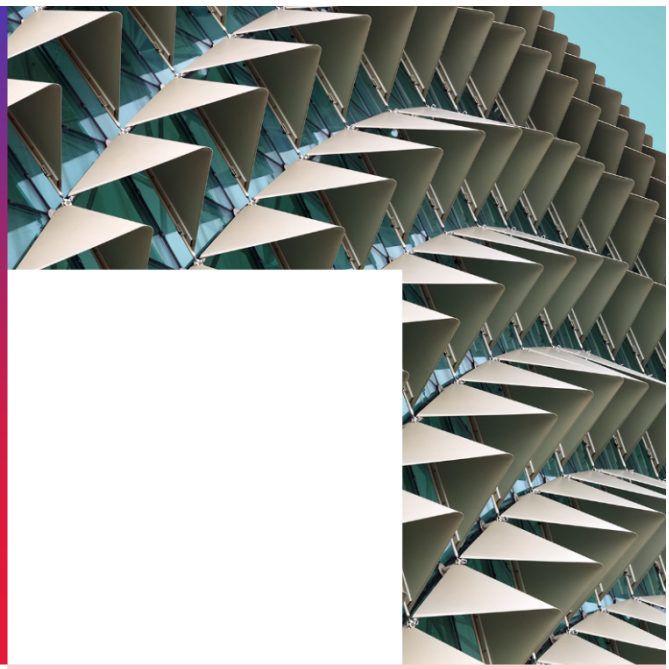


HV Phase Identification Report

HV Pilot

25/08/2026

Prepared by Iro Psarra, Thomas Wilson, George Williamson-Snape

Confidential © 2025 CGI Inc.

Amendment History

Date	Issue	Status
14/12/2025	0.1	Initial version
06/03/2026	0.2	Reviewed version
06/03/2026	1.0	Final version

Confidential

Contents

1 Introduction 5

1.1 Background 5

1.2 Document Purpose 5

1.3 Overview 6

1.4 Abbreviations 6

2 HV Phase Identification 8

2.1 HV Phase Identification Approaches 8

2.1.1 Background Information 8

2.1.2 Approach A: Correlation between SM and LV Monitoring data in HV Feeder Level 9

2.1.3 Approach B: HV Feeder Level Clustering 9

2.1.4 Approach C: Sequential Phase Identification 10

2.1.5 Project Outputs 11

2.2 Data Requirements 12

2.3 Pre-Calculation Validity checks 12

2.3.1 Smart Meter Voltage Data 12

2.3.2 MPAN-Substation Allocation Errors 13

2.3.3 LV Monitoring Data 13

3 Results 15

3.1 Data Preparation 15

3.1.1 Trial Area 15

3.1.2 Data Analysis and Cleaning 15

3.1.3 LV Monitoring Data – Clock Offset 17

3.1.4 Absence of SMs in single-phase transformer 17

3.2 Correlation between SM and LV Monitoring data in HV Feeder Level - Results 18

3.3 HV Feeder Level Clustering – Results 22

3.4 Approach C: Sequential Phase Identification – Results 25

4 Summary of Learning Points 29

5 Appendix 31

A.1 Pearson Correlation 39

A.2 K-means 39

A.3 Hierarchical Clustering 40

A.3.1 Linkage Function between clusters 40
A.4 Silhouette Score 41
A.5 Accuracy 42
A.6 Rand Index & Adjusted Rand Index 42
A.7 Fowlkes-Mallows Scores 43

Confidential

1 Introduction

1.1 Background

The HV Pilot aims to identify the phases to which single-phase distribution transformers and two-wire sections of HV overhead line are connected, enabling more accurate power-flow modelling and supporting future actions to minimise phase unbalance. Phase unbalance is known to increase losses, reduce utilisation of network assets, and affect voltage quality, though the phase connectivity across much of the rural HV network is incomplete or unknown.

At the moment, current modelling techniques 'roll' the phases within 11kV network models, which is effectively the best-case scenario from a network planning perspective. The reality is we are unaware of the phasing, where the HV Pilot scheme will help understand how models are actually connected (on a small trial sample). This will help to drive more accurate network models and ensure the network remains within operational limits.

The Smart Meter voltage data will be utilised, building on the learnings from the SMITN project, which successfully demonstrated the identification of phases in the LV network. In addition, imagery and LiDAR data will be used to build a virtual model of the network, enabling the physical tracing of wire connections. These techniques will be further developed and tested to evaluate their accuracy and cost-effectiveness as alternatives to traditional manual site inspections.

Early assessments of the network data, monitoring data, and smart meter coverage were carried out to determine which feeders were suitable for analysis using smart meter voltage data. In parallel, we conducted a structured review of the helicopter survey imagery and LiDAR data, examining the availability, coverage, and usability of the visual data for machine-vision processing. By bringing together the results of the network assessment and the visual data review, we refined the candidate list and selected a set of seven HV feeders to form the basis of the initial trial. These feeders provide the necessary data inputs for developing and testing both phase-identification approaches.

A dedicated Azure-based datastore integrates the network topology, CROWN data, smart meter inventory and voltage data, LV monitoring, and the aerial imagery. This supports consistent processing, repeatable analysis, and provides the framework for a potential future BAU implementation.

This report focuses on the use of smart meter voltage data to determine HV phase connections and outlines the analytical methods being developed, including clustering and other voltage-pattern analysis techniques applied across the selected feeders. The project therefore aims to demonstrate the viability of using smart meter data to identify HV phase connectivity, providing the validation required to assess accuracy and performance.

1.2 Document Purpose

The purpose of this document is to outline the approach to determining HV phase connections using smart meter voltage data and to describe the analytical methods that will be developed and evaluated as part of the HV Pilot project. Building on the findings and recommendations from the SMITN project, we will assess a range of techniques, including data preparation steps, clustering methods for grouping meters by voltage behaviour, the use of reference distribution substations, and other analytical approaches identified for HV phase identification.

We should note that this is an initial findings report, and additional techniques, refinements, and optimisation steps will be proposed as the project progresses. As we move further into the analysis, the approaches outlined here will be expanded and strengthened to improve robustness and overall performance.

1.3 Overview

Section 2 of this document presents an overview of the HV Pilot 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
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

Term	Definition
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)

Confidential

2 HV Phase Identification

This section provides a summary of the approaches that will be explored in the project, outlining their benefits and potential limitations. It also highlights key learnings from the SMITN NIA project that are relevant and can be leveraged to strengthen the methods developed for HV Phase Identification.

2.1 HV Phase Identification Approaches

2.1.1 Background Information

Single-phase distribution transformers are connected to two of the three conductors on an HV feeder. In some areas, some side HV branches have only two conductors in order to reduce installation costs, meaning that all distribution transformers on these branches are connected to the same phase pair. Across the HV feeder, many single-phase distribution transformers may end up using the same two conductors, contributing to phase-load unbalance. In this project we will apply several HV phase identification approaches to determine the HV phases of single-phase distribution transformers.

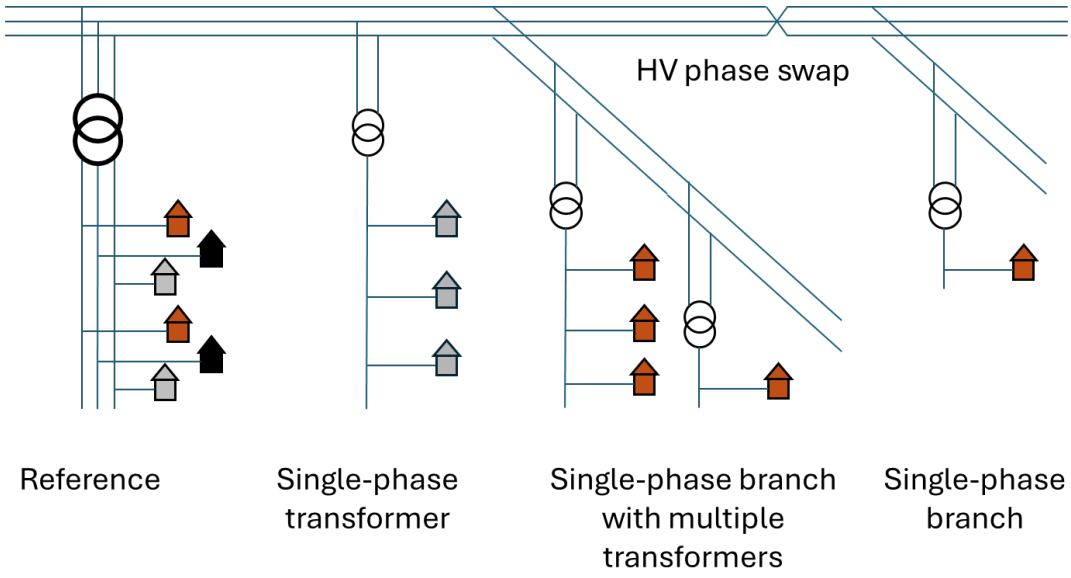


Figure 1: Example HV feeder with single-phase distribution transformers and single phase branches

We are testing the assumption that the HV phase supplying a single-phase transformer can be identified by analysing the voltage patterns recorded by smart meters from the LV side. Specifically, we compare the voltage time-series from customers connected to single-phase distribution transformers with the voltage patterns from customers connected to nearby three-phase distribution transformers (reference transformer). Because each LV line-to-neutral voltage reflects a specific HV line-to-line combination, the correlation between these voltage patterns allows us to infer which pair of HV conductors is feeding each single-phase transformer.

