analysed. The step changes are calculated by taking the difference of voltage between consecutive periods ($v(t+1) - v(t)$). It is expected that the use of voltage step changes will magnify the differences between the GridKey and the SM voltage data. Tables 6 and 7 present the summary results.

Resolution	Data	Total	True	Accuracy
1min	**Time-series voltage**	1347	1109	0.82
1min	**Step changes**	1347	1116	0.83

Table 6: time-series voltage data vs step changes

Resolution	Data	Total	True	Accuracy
30min	**Time-series voltage**	1577	1308	0.83
30min	**Step changes**	1577	1312	0.83

Table 7: time-series voltage data vs step changes

It is noted that the use of magnitude of differential step changes have a small positive impact in the correlation results, with higher impact in 1 minute data.

3.2.3.3 30 min vs 1 min

This section considers how the resolution of the data (i.e. 1 min/HH) may impact the correlation result, as well as the algorithms sensitivity to the duration of the input data. For a better comparison between the 1-minute

and HH voltage data, we selected the properties that have both HH and 1-minute voltage data available. For the purpose of this analysis, 1232 properties have been selected.

The algorithm breaks down the data into different time ranges. Initially the algorithm uses all of the available voltage data and then breaks them down by month, week, day and 6 hourly intervals and calculates the average accuracy and the standard deviation for each time period by SS. In Table 8, we can view the summary results for substation “942037”. For one minute data, the accuracy stays the same across all different intervals. However, using HH data we notice deviations in the results. Specifically, the accuracy decreases when utilising a lower number of samples.

			Avg accuracy	Avg Rand index	Accuracy std
Frequency	**Correlation**	**Interval**			
1min	Step Changes	All (2 months)	0.987	0.983	-
		month	0.980	0.975	0.0096
		week	0.985	0.981	0.0047
		day	0.984	0.981	0.0045
		6hour	0.985	0.981	0.0043
30min	Step Changes	All (4 months)	0.987	0.983	-
		month	0.977	0.970	0.0200
		week	0.984	0.980	0.0051
		day	0.949	0.937	0.0328
		6hour	0.793	0.801	0.1309

Table 8: Comparison of 1 minute and HH voltage data using different time intervals for SS 942037

Figure 8 illustrates the average results for all the selected SS. Using 1-minute data, the phase identification accuracy is maintained even with a measurement duration reduced to a period of 6 hours. However, the accuracy with HH data reduces if the duration is reduced below one week. This suggests that the additional burden of setting up the meters to record data at 1 minute intervals is worthwhile and that the meter could be reset to normal operation after a relatively short amount of time, say a week, to minimise the operational impact of the additional data retrieval.

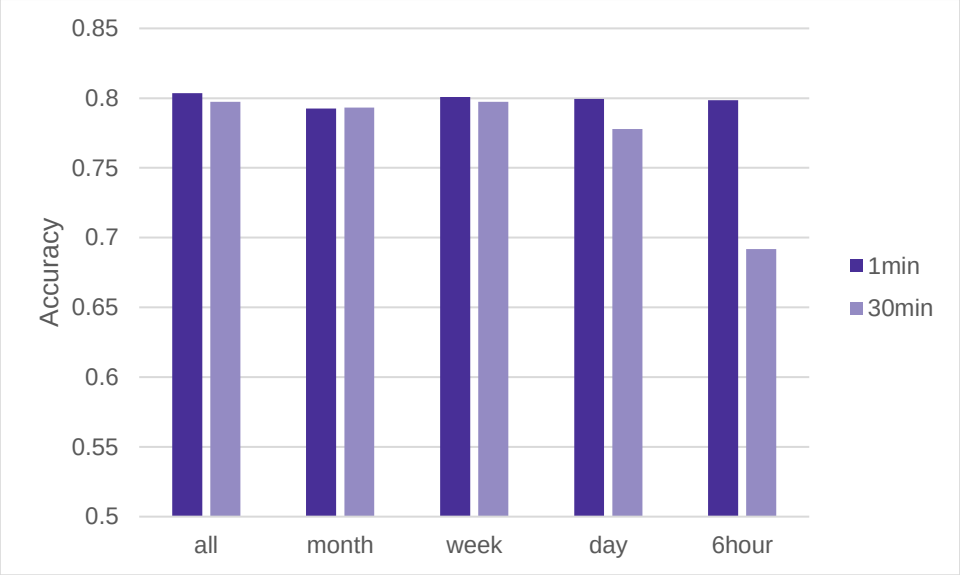


Figure 8: Summary results for all the substations using different time intervals (Approach A)

3.2.4 Synthetic data with different time intervals

In section 3.2.3, we saw that the accuracy decreased when the algorithm used the HH voltage data instead of the 1-minute data. The possible reasons are:

- the algorithm works better with higher resolution data, because higher resolution data captures more information,
- the impact of measurement interval synchronisation and clock offset in HH data (see section 2.3).

However, data with high resolution such as the 1-minute data are very expensive to store and maintain. In this section, we will analyse how different resolutions of the data may impact the results. We will resample the 1-minute data to synthesize 5, 10, 15 minute and HH voltage data. To compare the real HH data with the synthetic HH, we chose only the devices that are available both in 1-minute and HH data.

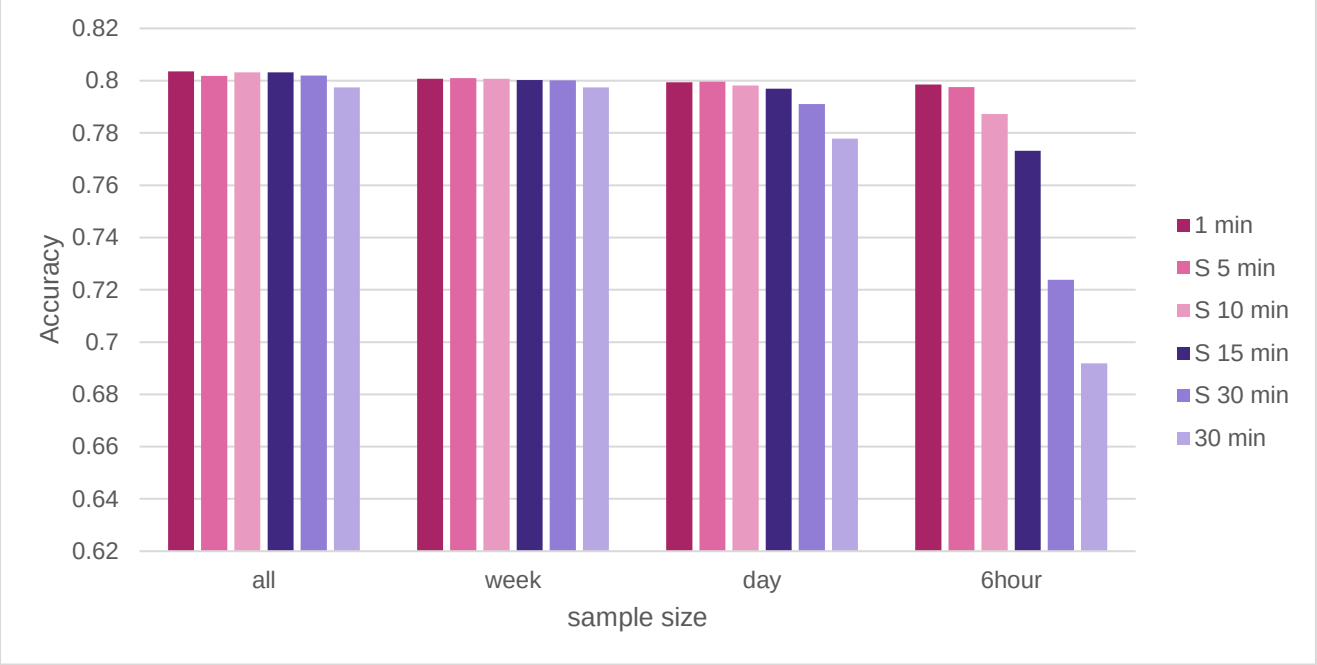


Figure 9: Summary results using synthetic and real data with different resolutions and time intervals (Approach A)

From Figure 9 we can observe that the accuracy drops as we decrease the resolution of the data, especially when the duration is reduced below one week. It is noted that the accuracy decreases significantly in both the synthetic HH and the real HH data, with further decrease in real HH data. This was expected due to the measurement interval synchronisation and clock offsets issues.