All the measurement data used in this analysis is taken from the LV side of the distribution transformers. Each LV line-to-neutral voltage is derived from a specific HV line-to-line pair through the transformer's delta-star configuration. The typical conductor ordering in HV Feeder is L1 on the right, L2 in the middle, and L3 on the left when facing the Primary. (This was mentioned IIRC when the Primary is at your back (Red, right, rear)). Under this arrangement, a Dy11 distribution transformer produces the following mapping between LV phases and HV line pairs:

- LV phase L1 corresponds to the HV line pair L1-L2

Confidential

- LV phase L2 corresponds to the HV line pair L2-L3
- LV phase L3 corresponds to the HV line pair L3-L1

In the SMITN NIA project, two main approaches were tested for identifying the phase of customers connected to three-phase distribution substations, which serve as the reference distribution transformers in our use case. The first approach used direct voltage correlation between smart meter voltage data and voltages from LV monitoring data from the distribution substations. The second approach applied unsupervised machine-learning techniques - clustering to group smart meters into three phase clusters based on the similarity of their voltage time-series. Multiple clustering techniques were evaluated, including k-means and hierarchical clustering, with the hierarchical approach providing more stable and reliable results. Both methods were tested using different data resolutions, such as 30-minute and 1-minute voltage data, to assess their sensitivity to sampling interval. Across the trial area, both approaches achieved accuracy levels above 90% when compared against HAYSYS validation data, providing a strong foundation for applying similar principles within the HV Pilot project. For more information about the algorithms developed in SMITN please refer to *Phase Identification Report, SMITN.*

Below is a description of the approaches we intend to test for HV phase identification, outlining the methods and the underlying assumptions to determining the HV phase connections of single-phase distribution transformers and two-wire branches.

2.1.2 Approach A: Correlation between SM and LV Monitoring data in HV Feeder Level

The first approach aims to test the assumption that the phase of all smart meters on an HV feeder can be identified using a single reference distribution substation, that belongs to the examined HV feeder. The first approach uses the LV monitoring data as reference. Building on the correlation approach used in SMITN, where customers at a distribution substation were correlated with the substation's monitoring data, we extend this concept to the HV level. Instead of correlating customers within one LV network, we use the LV monitor at the chosen reference distribution substation to represent the entire HV feeder. We then correlate the voltage time-series of every smart meter connected to single-phase distribution transformers along the feeder with the three phase voltages from lv monitoring data measured at this reference point. The expectation is that each SM will correlate strongly with only one of the three reference phases, allowing its HV phase connection to be inferred.

A key requirement of this approach is that each HV feeder included in the analysis must contain at least one distribution substation equipped with LV monitoring.

To validate the results, we will check whether all smart meters connected to the same single-phase distribution transformer correlate strongly with only one of the three reference phases. If this behaviour is observed, it increases our confidence in two important areas:

1. The distribution transformer recorded as single-phase in CROWN is indeed a true single-phase connection, and
2. We can reliably determine its HV phase label by using the LV monitoring data at the reference distribution substation.

2.1.3 Approach B: HV Feeder Level Clustering

This approach considers voltage correlations across the entire HV feeder by calculating the correlation between every pair of smart meters and then applying clustering to the full correlation matrix. All smart meters supplied by the same single-phase transformer, or located on the same two-wire HV branch, should belong to the same

phase cluster. If the distribution transformer is genuinely single-phase, the proportion of smart meters that fall into the same cluster provides a confidence metric for the inferred phase connection.

This method has the advantage that it does not require LV monitoring data, making it applicable to feeders where no monitoring is installed. In the absence of distribution substation monitoring, there is no direct information available to label the identified phase clusters as L1, L2, or L3. Instead, the clusters can be assigned generic labels such as A, B, and C. For assessing phase unbalance, this is not a limitation. The key requirement is to determine which group of customers shares the same phase and whether one phase is more heavily loaded than the others, rather than identifying the absolute phase label.

This approach is more sensitive to data issues. Smart meters that are incorrectly assigned to the HV feeder will show very low correlation to the smart meters and can distort the clustering. Likewise, time offsets, inconsistent timestamping, and poor synchronisation across devices can introduce noise that reduces clustering accuracy. For these reasons, the clustering method requires more extensive preprocessing to ensure reliable phase-identification results.

2.1.4 Approach C: Sequential Phase Identification

2.1.4.1 Phase Identification for three-phase distribution substations (Reference)

Instead of creating a single large clustering across the entire HV feeder, this method builds the phase identification from the bottom up. The first step is to determine the three phase groups at each three-phase distribution substation individually. Using either LV monitoring data (where available) or clustering of smart meters, we identify and label the three phase clusters within each three-phase distribution substation.

2.1.4.2 Phase Identification for single-phase distribution substation

Once the phases of the three-phase distribution transformers have been identified, we then extend the analysis to the single-phase distribution transformers. The smart meters connected to each single-phase distribution transformer are correlated against the voltage profiles from smart meters from the closest reference three-phase distribution substation. Since all smart meters at a single-phase distribution transformer are supplied from the same HV phase pair, in principle only one SM is required for phase identification. However, we will use all available smart meters to confirm that they consistently point to the same phase.

2.1.4.3 Selecting SM to include in the analysis

When correlating a smart meter from a single-phase distribution transformer to the three phases of the reference distribution substation, several options are available. One approach is to use all SMs from the reference distribution substation and compute the average correlation for each phase, selecting the phase with the highest average value. Alternatively, we may select specific smart meters that are expected to provide better voltage reference for example:

- Smart meters that are electrically closer to the reference distribution transformer (lowest accumulated resistance),
- Smart meters that show the highest correlation with LV monitoring data, or with other smart meters in the same phase group,
- or meters without EV or PV technologies, which can introduce local voltage distortions.

As the project progresses, we will evaluate the performance of these approaches and may introduce further refinements or new techniques based on the results observed. For more information on the algorithms we intend to test, please refer to the *HV Pilot WP3 D1 Selected Algorithms Report*.

2.1.5 Project Outputs

Following discussion and feedback from business stakeholders, the project outputs will be refined to directly support operational, planning, and modelling needs.

The project will improve understanding of how customers are actually connected to the available phases of each distribution transformer and will therefore provide insight into potential unbalance in LV load modelling. Moreover, the outputs will help identify distribution transformers that are three-phase or split-phase but where the majority of customers are supplied from only a single phase. This situation is common where a single-phase distribution transformer has been upgraded to three-phase without the LV customer connections being rebalanced. In the case of split-phase distribution transformers, the correlation results will appear similar to those of a single-phase distribution transformer. This is appropriate from an HV unbalance perspective, as both LV phases of a split-phase transformer contribute to one HV phase as if it were a single-phase transformer.

For each transformer analysed, the project outputs will include:

- counts of customers assigned to each phase (L1, L2, L3, L12, L23, L31, L123),
- percentage phase distribution to indicate the level of transformer unbalance (for example, 30% of customers connected to L1, 20% to L2, 30% to L3, and 10% three-phase L123), and
- a tabulated presentation suitable for direct inclusion in NGED's internal documentation.

At the HV feeder level, the outputs will summarise customer phase allocation across each feeder. For every feeder analysed, we report:

- counts of customers assigned to each phase (L1, L2, L3),
- percentage distribution across phases for each feeder, and
- a tabulated presentation suitable for direct inclusion in NGED's internal documentation.

Split-phase distribution transformers present a challenge, as it will be more challenging to resolve the two LV phases (i.e. the phases that are 180° out of phase) using this approach. As a result, care is needed when using these results to represent LV phase unbalance in such cases.

An aggregated analysis across all feeders will also be provided to support business-level decision making. This includes the mean number of customers on each phase (or the total) and the mean demand on each phase (or total) across the surveyed feeders, as well as the median used to summarise the distribution of phase unbalance. In addition, the analysis will identify typical versus atypical balance patterns and provide commentary on how representative the trial feeders are of the wider network.

For outlier identification and confidence assessment, the project will apply a structured approach to detect atypical smart meter behaviour. The project will:

- identify cases where one or a small number of meters disagree with the majority. This could be a disagreement in the phase correlation, or, at a lower level, might be a meter with different voltage characteristics to all neighbours. The disagreement could indicate being connected to a different feeder or distribution substation (a CROWN error) or spurious readings from the meter.
- define the rules and thresholds used to classify and filter outliers, and

- assign a confidence level to the classification on a per-feeder or per-substation basis (for example, high confidence where more than 90% of meters agree).

For operational use, the results are translated into HV phasing, rather than LV labels (L1–L2–L3). This includes a clear and consistent mapping from LV phase labelling to HV phase pairs, alignment with established engineering conventions and expected feeder orientation, and presentation in a form that engineers can readily reconcile.

Furthermore, to support understanding and validation, visual outputs will accompany the results, particularly:

- Colour-coded maps showing phase allocations for each feeder.
- Logical network layers separated by phase for comparison.
- Ability to toggle between layers to identify continuity and check alignment with field data.

Finally, our outputs will be produced in formats suitable for inclusion in both interim and final project reports, feeding directly into modelling and network planning processes. They will also support the development of case studies and clear statements of project outcomes, particularly highlighting the benefits of correcting inaccurate phasing assumptions.

2.2 Data Requirements

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

1. Smart meter voltage data, with 30-minute time resolution.
2. Smart meter voltage data, with 10-minute time resolution.
3. Distribution substation voltage monitoring data, from GridKey and VisNet (10-minute resolution).
4. Premises to distribution substations connectivity data from CROWN.
5. Distribution substation information including its connection to HV feeders, physical configuration (ground- or pole-mounted), geographical details, electrical phase details, and transformer capacity.
6. Electric Office network data.
7. Aggregated LV feeder demand from the SMs.

For more information about the HV Pilot Datastore, please refer to *WP2 – D1 HVPilot Data Architecture and DataStore Overview*.

2.3 Pre-Calculation Validity checks

Experience from the SMITN project highlighted the importance of robust data-preprocessing routines to ensure that voltage time-series are reliable, synchronised, and suitable for use in phase identification methods.

2.3.1 Smart Meter Voltage Data

2.3.1.1 Data Outliers

Smart meter voltage data occasionally contain anomalous values, such as extremely low readings (i.e. 0-30 V) or unrealistically high measurements which typically arise from temporary local faults or data-logging failures. These values can distort distance calculations, reduce the effectiveness of clustering, and negatively impact correlation measures. Because these samples do not accurately represent the HV feeder voltage behaviour, they should be removed from the dataset. In SMITN, unreasonable readings were either corrected where possible or removed. We will apply the same principle in this project, using defined voltage thresholds to screen and clean the incoming datasets.

2.3.1.2 Clock Offsets

Some smart meters occasionally reported timestamps that were slightly ahead of or behind the true measurement time. Experience from SMITN showed that these offsets were more noticeable in lower-resolution data (e.g., 1-minute readings), and even small misalignments can affect the accuracy of correlation between meters. To address this, we will implement an offset-correction method that compares each meter's voltage trace with the average trace of neighbouring meters and shifts the series until the maximum correlation is achieved. In this project we will use HH and 10-minute data, so the impact of such offsets is expected to be lower; however, even at this resolution, misalignment can still influence the results and therefore requires correction.

2.3.1.3 Measurement Interval Synchronisation.

HH smart meter readings can begin at arbitrary clock offsets such as 10 or 20 minutes past the hour while higher-resolution readings may start partway through a minute. Because our phase-identification methods rely on comparing voltages across aligned time intervals, even small misalignments can affect the analysis. To address this, we will use one of two synchronisation approaches: rounding timestamps to the nearest interval boundary or applying linear interpolation to align the data points.

2.3.2 MPAN-Substation Allocation Errors

Customer records in CROWN do not always record the correct distribution substation or LV feeder for each customer. In many cases these allocation errors occur within the same distribution substation, for example, a customer may be assigned to the wrong LV feeder, or they may in practice be connected to a different distribution substation. Our analysis shows that even when a smart meter is incorrectly mapped, if it is ultimately fed from the same HV feeder, its voltage pattern remains similar to the others and the correlation remains high. This observation was first identified during the SMITN project and was the key insight that led to the development of the smart meter approach.

Although the correlation for these smart meters can be high and their individual phase can be identified reliably, problems arise when a smart meter is incorrectly associated with the wrong single-phase transformer. In such cases, the presence of a meter from another phase can cause a genuinely single-phase distribution transformer to appear as if it is supplying multiple phases, leading to it being misclassified as three-phase. For this reason, these meters need to be identified and excluded.

Moreover, when a meter is mistakenly associated with a distribution substation on a different HV feeder, its voltage behaviour is driven by a completely different upstream network. In these cases, the correlation with the rest of the meters is very low. For the clustering approaches, it is essential to remove such meters, as they behave as outliers and can negatively impact the clustering results.

In SMITN, several algorithms were developed to detect and correct LV feeder and substation misallocations, and we will apply similar principles in this project. Our main intention is to exclude from the clustering any meters that are not connected to the same HV feeder. To do this, we will calculate correlation matrix between all devices in a distribution substation and remove the meters whose average correlation with the group is low.

2.3.3 LV Monitoring Data

2.3.3.1 Data Outliers

LV monitoring data can also contain anomalous voltage readings caused by device or logging faults. These outliers can skew the results of our approaches and reduce the reliability of the phase-identification analysis. As

part of this project, we will investigate such anomalies in the LV monitoring datasets and remove or correct any values that are not representative

2.3.3.2 Clock Offsets

Different monitoring device manufacturers may timestamp or average their 10-minute intervals in different ways, for example, reporting the average at the start, midpoint, or end of the window. These differences can create timing misalignments between the LV monitoring data and the smart meter voltage data, even when the underlying voltage behaviour is the same. We will therefore investigate how each device applies its averaging and logging process and develop an alignment logic to synchronise the two data sources before carrying out the phase identification analysis.

3 Results

3.1 Data Preparation

3.1.1 Trial Area

The initial trial area consists of seven HV feeders, comprising 970 distribution substations and serving approximately 8,692 customers. Of these customers, 53% had an installed smart meter as of the end of August 2025. Half-hourly voltage data was first requested for the period 1 July to 31 August 2025, and 3,735 devices responded, giving a success rate of 81%. A further request was then issued to all devices to change the Average Voltage Measurement Period from the default 30 minutes to 10 minutes. Our analysis shows that 2,412 devices successfully reconfigured to the new interval and were able to return 10-minute voltage data for the period from the beginning of September to 31 October 2025.

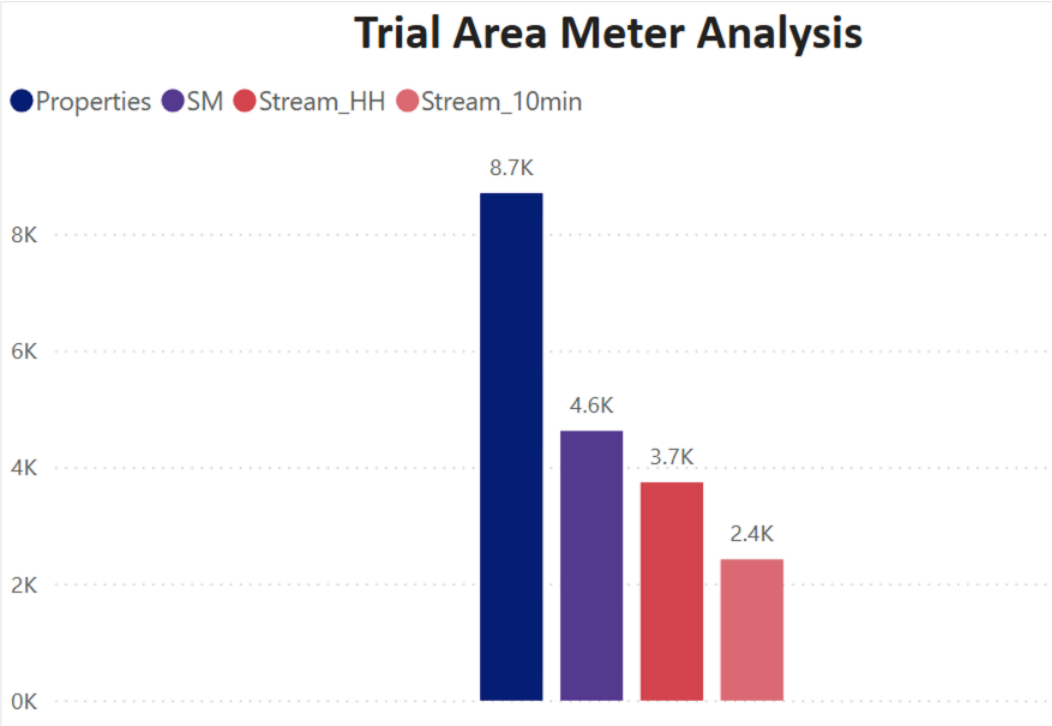


Figure 2: Trial Area Smart meter coverage

3.1.2 Data Analysis and Cleaning

3.1.2.1 Voltage Data Outliers

The scatter plot below shows the minimum and maximum recorded voltage values for each device across both the latest 30-minute and 10-minute datasets. The majority of devices fall within the expected LV operating range, typically recording voltages between 200-250 V.

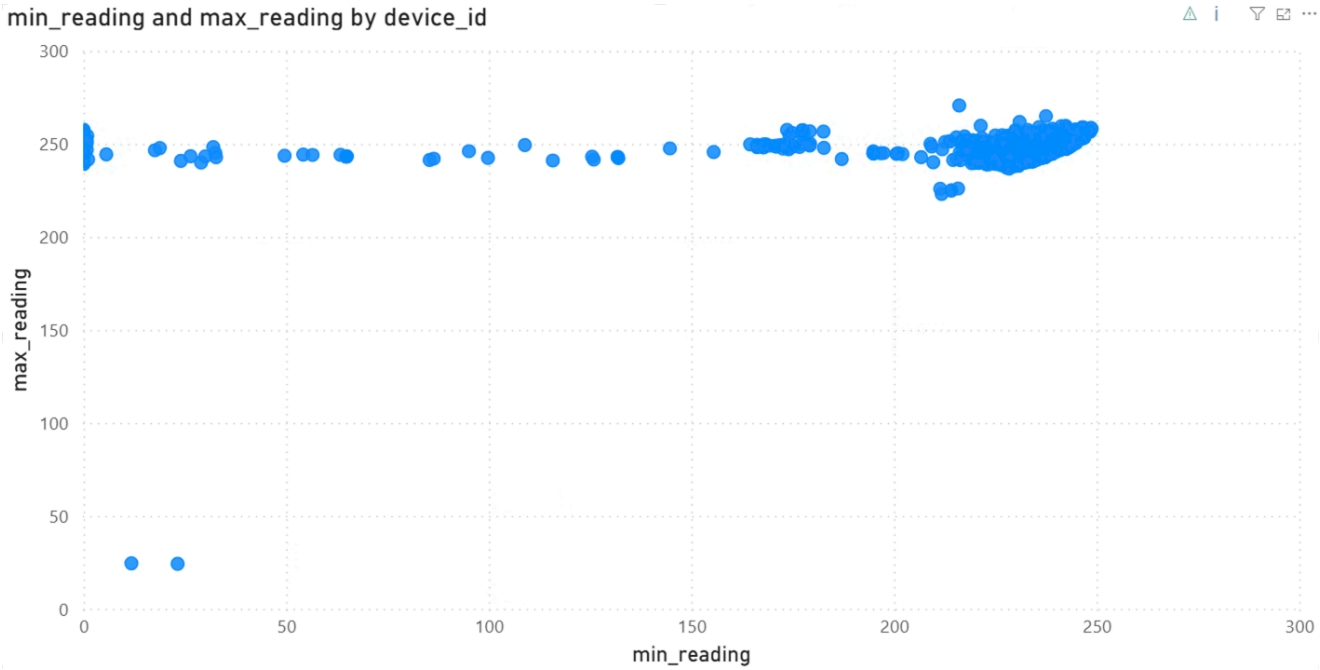


Figure 3: Minimum and maximum voltage values per smart meter

However, some unexpected events cause voltage readings to drop below 200 V. Results from the SMITN project showed that the analytical approaches used for phase identification are sensitive to outliers, and we want to avoid skewing the results due to the behaviour of individual devices. For this reason, we will remove all data points with voltage readings below 200 V.

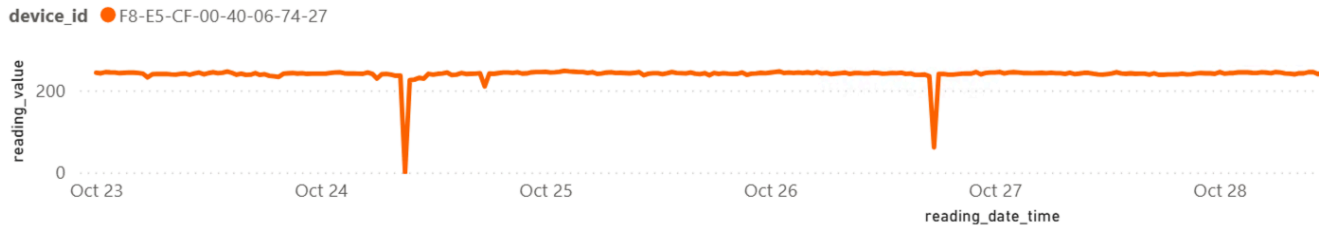


Figure 4: Example voltage outliers below 200 V identified in smart meter data

A small number of devices reported average voltages below 200 V with some showing consistently low values below 30 V. An example is shown in Figure 5. We suspect that these devices are affected by a decimal-point error in their voltage measurement and therefore applied a correction by multiplying the reported values by ten, as illustrated in the second image. As part of the analysis, we will test whether the phase-identification techniques used in this project confirm that this correction results in accurate and reliable voltage data.