The accuracy in the synthetic 5-minute data did not decrease even when we use only a 6 hour time interval, while there was a small drop in accuracy in the 10-minute synthetic data.

3.2.5 Summary Results per Substation

For Approach A "direct correlation between the SM and the SS voltage data", the proposed algorithm with overall best performance uses:

- the step changes of the voltage magnitude,
- 1 min data, as the results are more reliable even if we choose a small amount of data (6 hours). However, the use of 30-minute data with longer time intervals (i.e 3-4 months of data) are proved to be as sufficient as the 1-minute data.

The overall results per each substation are presented in Table 9. 60% of the substations have accuracy more than 0.8 (coloured green), 31% have accuracy between 0.8-0.5 (coloured yellow), and the 9% (3 SS) have accuracy lower than 0.5 (coloured orange) and they need further investigation.

Substation	Frequency	Data	Interval	Accuracy	Rand index	True	Total	Looped service accuracy
940337	1min	Step Changes	all	1.000	1.000	41	41	1.000
942839	1min	Step Changes	all	1.000	1.000	22	22	0.500
945113	1min	Step Changes	all	1.000	1.000	36	36	-

945487	1min	Step Changes	all	1.000	1.000	26	26	-
942037	1min	Step Changes	all	0.988	0.984	80	81	-
942661	1min	Step Changes	all	0.977	0.961	42	43	1.000
942774	1min	Step Changes	all	0.970	0.956	32	33	-
942755	1min	Step Changes	all	0.959	0.949	47	49	1.000
942368	1min	Step Changes	all	0.959	0.952	47	49	0.833
942084	1min	Step Changes	all	0.957	0.945	22	23	-
942681	1min	Step Changes	all	0.949	0.936	75	79	0.667
945237	1min	Step Changes	all	0.947	0.930	18	19	-
942840	1min	Step Changes	all	0.923	0.895	24	26	-
942687	1min	Step Changes	all	0.921	0.898	35	38	1.000
942071	1min	Step Changes	all	0.921	0.900	93	101	-
942756	1min	Step Changes	all	0.920	0.902	46	50	1.000
942367	1min	Step Changes	all	0.917	0.888	22	24	1.000
940717	1min	Step Changes	all	0.88	0.852	15	17	1.000
941916	1min	Step Changes	all	0.853	0.824	64	75	-
942647	1min	Step Changes	all	0.810	0.786	17	21	1.000
944922	1min	Step Changes	all	0.806	0.778	29	36	0.500
941985	1min	Step Changes	all	0.784	0.727	29	37	-
942657	1min	Step Changes	all	0.754	0.739	43	57	1.000
942695	1min	Step Changes	all	0.750	0.768	15	20	-
941987	1min	Step Changes	all	0.743	0.708	26	35	-
941988	1min	Step Changes	all	0.688	0.658	11	16	-
940458	1min	Step Changes	all	0.667	0.333	2	3	-
942683	1min	Step Changes	all	0.639	0.662	23	36	-
942369	1min	Step Changes	all	0.636	0.667	49	77	1.000
942651	1min	Step Changes	all	0.629	0.630	22	35	1.000
942649	1min	Step Changes	all	0.604	0.634	32	53	-
942680	1min	Step Changes	all	0.571	0.714	4	7	-
942086	1min	Step Changes	all	0.348	0.575	24	69	-
942842	1min	Step Changes	all	0.333	0.333	1	3	-
942757	1min	Step Changes	all	0.200	0.556	2	10	-

Table 9: Summary Results for each SS

We should note that the metrics in substations with very small numbers of SMs (i.e., less than 10) can often be very high or low (i.e. 942842 etc) and the low performance of the algorithm can be explained.

3.2.5.1 Comparison between EO phase, Looped services, HAYSYS and Selected Algorithms

In this section, EO phase information and the looped services EO network data are assessed using both the HAYSYS validation data as well as the selected algorithms results.

1 HAYSYS vs EO: To compare the HAYSYS and the EO phase information, properties with and without SMs were selected, which had phase information available from the two systems. HAYSYS agrees with EO phases for only 52% of the MPANs.

2 Selected Algorithms vs EO: In this case, properties with SMs only selected that have EO phase information populated. The selected algorithms also agree with EO for only 52% of the MPANs .

3 When the results from the selected algorithms were compared to the HAYSYS phase data these were in agreement for 83% of the MPANs.

4 HAYSYS vs Looped Services: The looped service metric is the number of looped services where all the devices are put in the same cluster, divided by the number of looped services, without taking into account the actual label of the phase. 89% of properties that belong to a loop service in EO has the same phase based on HAYSYS labels.

5 Selected Algorithms vs Looped Services: 90% of properties that belong to a loop service in EO have the same phase based on the results from the selected algorithms.

Table 10 shows the summary results of the comparison between HAYSYS, EO phase, looped Services and Selected Algorithms.

	HAYSYS	EO	Selected algorithms	Looped services
HAYSYS		52%	83%	89%
EO			52%	99%
Selected Algorithms				90%
Looped Services				

Table 10 HAYSYS vs EO vs Looped Services vs Selected Algorithms

3.3 Approach B: Clustering SMs into 3 groups - Results

3.3.1 Hierarchical Clustering VS K-means

In the absence of monitoring data in the SS, the algorithm groups the SMs into three groups using unsupervised machine learning. First, voltage correlation between each pair of SM's is calculated and then, two clustering algorithms, Hierarchical clustering and K-means have been applied. The summary results for all the substations are presented in Table 11.

To assess the performance of the clustering algorithms, three metrics have been used, accuracy, Rand Index and Fowlkes Mallows score. The Rand Index is similar to accuracy, but it is designed for comparing similarity of clusterings. It does this by going through all pairs of points and checking if they have the same label within each of the two clusterings (see Appendix A.7). Fowlkes Mallows score is another metric that measures the similarity of two clusterings of a set of points (see Appendix A.8). Both Rand Index and Fowlkes-Mallows score requires "truth" data, in our analysis, the HAYSYS data are used for validation. The score ranges from 0 to 1. A high value indicates a good similarity between two clusters.

As discussed in section 3.2.2, the Rand Index and Fowlkes-Mallows score can work accurately, even if the phase labels are changing but the groups of SMs remain the same. The "accuracy", however, compares the predicted label with the true label. For this reason, to calculate the accuracy, we had to assign in each resulted

group (i.e., 1,2,3) a phase label (L1,L2,L3). The HAYSYS data were used to assign labels in each cluster. The label for each cluster will be determined based on the greatest number of labels of MPANs within that cluster.

From the Table 11, we can observe that using the voltage step changes instead of the voltage time series data has improved the results significantly both in 1-minute and HH data. Moreover, the hierarchical clustering performs better than the K-means clustering. To optimise the hierarchical clustering, different variations of hierarchical clustering were used, with the highlighted line giving the best results. For more information see Appendix A.4.