Figure 5: Example of smart meter voltage with suspected decimal-point errors

Confidential

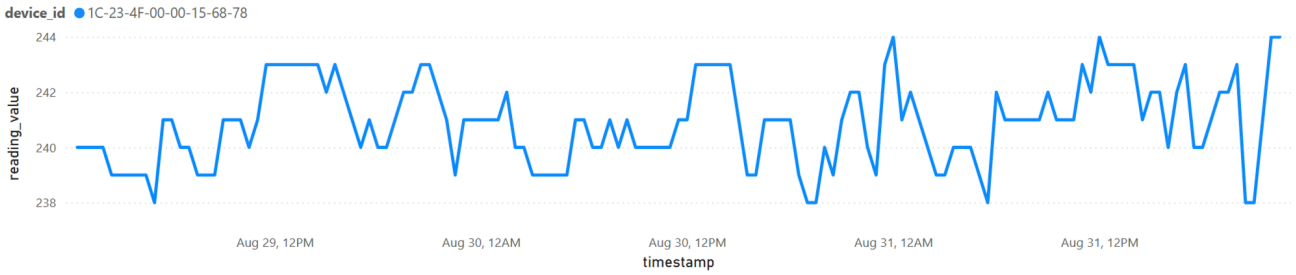


Figure 6: Example of corrected smart meter voltage behaviour after scaling

3.1.3 LV Monitoring Data – Clock Offset

The two monitors shown in Figure 7 are installed on the same HV feeder (770003/0009), and their voltage profiles should therefore follow similar pattern. However, the traces exhibit a noticeable time misalignment, suggesting a synchronisation issue between the devices. The monitors are from different manufacturers (776718 is VisNet, 776899 is GridKey), and they apply different methods for calculating and timestamping the 10-minute average voltage. From previous analysis at Parc Eirin, where both a VisNet and a GridKey monitor were installed at the same distribution substation, it was found that VisNet timestamps correspond to the start of the 10-minute averaging interval, while GridKey timestamps correspond to the end of the interval.

Understanding this offset is important, particularly when correlating distribution substation monitoring data with smart meter voltage measurements, where consistent timing is essential for accurate phase identification. At present, the LV monitoring data for the September–October period, when the smart meters were recording voltage at 10-minute intervals, is not yet available. Once these data are received, we will align the smart meter voltage data with the LV monitoring data and implement an offset-correction algorithm to account for the different timestamping conventions used by VisNet and GridKey devices.

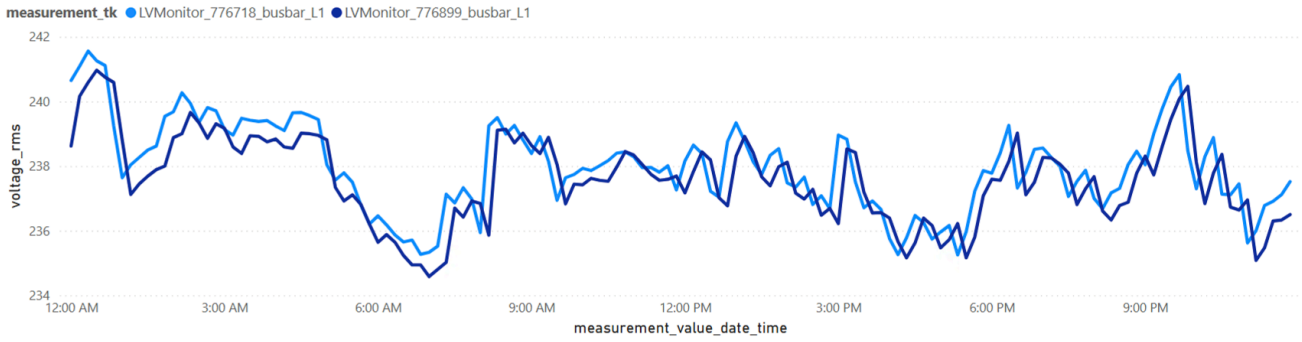


Figure 7: LV monitoring voltage time-series illustrating clock offset between devices

3.1.4 Absence of SMs in single-phase transformer

One major limitation of the approaches discussed, is the absence of smart meters on some single-phase distribution transformers. Although a single smart meter is sufficient to identify the phase of a transformer, if no smart meter is present then the phase cannot be determined using these methods. In the trial area, 574 out of 970 distribution substations have at least one smart meter installed, meaning that phase identification is currently possible for around 60% of the network. When considering the 10-minute data, this figure reduces to 469 distribution substations, or 48%.

We can identify the phase of single-phase distribution substations that do not have any smart meters if they are connected to the same single-phase branch on the HV feeder.

An example is shown in the image below. Substation 413361 does not have any SMs associated with it, so its phase cannot be identified directly. However, it is connected to the same two-wire branch as substation 412719, which has been identified as L1. Because all distribution transformers on a single-phase HV branch share the same HV conductor pair, substation 413361 can be confidently assigned to phase L1 as well.

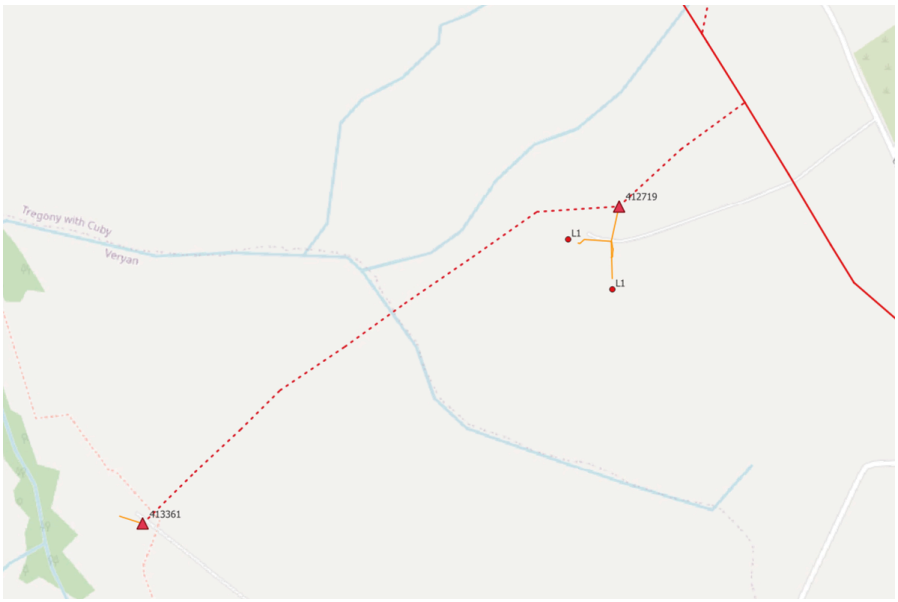


Figure 8: Example of phase inference for a distribution substation without smart meters using a shared single-phase HV branch

For this approach to work accurately, we need confidence that the Electric Office (EO) data correctly records the number of phases on each HV line, and that the sections shown as single-phase branches (such as the interrupted red line in the Figure 8) are indeed true two-wire sections. In the WP1 visual data assessment, we identified several discrepancies between EO records and what is visible in the aerial imagery. The Machine Vision solution developed in the project can help verify whether these spans are genuinely single-phase branches, providing an additional layer of validation before inferring phases for distribution substations without SMs.

As SM adoption increases, this limitation is likely to become less significant. Even so, this approach remains useful, as it provides an additional way to validate and strengthen the results across the network.

3.2 Correlation between SM and LV Monitoring data in HV Feeder Level - Results

This approach aims to determine the HV phase of single-phase distribution transformers by correlating their smart meter voltage data with the three phase voltages measured at a monitored three-phase reference distribution substation. The phase that produces the highest correlation is taken to be the phase supplying that transformer. For more information please refer to Section 2.1.2.

When multiple smart meters are connected to the same single-phase transformer, the consistency of their correlation results provides a strong indicator of the reliability of the identified HV phase. An example can be found in Figure 9. On the left, we can observe that all the smart meters connected to 762803 are all correlated with one phase, L1. Also, we can observe that the substation 760965 is connected on the same single-phase branch with the 762803, and all of the SM has been connected to L1 as well. On the right, the aerial image provides physical confirmation of this result. The single-phase branch is connected to the right and middle conductors of the three-wire HV line, which corresponds to phase L1.

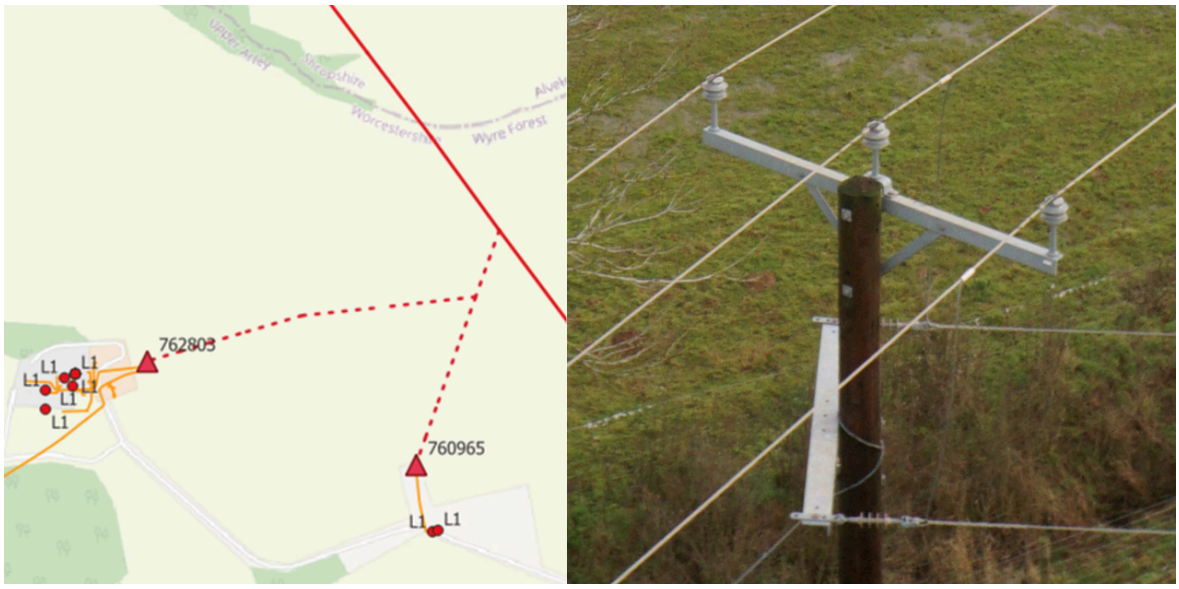


Figure 9: Correlation-based phase identification validated by aerial imagery for a single-phase branch

To quantify this, the following statistics can be used.

For each transformer:

$$\text{Phase Distribution Ratio (PDR)} = \frac{\text{Number of SMs correlating with the same phase}}{\text{Total number of SMs at the transformer}}$$

If the Phase Distribution Ratio (PDR) is distributed across all three phases, the distribution transformer is classified as three-phase (3PH). If more than 90% of the PDR is concentrated on a single phase, the distribution transformer is classified as single-phase (1PH) and assigned to that phase. An example of the classification results is shown in the table 1. All results from Approach A are provided in Appendix A.1.

HV_Feeder	substation	No SM L1	No SM L2	No SM L3	PDR_L1 (L1-L2) %	PDR_L2 (L2-L3) %	PDR_L3 (L3-L1) %	Approach Phase Classification	No of SM	Total No of Customers	CROWN phase info
417560_0041	410024	8	5	6	42	26.32	31.58	3PH	19	51	3PH
417560_0041	410293	0	0	3	0	0	100	L3	3	7	1PH
417560_0041	410743	0	2	0	0	100	0	L2	2	4	1PH
417560_0041	410816	2	0	0	100	0	0	L1	2	7	1PH
417560_0041	410820	2	0	0	100	0	0	L1	2	5	1PH
417560_0041	410955	2	0	0	100	0	0	L1	2	5	1PH
417560_0041	411100	4	0	0	100	0	0	L1	4	7	1PH
417560_0041	411116	2	0	0	100	0	0	L1	2	4	1PH
417560_0041	411120	0	0	2	0	0	100	L3	2	5	1PH
417560_0041	411653	22	16	12	44	32	24	3PH	50	97	3PH
417560_0041	411766	16	14	18	33	29.17	37.5	3PH	48	119	3PH
417560_0041	412121	0	4	3	0	57.14	42.86	3PH	7	10	3PH
417560_0041	412570	12	16	15	28	37.21	34.88	3PH	43	85	3PH
417560_0041	412639	1	0	15	6	0	93.75	L3	16	34	1PH
417560_0041	412719	2	0	0	100	0	0	L1	2	3	1PH
417560_0041	413223	0	2	0	0	100	0	L2	2	6	2PH
417560_0041	413232	27	7	11	60	15.56	24.44	3PH	45	96	3PH

417560_0041	413234	3	2	7	25	16.67	58.33	3PH	12	30	3PH
417560_0041	410024	8	5	6	42	26.32	31.58	3PH	19	51	3PH
417560_0041	410293	0	0	3	0	0	100	L3	3	7	1PH
417560_0041	410743	0	2	0	0	100	0	L2	2	4	1PH

Table 1: Example of Phase Distribution Ratio (PDR) and phase classification results per distribution substation

In some cases, only a single smart meter is available. Where CROWN indicates that the distribution substation is single-phase, this single meter is sufficient to identify the phase of the substation.

The table below compares the distribution transformer classification from CROWN with the classification produced by our approach. Our proposed approach agrees with CROWN for approximately 88% of substations.

CROWN Phasing	Approach Classification	No of Substations
3PH	3PH	84
1PH	1PH	85
1PH	3PH	17
3PH	1PH	5

Table 2: Comparison of CROWN phase classification and Approach A results

For these statistics, we excluded distribution substations with only a single smart meter, as it is not possible to confirm whether such cases are truly single-phase or three-phase. We also note that where only a small number of smart meters are available, it remains the possibility that both meters are connected to the same phase even if the distribution substation is actually three-phase.

Some of the observed discrepancies can be explained by changes in network configuration over time. An example is shown in the graphs and images below for distribution transformer 606008. In CROWN, this transformer is a three-phase transformer. However, the voltage time-series for the associated smart meters follow the same trend, indicating that they are supplied from the same phase. This behaviour is supported by the available LiDAR imagery, which was captured on 10/01/2024 and shows the distribution transformer connected to a single-phase HV branch.

Further investigation, including a review of Google Maps imagery and GIS records, indicates that the line is now three-wire, with an upgrade recorded in 2024. This suggests that the network configuration may have changed since the LiDAR survey was undertaken. However, despite the upgrade to a three-phase configuration, the associated smart meter voltage time-series still follow the same trend, suggesting that all customers remain connected to a single phase. This is likely because the LV connections have not yet been rebalanced following the upgrade. It should also be noted that phase rebalancing may not necessarily occur following an upgrade. In some cases, existing distribution transformers may remain connected as originally configured, with an additional third wire added downstream to enable a three-phase connection. New connections may then be allocated to the new phase.

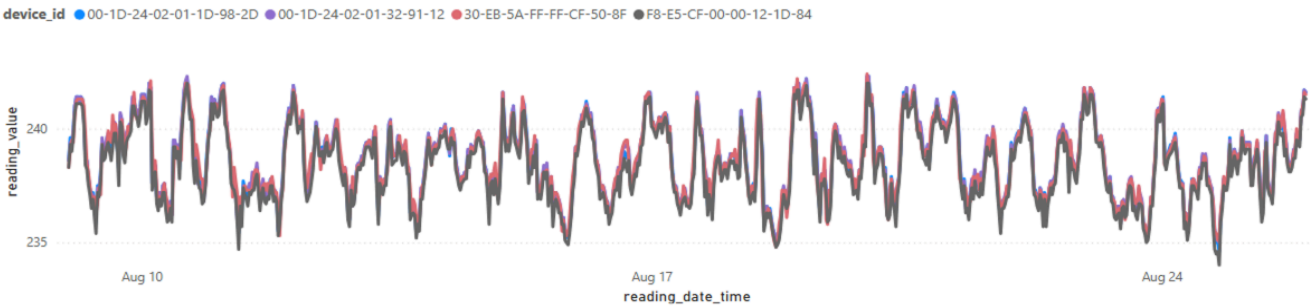


Figure 10: a) Example voltage time-series for distribution transformer 606008 indicating single-phase behaviour.



Figure 11: LiDAR imagery confirming single-phase HV branch connection for distribution transformer 606008

The table below summarises the phase-identification results per HV feeder, showing the number of distribution substations and associated smart meters classified as three-phase (3PH) or assigned to each single phase (L1, L2, L3). This provides an overview of the phase unbalance across each feeder and highlights how distribution substations and customer load are distributed between phases.

HV_Feeder	Phase Results	No of substations	No of SM	Total No of customers
417560_0041	3PH	23	556	1461
417560_0041	L1 (L1-L2)	22	34	
417560_0041	L2 (L2-L3)	4	6	
417560_0041	L3 (L3-L1)	10	15	
600017_0031	3PH	12	193	803
600017_0031	L1 (L1-L2)	5	5	
600017_0031	L2 (L2-L3)	29	47	
600017_0031	L3 (L3-L1)	29	97	
600017_0034	3PH	12	252	1030
600017_0034	L1 (L1-L2)	10	16	
600017_0034	L2 (L2-L3)	5	10	
600017_0034	L3 (L3-L1)	63	107	
770003_0001	3PH	15	371	972
770003_0001	L1 (L1-L2)	8	20	
770003_0001	L2 (L2-L3)	4	6	

770003_0001	L3 (L3-L1)	11	19	
770003_0009	3PH	41	1008	2622
770003_0009	L1 (L1-L2)	33	66	
770003_0009	L2 (L2-L3)	8	11	
770003_0009	L3 (L3-L1)	28	59	

Table 3: Summary of phase-identification results per HV feeder

These results are based on half-hourly (HH) smart meter data, as we have not yet received the 10-minute LV monitoring data for the September - October period, during which the smart meters were configured to record 10-minute voltage data. We therefore expect the results to improve further once the corresponding 10-minute LV monitoring data becomes available.

Two key conclusions can be drawn from this analysis:

1. Within the single-phase distribution substations, there is clear evidence of phase unbalance, with a large proportion of customers often connected to the same phase.

2. The number of smart meters connected to single-phase distribution substations is significantly lower than the number connected to three-phase distribution substations. As a result, if the three-phase distribution substations on a feeder are relatively well balanced, the overall feeder-level phase unbalance may remain moderate.

3.3 HV Feeder Level Clustering – Results

This approach identifies the phases by analysing voltage correlations across the entire HV feeder and clusters all smart meters into 3 phase groups. Smart meters supplied by the same single-phase distribution transformer or located on the same two-wire HV branch are expected to fall within the same cluster. For a detailed description of this method, see Section 2.1.3.

HV feeder level clustering is sensitive to outliers. Smart meters that are incorrectly allocated to the feeder, or affected by timing misalignment can have voltage patterns that are poorly correlated with the rest of the devices on the feeder. In addition, localised effects such as EV charging, PV generation, or data-quality issues can introduce atypical voltage behaviour. When such meters are included in a HV feeder-wide clustering analysis, they can distort the correlation results, reduce the separation between true phase groups, and ultimately degrade the clustering results.

To improve the robustness of the clustering and prevent the results from being skewed by outliers, we applied a filtering step prior to clustering. Specifically, we calculated the average correlation of each smart meter with all other meters on the same HV feeder and excluded from the clustering any devices with an average correlation below 0.4.