Frequency	Data	Algorithm	Distance type	Linkage method	Avg Rand index	Avg Fowlkes-Mallows score	Avg accuracy
1M	Voltage	hierarchical	squareform	ward	0.7315	0.6519	0.7383
1M	Voltage	hierarchical	squareform	complete	0.6984	0.6318	0.7011
1M	Voltage	k-means	-	NULL	0.6954	0.5991	0.6968
1M	Voltage	hierarchical	pdist	ward	0.6922	0.6032	0.6887
1M	Voltage	hierarchical	pdist	complete	0.6384	0.5876	0.6534
1M	Step changes	hierarchical	squareform	complete	0.7845	0.7010	0.7913
1M	Step changes	hierarchical	squareform	ward	0.7816	0.6771	0.7776
1M	Step changes	k-means	-	NULL	0.7632	0.6656	0.7596
1M	Step changes	hierarchical	pdist	ward	0.7422	0.6408	0.7442
1M	Step changes	hierarchical	pdist	complete	0.7276	0.6341	0.7299
HH	Voltage	hierarchical	squareform	ward	0.6945	0.6098	0.6871
HH	Voltage	hierarchical	squareform	complete	0.6489	0.5675	0.6407
HH	Voltage	k-means	-	NULL	0.6383	0.5478	0.6316
HH	Voltage	hierarchical	pdist	ward	0.6292	0.5498	0.6169
HH	Voltage	hierarchical	pdist	complete	0.5899	0.5445	0.5897
HH	Step changes	hierarchical	squareform	ward	0.7072	0.6084	0.7004
HH	Step changes	hierarchical	squareform	complete	0.6761	0.5964	0.6791
HH	Step changes	k-means	-	NULL	0.6667	0.5453	0.6396
HH	Step changes	hierarchical	pdist	ward	0.6578	0.5400	0.6352
HH	Step changes	hierarchical	pdist	complete	0.6159	0.5377	0.6066

Table 11: Clustering Optimisation

3.3.2 Fisher Z transformation

After calculating the NxN matrix with the voltage correlation for each SM pair, Fisher Z transformation has been applied. This transformation accentuates the differences between correlations that are close to unity and gives transformed values that extend to infinity (Appendix A.9). The results are presented in Table 12. The application of the Fisher Z transformation didn't improve the performance of the model.

Frequency	Data	Algorithm	Fisher Z	Avg Rand index	Avg accuracy
1M	Step Changes	k-means	No	0.7632	0.7596
1M	Step Changes	k-means	Yes	0.7539	0.7400
1M	Step Changes	hierarchical	No	0.7816	0.7778
1M	Step Changes	hierarchical	Yes	0.6962	0.6844
HH	Step Changes	k-means	Yes	0.6667	0.6396
HH	Step Changes	k-means	No	0.6780	0.6580
HH	Step Changes	hierarchical	No	0.7072	0.7003
HH	Step Changes	hierarchical	Yes	0.6498	0.6435

Table 12: Clustering Results with and without Fisher Z transformation

3.3.3 1 min vs HH based on different time intervals

In Approach A, we discussed that the HH voltage data are as effective as the 1-minute data, only when we use longer time intervals for example 4 months of data. However, in Approach B, the average accuracy per substation of the clustering has decreased significantly, from 0.78 to 0.71 even when we use all of the available HH data. In Figure 10, it is observed that the performance of the model was 0.78 when 1-minute data are used as input to the model and the performance stays high even when we decreased the data points that are used as an input (i.e., 6 hours). For the HH data, the performance decreases even further as we decrease the time intervals. Again, this suggests that the additional effort required to obtain 1 minute voltage data would be worthwhile due to the improved accuracy.

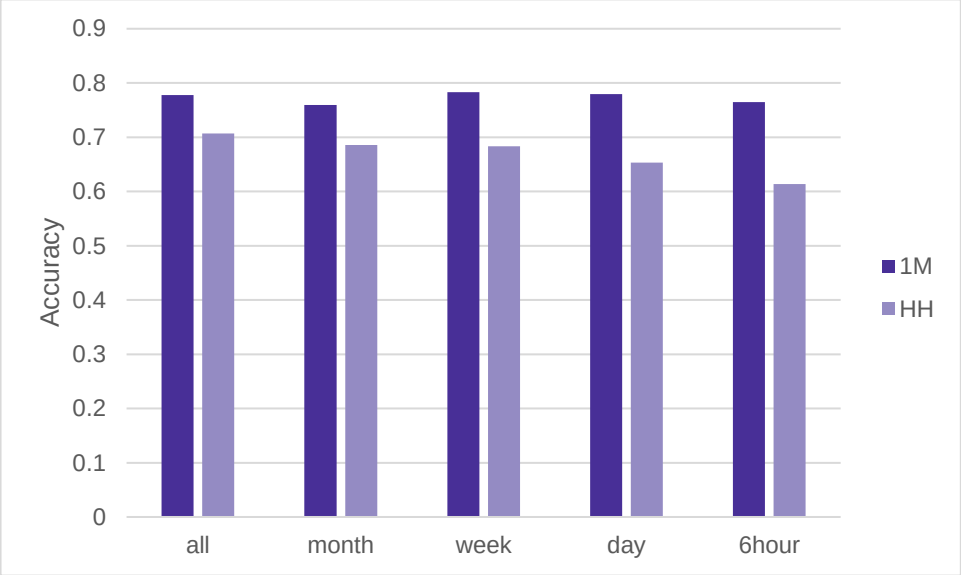


Figure 10:HH vs 1- minute data using different time intervals (Approach B)

It is noted that to be able to compare the performance of the model using the 1 minute and HH voltage readings, we are using only the devices that have both 1 minute and HH data available.

3.3.4 Synthetic data with different time intervals

Higher resolution data gave more accurate results especially when used as an input into the Approach B. 5-minute or 10-minute average voltage data may provide an optimum balance between accuracy and data management effort (see section 3.2.4). For this reason, we re-ran the clustering algorithms using synthesised 5, 10, 15 minute and HH voltage data and compared the real HH results with the synthetic HH data.

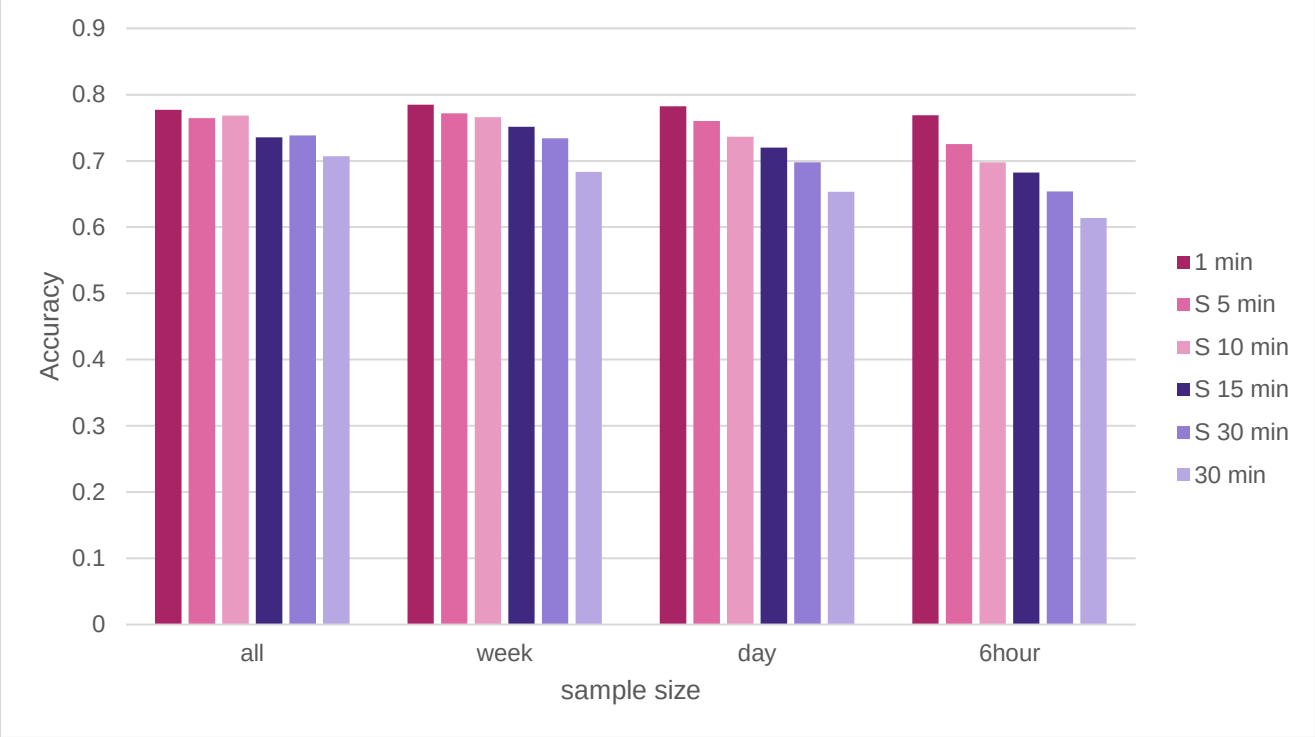


Figure 11: Accuracy using real and synthetic voltage data with different resolution

The performance of the clustering algorithm is decreasing as we decrease the resolution of the voltage data even when we use larger samples. However, the performance of the algorithm using synthetic 5-minute and 10-minute data is very close to 1-minute data when the duration is above one week. Moreover, we can observe that the accuracy in real HH data is lower than the synthetic HH data. This highlights the impact of the measurement interval synchronisation and clock offsets in phase identification approaches.

3.3.5 Approach B – Summary Results

For the Approach B "Clustering SMs into 3 groups", the proposed algorithm with overall best performance uses:

- the step changes of the voltage magnitude,
- Hierarchical Clustering,
- 1-minute voltage data.

Table 13 shows the summary results from the clustering algorithm per SS. 53% of the SS have accuracy more than 0.8, 44% have accuracy from 0.5-0.8 and only 1 SS (3%) scored less than 0.5.

Approach	Percentage of Substations (%)	Accuracy Score
B	53%	> 0.8
B	44%	0.5-0.8
B	3%	< 0.5

Table 13: Summary results for Core Area

Figure 12 shows an example of the optimised clustering results for “942084” SS. On the right, there is a heatmap, which shows the correlation results for each pair of SM. The green colours represent the pair of SMs that are highly correlated, while the red colours are the SMs with very weak correlation. In this figure, the SMs have been ordered based on the clustering group that the algorithm assigned them (1,2,3). We can observe, how the algorithm managed to create three very distinguished clusters, one for each phase.

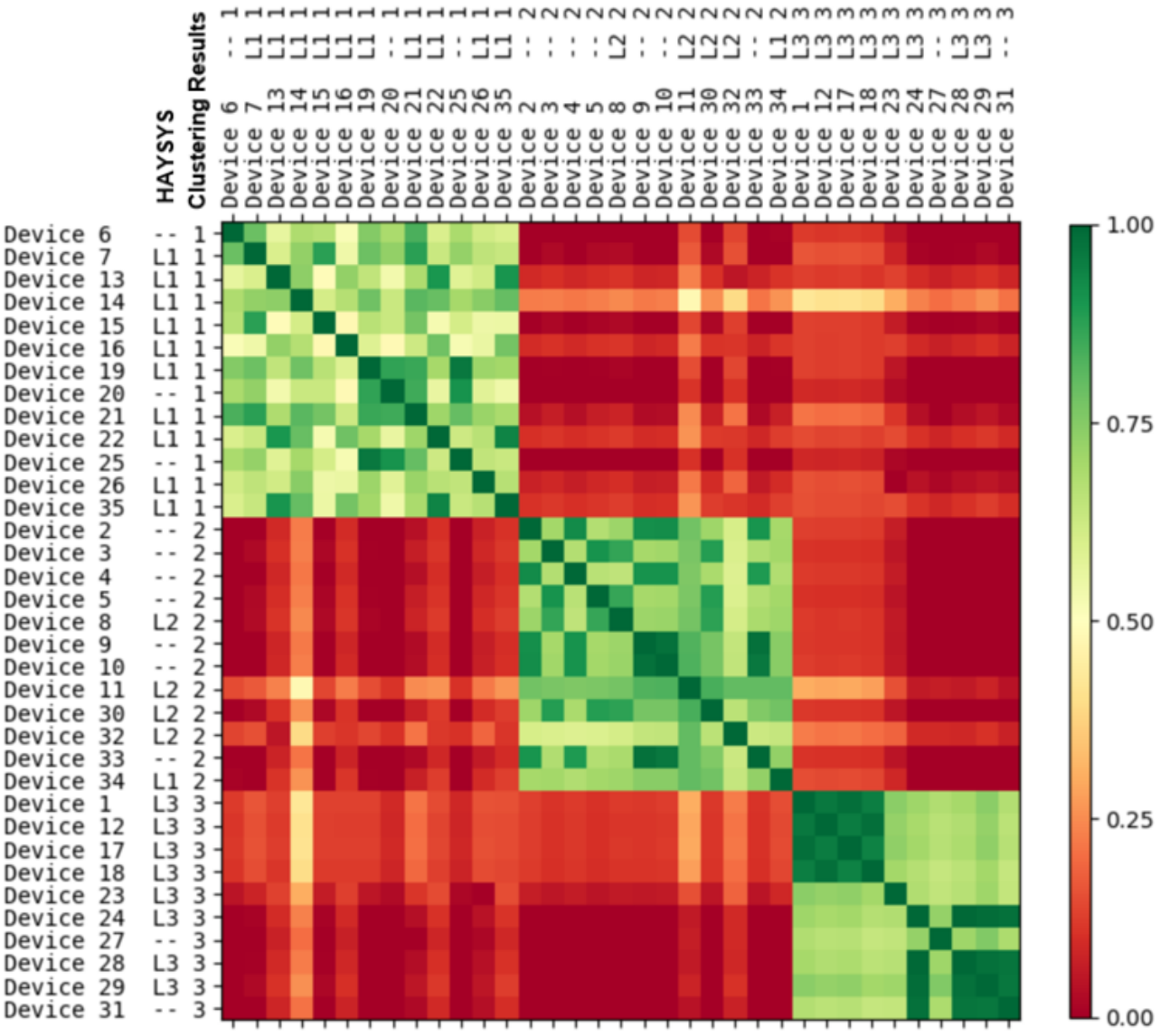


Figure 12: Correlation Heatmap and Clustering results

3.3.6 Approach A vs Approach B

To compare the Approach A and Approach B, we selected the SM devices that can be used for both approaches and fulfil the criteria discussed in section 3.2.1. In Approach A, from a total of 1347 devices available in 1-minute data, 1116 has the same phase information as the HAYSYS validation data, while Approach B scored less. Table 14 shows the optimised summary results for the two Approaches using the HAYSYS validation data, EO and looped Services. The much lower accuracy values using EO data for

Validation appears to be an anomaly. This suggests further validation of the accuracy of the EO data is required to establish whether this low accuracy is due to inaccuracy within the EO data.

Approach	Resolution	Validation	Total	Correct	Accuracy
A	1 min	HAYSYS	1347	1116	**0.83**
B	1 min	HAYSYS	1347	1076	**0.80**
A	1 min	EO	414	227	0.55
B	1 min	EO	414	199	0.48
A	1 min	Looped Services	54	49	0.89
B	1 min	Looped Services	54	50	0.92

Table 14: Approach A vs Approach B per device

Figure 13 compares the performance of the 2 approaches for each substation. We can observe that Approach B scores slightly less in most of the substations.