Initially, we focused our analysis on HV feeder 417560/0041 to develop, test, and optimise the approach. As the project progresses, we plan to apply the refined methodology to the remaining feeders. We also used the available 10-minute data for this approach. The figure 12 illustrates the behaviour of the HV feeder-level clustering for a three-phase distribution substation. In this example, the smart meters connected to distribution substation 411766 are clearly separated into three distinct clusters, each corresponding to one of the three phases (L1, L2, and L3). The voltage time-series within each cluster are well aligned, showing consistent trends over time, which indicates that the clustering algorithm is performing as expected when the underlying data is coherent and free from major anomalies.

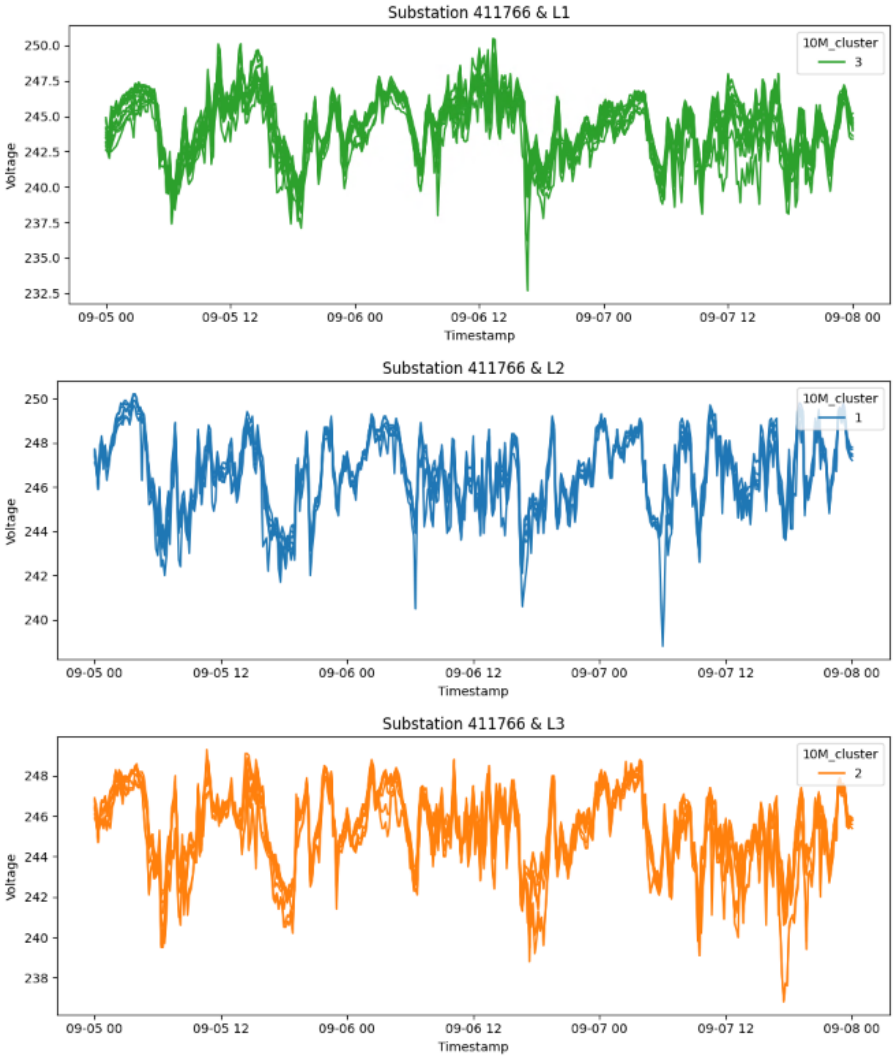


Figure 12: HV feeder-level clustering results for a three-phase distribution substation with well-defined phase groups

The figure below shows an example of a single-phase distribution transformer (412639) where the approach successfully assigns all associated smart meters to the same phase group.

Confidential

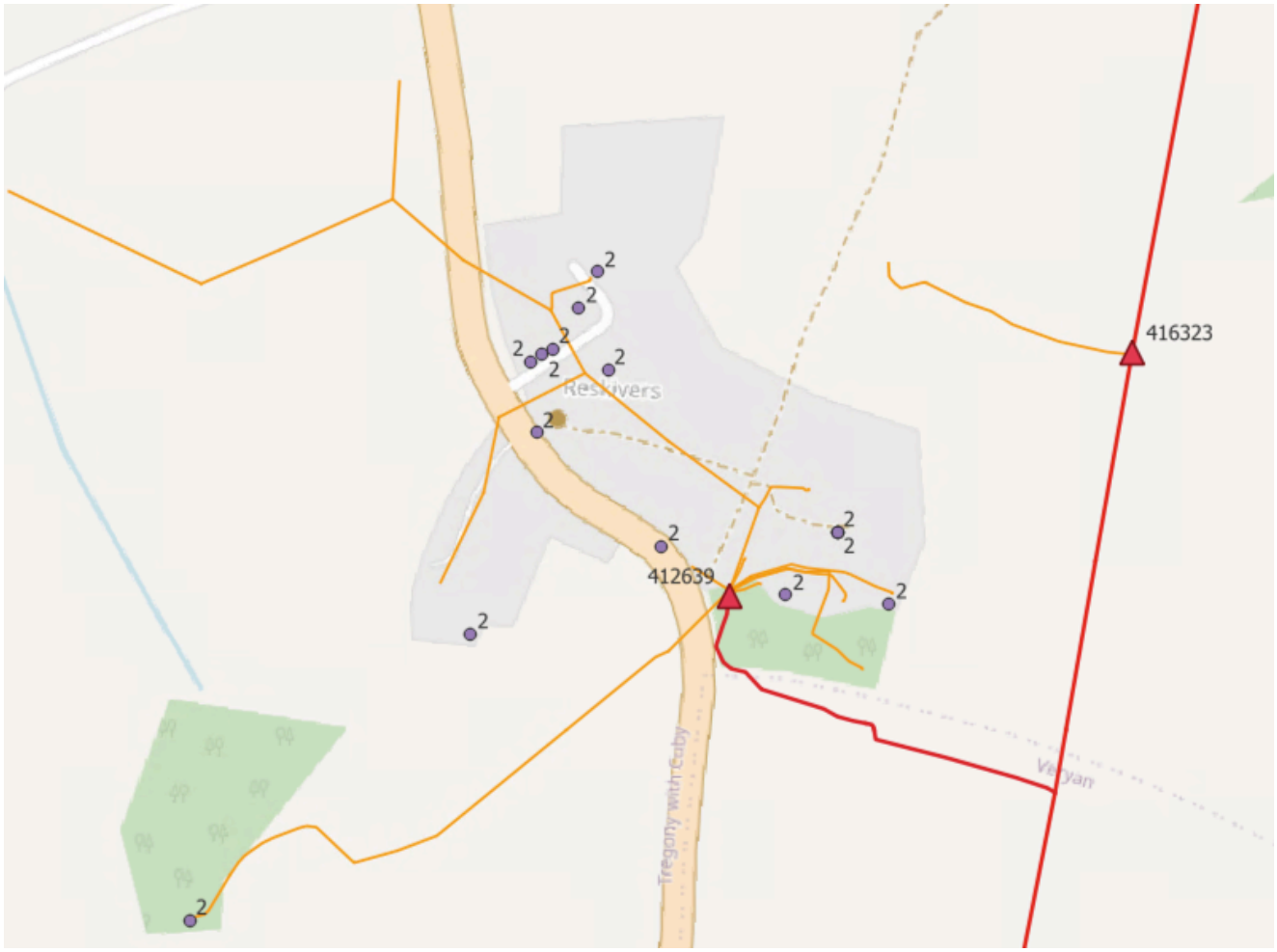


Figure 13: HV feeder-level clustering results for a single-phase distribution transformer showing consistent phase assignment

The table below compares the phase assignment results from Approach A (reference-based correlation using LV monitoring) and Approach B (HV feeder-level clustering) for single-phase distribution substations.

Each row shows how a phase identified by Approach A (L1, L2, or L3) corresponds to a cluster identified by Approach B (Group_1, Group_2, or Group_3), along with the number of distribution substations for which this mapping occurs.

Approach B	Approach A	Count of substations
Group_3	L1 (L1-L2)	13
Group_1	L2 (L2-L3)	2
Group_1	L3 (L3-L1)	3
Group_2	L3 (L3-L1)	4

Table 4: Comparison of phase assignments between Approach A and Approach B for single-phase substations

The results show that phase L1 is very clearly defined, with all distribution substations identified as L1 by Approach A consistently assigned to Group_3 by Approach B. Group_2 contains only distribution substations identified as L3, while Group_1 includes a mixture of distribution substations identified as L2 and L3. This indicates that, although the clustering approach reliably separates L1, the distinction between L2 and L3 is less clear in some cases.

3.4 Approach C: Sequential Phase Identification – Results

Approach C is built upon the correlation and clustering approaches. We utilise the devices seen in the clustering results of a given distribution substation and apply them to the whole feeder as a “ground truth” where each device in the feeder is correlated against the substation devices to find a high average phase correlation. (see Section 2.1.4 for more information).

We started by focusing on the same HV feeder used in Approach B, and first applied clustering at the level of a three-phase distribution substation. For the initial iteration of this approach, we used the available half-hourly (HH) data and we selected substation 411766 to be the reference distribution substation, as its clustering results showed a clear separation between the three phases and aligned well with the phase assignments obtained using Approach A. Also, the clustering at substation 411766 showed a Silhouette Score of 0.3425, which indicates moderate cluster separation. This confirms that devices within each phase group are highly correlated with one another, while devices belonging to different phases are well separated.

We then used the smart meters across the HV feeder and calculated their correlation with all smart meters associated with substation 411766. For each device on the feeder, we computed the average correlation per phase using the phase groupings identified at the reference distribution substation, and assigned the device to

the phase with the highest average correlation.