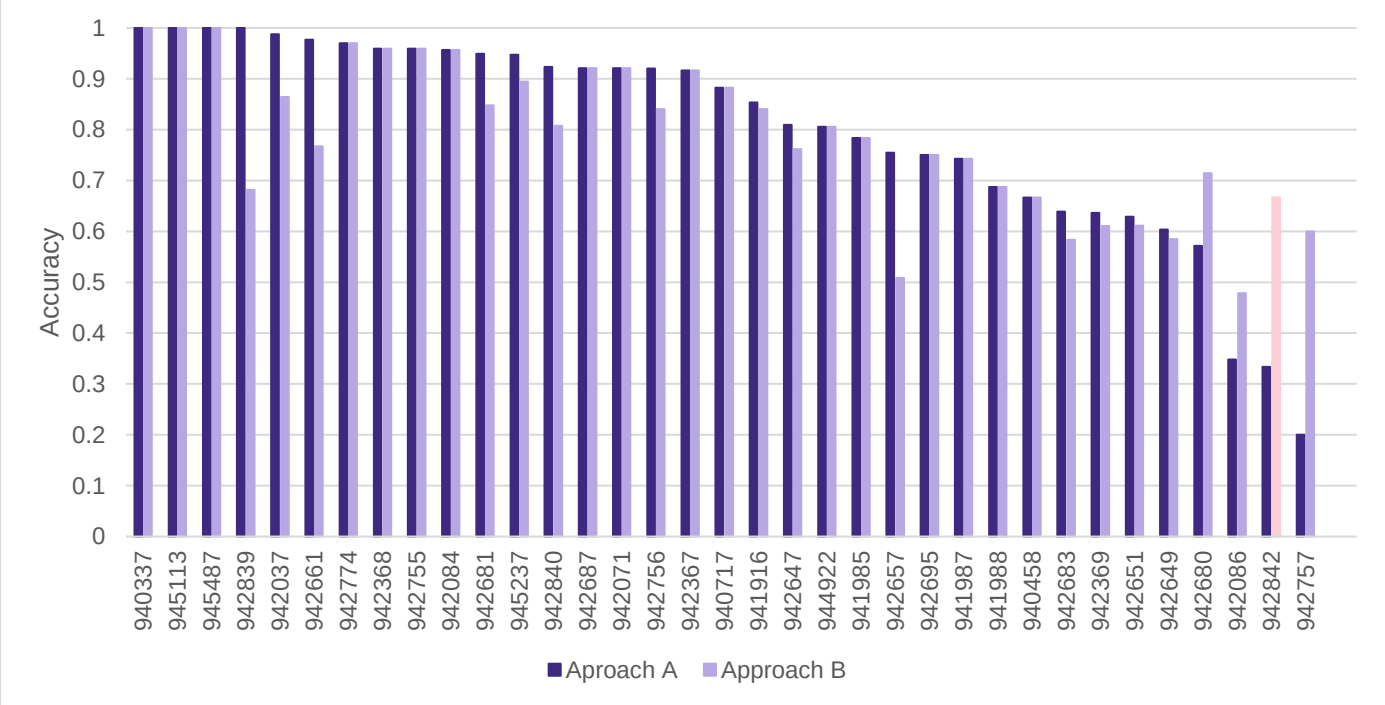


Figure 13: Approach A vs Approach B per substation

It is noted that, the “942842” SS has a very small number of devices available, less than 4, so the results are not reliable especially in the clustering algorithms and should be discarded from the analysis.

3.3.7 Approach A and B agreement

From the summary results, we can observe that the results for some SS are more reliable than others. The purpose of this section is to assess if the results from Approach A and B consistently agree. The accuracy of the Approach B is calculated using the results from the Approach A as validation set.

Moreover, Approach A will work well regardless of the number of SMs per SS, as we apply direct comparison between each SM and GridKey data, but Approach B could be compromised if there are insufficient neighbours to form a cluster group.

In Figure 14 we can observe that the results from the Approaches A and B are consistent in more than 90% of the SS. Further investigation is needed in the 3 SS where the performance of Approach B is lower than 0.8. Moreover, there is no evidence that the performance of the approach B is affected in SS with lower number of SMs.

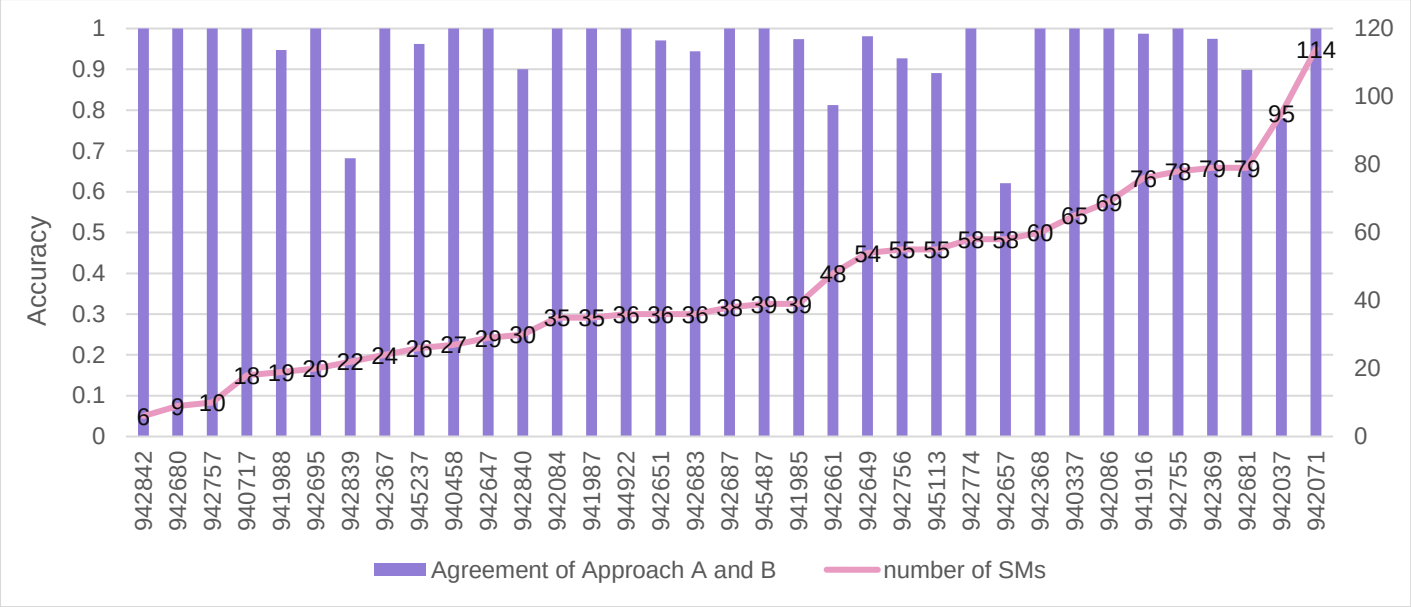


Figure 14: Approach A and B agreement

4 Summary of Learning Points

This analysis has presented work on identifying the phase that each property is connected on the network in 46 SS in the Milton Keynes area using SM voltage data. The main findings from this analysis are:

i. Voltage data from SMs can be used to effectively identify the phase to which each property is connected.

ii. Voltage correlation between SMs and SS monitoring has proven to be a promising technique and phases were identified with around 83% accuracy. However, it relies on SS monitoring devices to be installed correctly in each SS. More substation monitoring will be rolled out in the next five years so phase data could initially be provided by using the clustering method and then revisited to use correlation where monitoring is installed.

iii. Clustering SMs into three groups has been proven to be an efficient technique with accuracy 80%. However, a limitation of this technique is that the actual phase label is unknown and other data sources need to be used to label the resulted classes. Hierarchical clustering performs better than the K-means clustering.

iv. The results from the voltage correlation between SMs and SS and the clustering techniques consistently agree in more than 90% of the substations with accuracy higher than 0.9. This gives an extra confidence that the results from the two approaches are reliable. However, further investigation is needed for the substations that approach A and B are consistent but scored less than 0.8 using HAYSYS survey data.

v. Higher resolution voltage data can give more accurate results, especially when used as an input into the clustering algorithms. Data with lower resolution requires a greater number of time samples to produce more stable results but even with longer data sets the accuracy tends to be lower than with higher resolution data. The analysis of using synthetic 5-minute and 10-minute average data shows that it could be used as an optimum balance between accuracy and data management effort and this is an area for further investigation.

vi. Using the magnitude of the step changes of voltages works more efficiently than using the time-series voltage data as input in the algorithms.

vii. EO phase information is around 50% accurate based on the HAYSYS validation data and the results from the selected algorithms. However, most of the looped services that exist in the network data from EO agree with HAYSYS and the selected algorithms, which makes the network data more reliable.

viii. The trials have demonstrated that the transfer of voltage data from smart meters is only partially successful, and that some meters do not respond to reconfiguration commands to change their measurement time resolution. These issues reduce the scope for accurate phase identification and would ideally be investigated and addressed.

ix. The voltage data from many of the smart meters uses measurement intervals that are not aligned to clock minutes of half-hours, such that 30-minute samples may begin at any time within the half-hour, and 1-minute samples many begin at any time within the minute. Using higher resolution data therefore provides greater time synchronisation between the measurement intervals, in addition to giving greater visibility of the voltage variations. Re-running the analysis with synthetic HH data, which are derived from the 1-minute data, has shown an improvement in the results in comparison to real HH data. This is an indicator that the phase identification results would be improved if future smart metering

specifications defined that measurement intervals should be synchronised to clock half-hours or minutes.

Future Work

1. Determine if 1 minute voltage data gathering could impose too high a burden on the smart meter management and reading processes. If so determine whether there is value in investigating the accuracy for 5 or 10 minute average data.

2. Carry out a separate assessment of the accuracy of EO database phase data, for example by comparing to text data on the LV geoschematic diagrams.

3. The coloured labels seen in some of the photographs of the meter box interiors should also be compared to the other sources of phase data to determine if this is a useful source of information

4. From the summary results, we can observe that the results for some SS are more reliable than others. Further analysis is needed to understand the issues that may degrade the performance of the models in some substations. Potential issues are:

    - The labels from validation data from HAYSYS,
    - Balanced network, where there is little difference between phase voltages,
    - Local effects on voltage by LCTs.
5. Three-phase meters will be added into the analysis, in order to find the actual phase of each phase of the device.

5 Appendix

A.1 Reverse Rotation and HAYSYS Phase labelling

Within a three-phase electricity supply, each phase can be considered a phasor with phasors rotating anti-clockwise considered positive rotation and clockwise considered negative rotation.

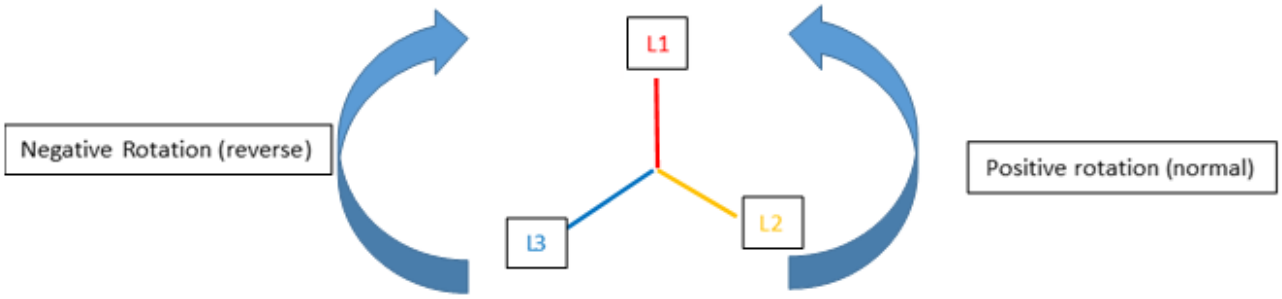


Figure 15:Phase Rotation

The HAYSYS phase finder establishes the phase by using a reference phase and then the sequence of the other phases are assigned based on positive rotation e.g. if the reference phase is set as L1 then the next voltage peak will be L2 followed by the L3 peak. The global reference used is L1 but this can be set locally by identifying any phase at the Distribution substation and setting this as a reference.

With positive rotation we get

- L2 lagging L1 by 120deg (or leading L1 by 240deg)
- L3 lagging L1 by 240deg (or leading L1 by 120deg)

With negative rotation we get

- L3 lagging L1 by 120deg (or leading L1 by 240deg)
- L2 lagging L1 by 240deg (or leading L1 by 120deg)

The HAYSYS Phase Finder Unit uses one phase conductor as a reference and makes an assumption that the other two phase conductors will have the phases that would arise with conventional positive rotation. Since only one conductor is used to define the reference, it is not possible to determine whether the substation has positive or negative rotation. If the system has negative rotation, the incorrect phase is dependent on the reference voltage used.

- If the reference voltage is L1 then under negative rotation it will detect L2 as L3 and L3 as L2.
- If the reference voltage is L2 then under negative rotation it will detect L3 as L1 and L1 as L3.
- If the reference voltage is L3 then under negative rotation it will detect L2 as L1 and L1 as L2.

A.2 Pearson Correlation

Pearson Correlation measures the strength of the linear relationship between two variables. It has a value between -1 and 1 with a value of -1 meaning a total negative linear correlation, 0 being no correlation and +1 meaning a total positive correlation.

In other words, if the value is in the positive range, the relationship between variables is positively correlated, and both values decrease or increase together. On the other hand, if the value is in the negative range, it shows that the relationship between variables is negatively correlated, and both values will go in the opposite direction.

Pearson's correlation formula is as follows.

$$\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

A.3 K-means

K means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. k-means clustering minimizes within-cluster variances.

The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may prevent the algorithm from converging. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

Given an initial set of k means m1(1),...,mk(1) the algorithm proceeds by alternating between two steps.

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.)

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)\dashv} \right\|^2 \ \forall_j, 1 \leq j \leq k \right\}$$

Where each xp is assigned to exactly one S(t) even if it could be assigned to two or more of them,

Update step: Recalculate means (centroids) for observations assigned to each cluster

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \epsilon S_i^{(t)} {}_J^x}$$

The algorithm has converged when the assignments no longer change. The algorithm is not guaranteed to find the optimum.

A.4 Hierarchical Clustering

Hierarchical clustering methods are then divided in two categories, agglomerative and divisive. Agglomerative hierarchical clustering is a “bottom up” approach. In this method, each object is a different cluster in the beginning. Then one pair of clusters is merged at a time and the method continues until all clusters are merged into one cluster. On the other hand, divisive hierarchical clustering is a top-down approach. In this approach, all objects are initially in one big cluster and then split into smaller clusters until each object forms an individual cluster. A dendrogram is used in order to present the results of hierarchical clustering. Figure 13 illustrates an example of the way the agglomerative and divisive hierarchical clustering work.