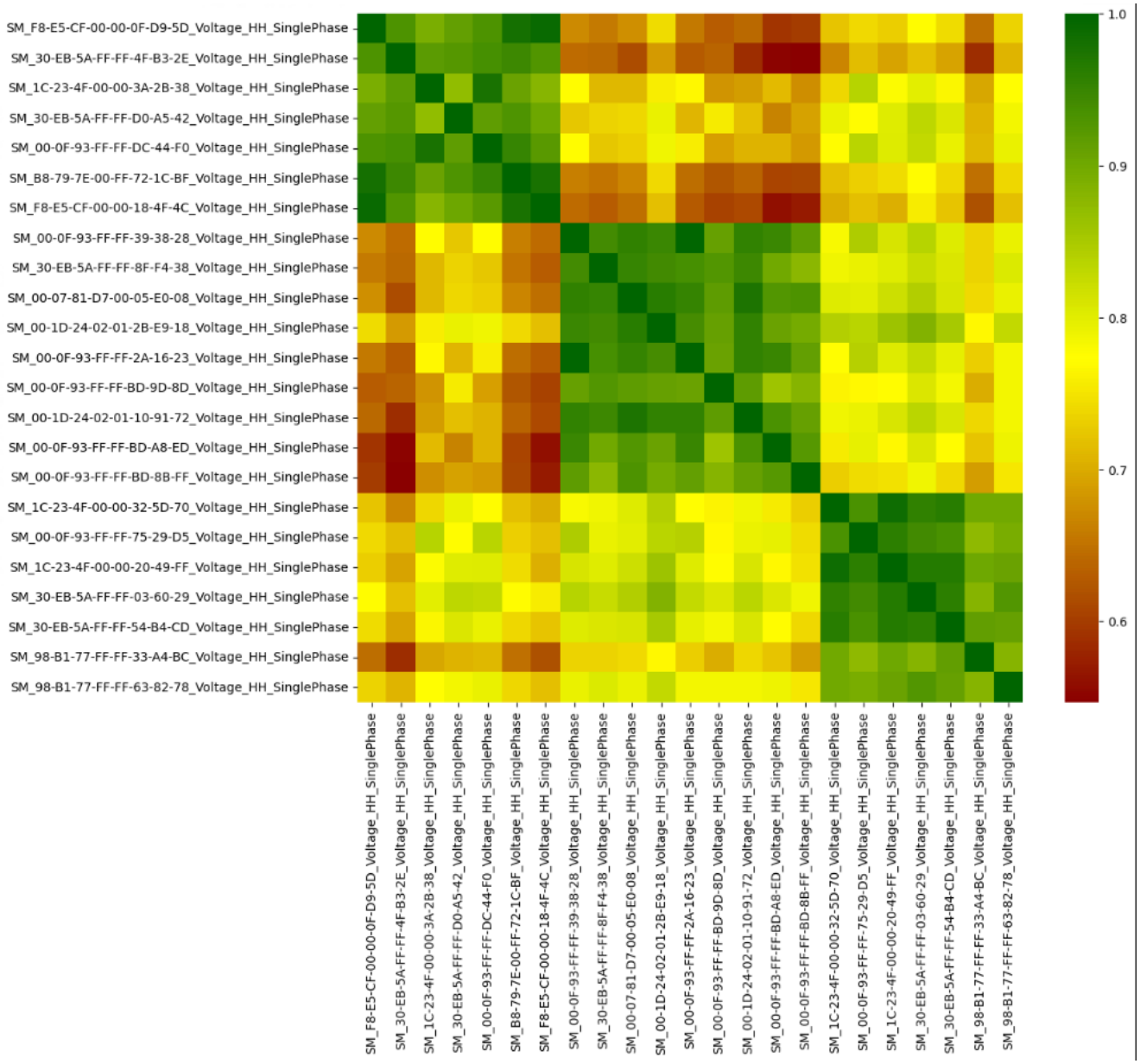


Figure 14: Example smart meter voltage correlation matrix used for hierarchical clustering

The table below summarises the comparison between Approach C and Approach A for single-phase distribution substations. The results show a strong level of consistency between the two approaches. Group 1 in Approach C aligns predominantly with phase L1, with 74% of devices assigned to L1 and 26% of devices aligning with L3 or L2. Group 2 aligns exclusively with phase L2. Group 3 is mainly associated with phase L3, with 73% of devices aligned this way. Additionally, 27% of the devices are aligning with L1 or L2. Overall, the results shows a strong foundation for this approach going forward.

Approach C	Approach A	Count of Devices
Group_1	L1	186
Group_1	L2	49
Group_1	L3	16
Group_2	L2	167
Group_3	L1	14

Confidential

Group_3	L2	38
Group_3	L3	141

Table 5: Comparison of phase assignments between Approach A and Approach C for three-phase distribution substations

We can assess the reliability of Approach C by evaluating its performance on single-phase distribution substations. Table 6 summarises the results of this analysis.

Distribution Transformer Type	Number of phases identified	Count
1PH	1	34
1PH	2	2
1PH	3	1
2PH	1	1
3PH	1	3
3PH	2	3
3PH	3	14

Table 6: Comparison between CROWN classification and Approach C

The findings show over 91% of single-phase distribution transformers were correctly identified as containing only one phase, indicating a high level of consistency. Only a small number of distribution substations (3 in total) were incorrectly identified as being three-phase distribution transformers. These misclassifications may be attributed to several factors, including measurement noise, limited data availability, or issues related to the crown data.

The next table summarises the results for HV feeder 770003/0009 using as reference substation the 776718. Approach A had identified a significant phase unbalance on this feeder. The results from Approach C confirm the same unbalance, showing a high level of agreement between the two approaches.

Distribution Transformer Type	Number of phases identified	Count
1PH	1	61
1PH	2	4
3PH	1	11
3PH	2	11
3PH	3	23

Table 7: Comparison between CROWN classification and Approach C for substation 776718

Approach C	Approach A	Count of Devices
Group_1	L1	15

Group_1	L2	314
Group_2	L1	445
Group_2	L3	7
Group_3	L1	12
Group_3	L2	2
Group_3	L3	349

Table 8: Comparison of phase assignments between Approach A and Approach C for devices in three-phase distribution substations in feeder 770003/0009

As a next step, we will investigate the most effective strategy for selecting reference distribution transformers, for example, choosing reference distribution substations based on geographic proximity, or prioritising those that exhibit the strongest and most stable clustering results. We also plan to repeat this analysis using 10-minute voltage data, which we expect will further improve phase separation and overall confidence in the results. Finally, as the project progresses, longer-duration smart meter voltage data will become available, which are expected to further strengthen the analysis.

4 Summary of Learning Points

The analysis demonstrates that smart meter voltage data can be effectively used to identify HV phase connections for single-phase distribution transformers. Correlation-based methods using LV monitoring data (Approach A) provide reliable and interpretable phase assignments where monitoring is available, while feeder-level clustering (Approach B) and Approach C offers a viable alternative in the absence of monitoring, albeit with greater sensitivity to data quality issues.

The comparison between approaches shows promising and largely consistent results, indicating that the methods are effective in identifying HV phase connections. However, some discrepancies remain and require further investigation. These cases will be used to refine and improve the approaches, which will then be applied across the remaining HV feeders to assess their robustness and scalability.

Within the single-phase distribution substations there is clear evidence of phase unbalance. However, the number of smart meters connected to single-phase distribution substations is much lower than the number connected to three-phase distribution substations. As a result, if the three-phase distribution substations on a feeder are well balanced, the overall phase unbalance across the HV feeder may not be too severe. This assumption will need to be examined further as the project progresses.

Data quality issues, such as voltage outliers, clock synchronisation errors, timestamp misalignment can significantly degrade phase-identification performance, particularly for clustering-based approaches. When clustering algorithms are constrained to form exactly three groups, the presence of outliers can distort cluster boundaries and lead to incorrect phase assignments. Our analysis shows that removing outliers at an early stage, for example by filtering devices with low average correlation to the rest of the feeder, substantially improves clustering stability. In addition to outlier removal, future work could explore increasing the number of clusters initially, allowing anomalous devices to form separate clusters that can be excluded or analysed independently.

Another key learning is that a significant proportion of single-phase distribution substations do not currently have any smart meters installed, which means their phases cannot be identified directly using our proposed methods. To address this gap, we can utilise Electric Office (EO) HV OHL data to identify distribution substations connected along the same single-phase HV branches and infer their phase from neighbouring distribution substations with known phase assignments. This approach assumes that the EO records correctly represent the number of conductors. To increase confidence in these inferences, the Machine Vision approach can be used to validate whether the identified branches are genuinely single-phase.

Next Steps

The next phase of the project will focus on refining the existing approaches, extending them to additional HV feeders, and strengthening confidence metrics and result presentation.

- We will apply 10-minute smart meter voltage data in Approach A and assess whether the phase-identification results improve or change.
- We will further refine Approach B by experimenting with a higher number of clusters to better isolate outliers before consolidating results into the three physical phases.
- We will enhance Approach C by moving to 10-minute data, exploring different choices of reference substations (for example based on proximity), and refining the selection of smart meters used for correlation.
- We will extend Approaches B and C across the remaining HV feeders to test scalability and consistency.

- We will develop improved confidence metrics and refine result visualisations based on stakeholder feedback to ensure outputs are clear and actionable.