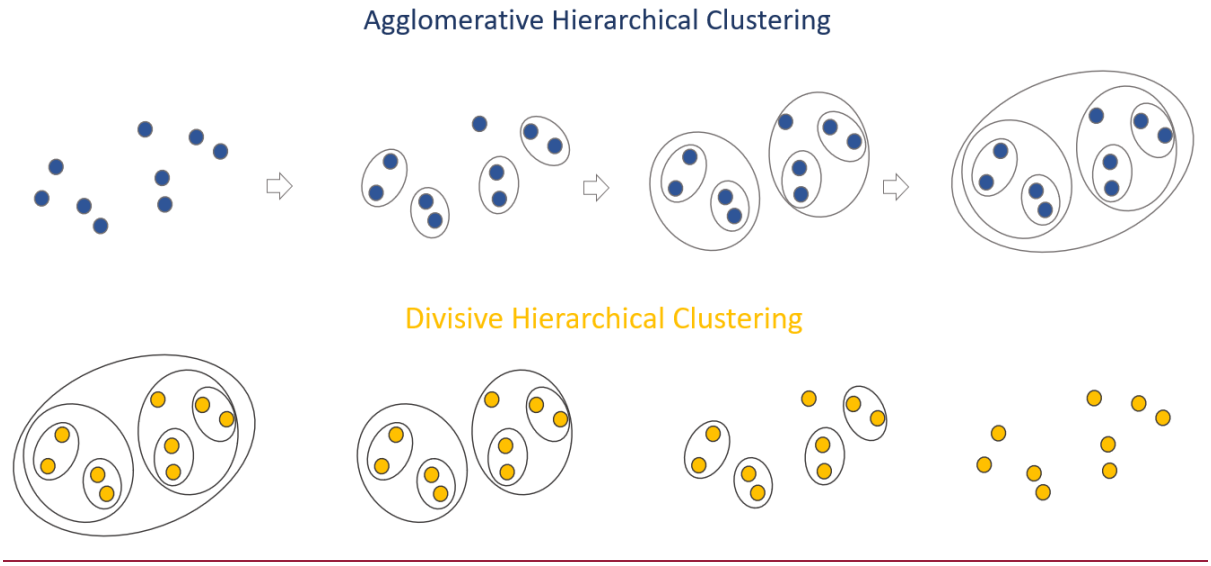


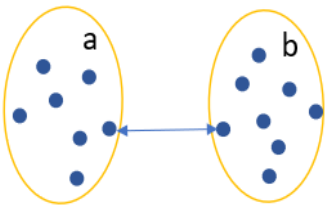
Figure 16: Agglomerative vs Divisive Hierarchical Clustering

Squareform and pdist are two different methods of calculating distances in hierarchical clustering. Squareform is used when all the data points are arranged in a square matrix and is more efficient in hierarchical clustering as it uses a single distance matrix to calculate all the distances between all the points, while pdist requires separate calculation for each pair of points. On the other hand, pdist is more accurate as it takes into account the actual distances between each pair of points.

A.4.1 Linkage Function between clusters

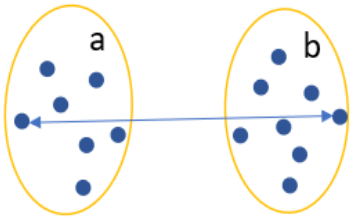
The creation of the hierarchical cluster tree requires a distance metric which calculates the distance between clusters. There are multiple linkage methods, the single, complete, average, centroid and Ward’s linkage method.

Single Linkage



In the single linkage method, the distance between two clusters is defined as the shortest distance between two points in each cluster. In Figure 7, we can see that the distance between two clusters, a and b is the length of the arrow that unites the two clusters' closest points.

$$L(a,b) = min\left(D(x_{ai}, x_{bj})\right) \quad (2.3)$$



Complete Linkage

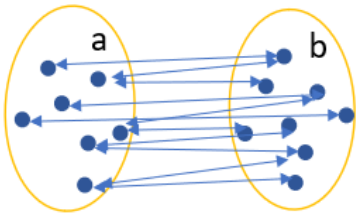
On the other hand, in complete linkage hierarchical clustering, the distance between two clusters a and b is the length of the arrow that connects the two points in the two clusters which are as far away as possible.

$$L(a,b) = max\left(D(x_{ai}, x_{bj})\right)$$

Average Linkage

In average linkage method, the distance between two clusters is defined as the average of all pairwise distances across the two clusters.

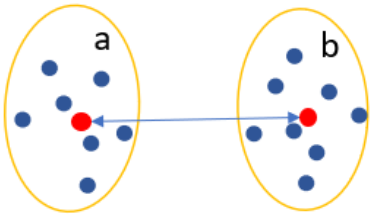
$$L(a,b) = \frac{1}{n_a n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} D(x_{ai}, x_{bj}) \quad (2.5)$$



Centroid Linkage

In the centroid linkage method, the distance between two clusters is the distance between the two mean vectors of the clusters. At each stage, the two clusters with the smallest centroid distance are merged.

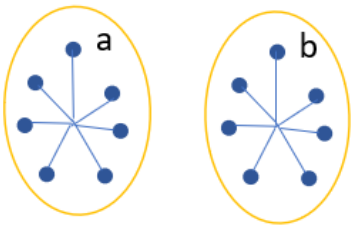
$$L(a,b) = D((\frac{1}{n_a}\sum_{i=1}^{n_a} x_{ai}), (\frac{1}{n_b}\sum_{j=1}^{n_b} x_{bj})) \quad (2.6)$$



Ward's minimum variance method

The Ward's linkage method is the only method that is based on sum-of-squares criterion. At each stage, two clusters merge that provide the smallest increase in the combined error sum of squares.

$$L(a,b) = D(\{x_{ai}\}, \{x_{bj}\}) = \|x_{ai} - x_{bj}\|^2$$



A.4.2 Hierarchical Clustering using SMs

The clustering algorithm starts with each device in its own singleton cluster. The algorithm takes the clusters with the smallest distance between them (from the linkage method) and merges them into a single cluster. Then the distances between this new combined cluster and all the other clusters are recalculated, and then the algorithm repeats until all devices are in one cluster. This algorithm then returns the history of all its merges and what distances were between each of the clusters it merged.

From this history, we can pick a threshold where we ignore all merges that occurred where the distance between the merged clusters was above that threshold. This will give us a variable number of clusters depending on the threshold and the distance matrix. We can also choose a fixed number of clusters, by not

merging any more clusters once it has made the number of clusters that we want. In phase identification, we pick 3 clusters.

A.4.2.1 An example of Hierarchical Clustering

In Table 15, we can observe an example of a correlation matrix between device's voltage streams.

	Device 1	**Device 2**	**Device 3**	**Device 4**	**Device 5**	**Device 6**
Device 1	1	0.2	0.8	-0.1	0.1	0
Device 2	0.2	1	0	0.9	0.8	0.1
Device 3	0.8	0	1	0	0	0.2
Device 4	-0.1	0.9	0	1	0.9	-0.2
Device 5	0.1	0.8	0	0.9	1	0.2
Device 6	0	0.1	0.2	-0.2	0.2	1

Table 15: correlation matrix

From the correlation matrix, the distance matrix is calculated. Using $D_{i,j} = 1 - C_{i,j}$ the distance between devices 1 and 2 is 0.8, between devices 3 and 3 would be 0, between devices 2 and 4 would be 0.1, etc.

Starting the clustering algorithm, each device is in its own cluster. So, device 1 is in cluster 1, device 2 in cluster 2, etc. The algorithm checks the distance between each pair of clusters and picks the minimum. At the start, this linkage method is the same as the distances between the points (because in $d(u, v) = \max\left(\mathrm{dist}(u_i, v_j)\right)$ there is only one point in each cluster, so it is the maximum over only one distance), so it will just merge the clusters of the closest points. It is noted that the example uses the 'complete' linkage method for simplicity.

Iteration 1: The closest clusters, clusters 2 and 4, have distance 0.1. Clusters 4 and 5 also have distance 0.1. We will pick only one pair of clusters to merge, 2 and 4. Let the new cluster containing devices 2 and 4 be called cluster 7.

Iteration 2: Cluster 4 and 2 got merged into cluster 7, and so now we need to check the distances between this cluster and the other clusters. The distance between cluster 5 and cluster 7 is now $\max(\mathrm{dist}(Device\ 5, Device\ 2), \mathrm{dist}(Device\ 5, Device\ 4)) = \max(0.2, 0.1) = 0.2$. After calculating the new distances between the new cluster 2 and the other clusters, the closest clusters are now 7 and 5 (linkage distance 0.2), and 1 and 3 (0.2). We'll randomly pick 2 and 5, so now devices 2, 4, 5 are now in the new cluster 8.

Iteration 3: After recalculating the distances between clusters, clusters 1 and 3 have the smallest distance (0.2) so they will be merged into the new cluster 9.

Iteration 4: The distance between clusters 8 and 9 is 1.1, between 9 and 6 it is 1, and between 8 and 6 it is 1.2. So, the clusters 9 and 6 get merged, and now they become cluster 10 containing 1, 3, 6.

Iteration 5: There are only 2 clusters left, now they merge into a single cluster.

From this history, if we want 3 clusters (e.g., one for each phase), then we will take the state after iteration 3 as our clusters, (1,3), (2,4,5), and (6). This looks like a good clustering because the correlations between the devices within each cluster are high, and the correlations of devices in different clusters are low.

Here is a dendrogram diagram that illustrates the process. Note that the numbers along the bottom are the distances between the clusters when they merged.

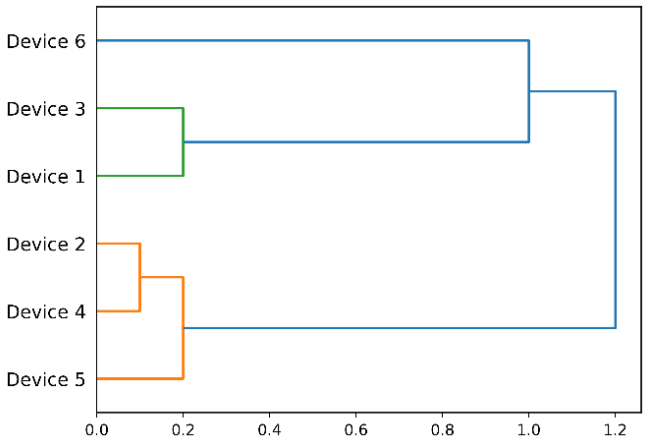


Figure 17: Dendrogram using complete linkage function

Here is the same dendrogram but made with Ward's linkage method instead of complete. Note that the distances at which clusters merge are different.

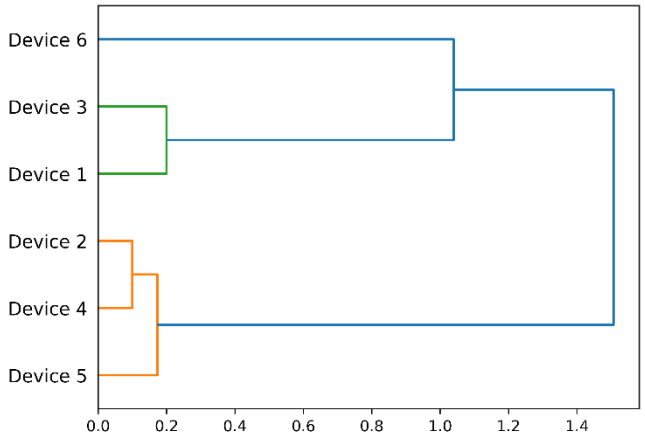


Figure 18: Dendrogram using complete linkage function

A.5 Silhouette Score

Silhouette Coefficient or Silhouette Score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1 means cluster are well apart from each other and clearly distinguished. 0 means clusters are indifferent, or we can say that the distance between clusters is not significant. -1 means clusters are assigned in the wrong way.

The formula for the calculation of the Silhouette Score is:

$$Silhouette\ Score = (b - a) / \max(a, b)$$

Where,

a = average intra-cluster distance i.e., the average distance between each point within a cluster.

b = average inter-cluster distance i.e., the average distance between all clusters.

A.6 Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points.

Accuracy has the following definition:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

A.7 Rand Index & Adjusted Rand Index

Rand Index is a clustering metric that computes a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned to the same or different clusters in the predicted and true clustering.

The formula of the Rand Index is:

$$RI = \frac{Number\ of\ Agreeing\ Pairs}{Number\ of\ Pairs}$$

The RI can take values from 0 (completely dissimilar) to 1 (a perfect match).

The Rand Index score is then adjusted for chance using the formula below:

$$ARI = \frac{(RI\ -\ Expected\ RI)}{Max(RI)\ -\ Expected\ RI}$$

This is the adjusted Rand Index, with 0 having a Rand Index the same as an average random labelling, and 1 when the clusters are identical.

A.8 Fowlkes-Mallows Scores

Like the Rand Index, the Fowlkes Mallows scores measure the correctness of the cluster assignments using pairwise precision and recall. A higher score signifies higher similarity.

Fowlkes-Mallows Index (FMI) is a geometric mean of pairwise precision and recall, using True Positive (TP), False Positive (FP) and False Negative (FN).

Fowlkes-Mallow's score does not take into account True Negative (TN), it will not be affected by chance adjustments, unlike Rand Index.

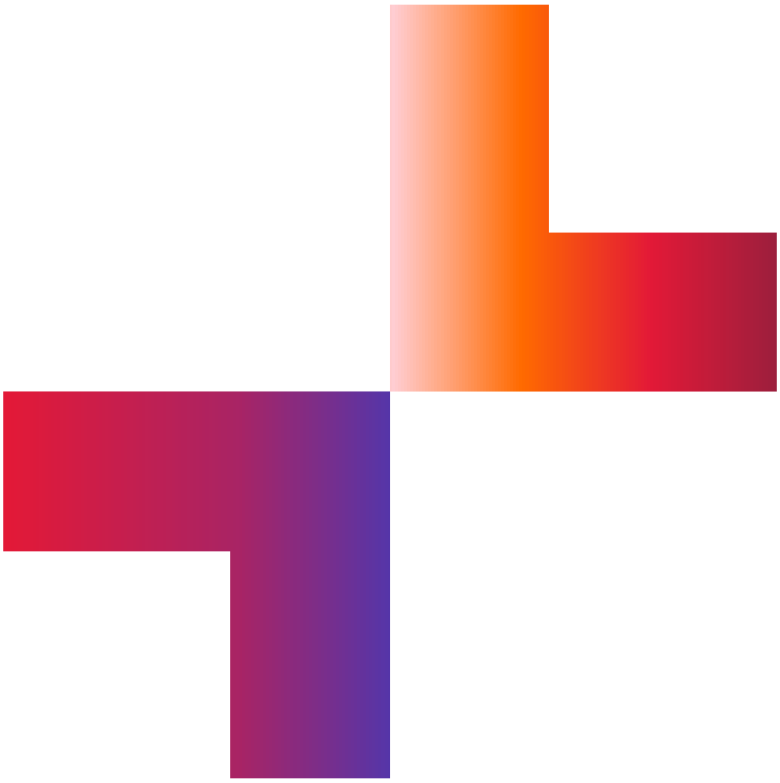
The formula for Fowlkes-Mallows is:

$$FMI = \frac{TP}{\sqrt{(TP + FP\ ) \times (TP + FN)}}$$

A.9 Fisher Z transformation

The Fisher transformation (or Fisher z-transformation) of a Pearson correlation coefficient is its inverse hyperbolic tangent. When the sample correlation coefficient r is near 1 or -1, its distribution is highly skewed, which makes it difficult to estimate confidence intervals and apply tests of significance for the population correlation coefficient ρ. The Fisher transformation solves this problem by yielding a variable whose distribution is approximately normally distributed, with a variance that is stable over different values of r.The formula we can use to transform Pearson's correlation coefficient (r) into a value that can be used to calculate a confidence interval for Pearson's correlation coefficient is introduced below:

$$Z_r = \ln\big((1 + r)/(1 - r)\big)/2$$



cgi.com