Confidential

5 Appendix

A.1 Approach A – Results

HV_Feeder	substation	L1	L2	L3	PDR-L1 (L1-L2) %	PDR-L2 (L2-L3) %	PDR-L3 (L3-L1) %	Approach Phase Classification	no of SM	No of Customers	CROWN phase
417560_0041	410024	8	5	6	42.11	26.32	31.58	3PH	19	51	3PH
417560_0041	410144	0	3	1	0	75	25	3PH	4	7	3PH
417560_0041	410293	0	0	3	0	0	100	L3	3	7	1PH
417560_0041	410743	0	2	0	0	100	0	L2	2	4	1PH
417560_0041	410771	0	0	1	0	0	100	L3	1	3	1PH
417560_0041	410794	4	0	4	50	0	50	3PH	8	13	1PH
417560_0041	410816	2	0	0	100	0	0	L1	2	7	1PH
417560_0041	410820	2	0	0	100	0	0	L1	2	5	1PH
417560_0041	410955	2	0	0	100	0	0	L1	2	5	1PH
417560_0041	411000	1	0	0	100	0	0	L1	1	3	1PH
417560_0041	411004	1	0	0	100	0	0	L1	1	2	1PH
417560_0041	411065	0	1	0	0	100	0	L2	1	4	1PH
417560_0041	411099	1	0	0	100	0	0	L1	1	13	1PH
417560_0041	411100	4	0	0	100	0	0	L1	4	7	1PH
417560_0041	411116	2	0	0	100	0	0	L1	2	4	1PH
417560_0041	411120	0	0	2	0	0	100	L3	2	5	1PH
417560_0041	411256	0	0	1	0	0	100	L3	1	3	1PH
417560_0041	411307	1	0	0	100	0	0	L1	1	3	1PH
417560_0041	411384	1	0	0	100	0	0	L1	1	6	1PH
417560_0041	411559	1	0	0	100	0	0	L1	1	6	3PH
417560_0041	411600	1	0	0	100	0	0	L1	1	4	1PH
417560_0041	411653	22	16	12	44	32	24	3PH	50	97	3PH
417560_0041	411766	16	14	18	33.33	29.17	37.5	3PH	48	119	3PH
417560_0041	411852	0	0	1	0	0	100	L3	1	1	1PH
417560_0041	412121	0	4	3	0	57.14	42.86	3PH	7	10	3PH
417560_0041	412264	1	0	0	100	0	0	L1	1	3	1PH
417560_0041	412303	1	0	0	100	0	0	L1	1	1	1PH
417560_0041	412354	1	1	0	50	50	0	3PH	2	3	3PH
417560_0041	412412	1	0	0	100	0	0	L1	1	2	1PH
417560_0041	412570	12	16	15	27.91	37.21	34.88	3PH	43	85	3PH
417560_0041	412639	1	0	15	6.25	0	93.75	L3	16	34	1PH
417560_0041	412719	2	0	0	100	0	0	L1	2	3	1PH
417560_0041	412774	1	0	1	50	0	50	3PH	2	36	1PH
417560_0041	413142	1	0	0	100	0	0	L1	1	1	1PH
417560_0041	413223	0	2	0	0	100	0	L2	2	6	2PH
417560_0041	413232	27	7	11	60	15.56	24.44	3PH	45	96	3PH
417560_0041	413234	3	2	7	25	16.67	58.33	3PH	12	30	3PH
417560_0041	413238	0	1	0	0	100	0	L2	1	8	
417560_0041	413428	1	0	4	20	0	80	3PH	5	6	1PH
417560_0041	413473	0	0	2	0	0	100	L3	2	4	1PH

Confidential

417560_0041	413584	21	15	13	42.86	30.61	26.53	3PH	49	107	3PH
417560_0041	413609	4	0	0	100	0	0	L1	4	10	1PH
417560_0041	413610	0	0	2	0	0	100	L3	2	3	1PH
417560_0041	413793	15	7	3	60	28	12	3PH	25	56	3PH
417560_0041	413818	1	0	0	100	0	0	L1	1	10	1PH
417560_0041	413968	0	0	1	0	0	100	L3	1	3	1PH
417560_0041	414004	2	0	4	33.33	0	66.67	3PH	6	10	1PH
417560_0041	414214	2	0	0	100	0	0	L1	2	3	1PH
417560_0041	414473	0	0	1	0	0	100	L3	1	2	1PH
417560_0041	414491	14	18	17	28.57	36.73	34.69	3PH	49	82	3PH
417560_0041	414567	35	21	17	47.95	28.77	23.29	3PH	73	140	3PH
417560_0041	414626	1	0	0	100	0	0	L1	1	10	3PH
417560_0041	414629	5	9	7	23.81	42.86	33.33	3PH	21	52	3PH
417560_0041	415427	8	4	4	50	25	25	3PH	16	48	3PH
417560_0041	415802	2	3	2	28.57	42.86	28.57	3PH	7	20	3PH
417560_0041	415813	18	13	11	42.86	30.95	26.19	3PH	42	75	3PH
417560_0041	416044	1	3	3	14.29	42.86	42.86	3PH	7	16	3PH
417560_0041	416108	0	0	1	0	0	100	L3	1	2	1PH
417560_0041	416208	1	0	0	100	0	0	L1	1	2	1PH
600017_0031	605900	0	1	0	0	100	0	L2	1	3	1PH
600017_0031	605903	1	0	0	100	0	0	L1	1	2	1PH
600017_0031	605904	1	0	0	100	0	0	L1	1	2	1PH
600017_0031	605905	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	605906	1	0	0	100	0	0	L1	1	4	1PH
600017_0031	605907	7	10	5	31.82	45.45	22.73	3PH	22	39	3PH
600017_0031	605908	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	605909	1	0	2	33.33	0	66.67	3PH	3	4	3PH
600017_0031	605910	0	2	0	0	100	0	L2	2	10	3PH
600017_0031	605911	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	605912	0	2	0	0	100	0	L2	2	7	1PH
600017_0031	605916	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605917	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	605919	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	605921	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605922	0	1	0	0	100	0	L2	1	3	1PH
600017_0031	605924	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	605927	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605928	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605931	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605933	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	605935	0	1	0	0	100	0	L2	1	3	1PH
600017_0031	605936	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	605937	0	1	0	0	100	0	L2	1	6	2PH
600017_0031	605938	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	605941	0	7	0	0	100	0	L2	7	13	1PH
600017_0031	605943	1	0	0	100	0	0	L1	1	5	3PH
600017_0031	605945	1	1	0	50	50	0	3PH	2	6	2PH
600017_0031	605947	0	1	0	0	100	0	L2	1	3	1PH

Confidential

600017_0031	605948	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	605949	0	4	0	0	100	0	L2	4	10	1PH
600017_0031	605953	0	3	0	0	100	0	L2	3	5	2PH
600017_0031	605954	1	0	0	100	0	0	L1	1	7	3PH
600017_0031	605957	0	3	0	0	100	0	L2	3	5	1PH
600017_0031	605959	0	0	1	0	0	100	L3	1	3	1PH
600017_0031	606007	0	0	1	0	0	100	L3	1	5	2PH
600017_0031	606008	0	0	4	0	0	100	L3	4	7	3PH
600017_0031	606009	0	2	0	0	100	0	L2	2	5	1PH
600017_0031	606010	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	606011	0	3	0	0	100	0	L2	3	8	2PH
600017_0031	606014	5	10	10	20	40	40	3PH	25	58	3PH
600017_0031	606015	0	1	0	0	100	0	L2	1	2	1PH
600017_0031	606016	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	606019	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	606021	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	606050	0	0	6	0	0	100	L3	6	8	1PH
600017_0031	606051	0	0	1	0	0	100	L3	1	3	1PH
600017_0031	606055	0	0	9	0	0	100	L3	9	26	3PH
600017_0031	606058	0	0	1	0	0	100	L3	1	2	2PH
600017_0031	606059	5	16	7	17.86	57.14	25	3PH	28	45	3PH
600017_0031	606060	0	0	1	0	0	100	L3	1	4	1PH
600017_0031	606062	0	1	28	0	3.45	96.55	L3	29	83	1PH
600017_0031	606063	0	0	3	0	0	100	L3	3	7	1PH
600017_0031	606067	0	0	2	0	0	100	L3	2	4	1PH
600017_0031	606068	0	0	1	0	0	100	L3	1	4	1PH
600017_0031	606069	0	0	1	0	0	100	L3	1	2	1PH
600017_0031	606070	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	606072	0	0	1	0	0	100	L3	1	2	1PH
600017_0031	606073	0	0	1	0	0	100	L3	1	2	1PH
600017_0031	606074	1	0	2	33.33	0	66.67	3PH	3	9	3PH
600017_0031	606076	0	0	34	0	0	100	L3	34	67	1PH
600017_0031	606078	0	0	6	0	0	100	L3	6	9	1PH
600017_0031	606081	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	606084	0	0	3	0	0	100	L3	3	4	1PH
600017_0031	606086	0	0	1	0	0	100	L3	1	3	1PH
600017_0031	606087	0	0	1	0	0	100	L3	1	3	1PH
600017_0031	606088	0	2	1	0	66.67	33.33	3PH	3	8	3PH
600017_0031	606091	0	2	1	0	66.67	33.33	3PH	3	2	3PH
600017_0031	606094	0	2	1	0	66.67	33.33	3PH	3	15	3PH
600017_0031	606095	3	1	1	60	20	20	3PH	5	9	3PH
600017_0031	606540	0	0	1	0	0	100	L3	1	1	1PH
600017_0031	606601	0	0	6	0	0	100	L3	6	14	1PH
600017_0031	606619	15	26	26	22.39	38.81	38.81	3PH	67	107	3PH
600017_0031	606787	0	1	0	0	100	0	L2	1	1	1PH
600017_0031	606852	0	0	5	0	0	100	L3	5	7	1PH
600017_0034	601085	16	9	9	47.06	26.47	26.47	3PH	34	67	3PH
600017_0034	601086	15	20	16	29.41	39.22	31.37	3PH	51	105	3PH

Confidential

600017_0034	601087	19	16	19	35.19	29.63	35.19	3PH	54	102	3PH
600017_0034	601088	1	1	2	25	25	50	3PH	4	19	3PH
600017_0034	601089	0	0	2	0	0	100	L3	2	4	1PH
600017_0034	601090	0	0	1	0	0	100	L3	1	9	1PH
600017_0034	601091	0	0	2	0	0	100	L3	2	3	1PH
600017_0034	601097	1	0	0	100	0	0	L1	1	1	1PH
600017_0034	601098	4	2	4	40	20	40	3PH	10	22	3PH
600017_0034	601100	0	1	0	0	100	0	L2	1	2	1PH
600017_0034	601102	1	0	0	100	0	0	L1	1	4	1PH
600017_0034	601104	0	0	2	0	0	100	L3	2	6	2PH
600017_0034	601107	0	0	1	0	0	100	L3	1	1	2PH
600017_0034	601110	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601111	0	0	1	0	0	100	L3	1	4	1PH
600017_0034	601113	0	1	1	0	50	50	3PH	2	4	3PH
600017_0034	601114	0	0	2	0	0	100	L3	2	6	1PH
600017_0034	601118	0	0	2	0	0	100	L3	2	3	1PH
600017_0034	601119	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601121	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601122	4	0	0	100	0	0	L1	4	5	1PH
600017_0034	601123	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601124	0	0	1	0	0	100	L3	1	2	3PH
600017_0034	601125	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601128	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601131	1	0	0	100	0	0	L1	1	3	1PH
600017_0034	601132	0	0	2	0	0	100	L3	2	4	
600017_0034	601134	0	0	2	0	0	100	L3	2	2	1PH
600017_0034	601135	0	0	3	0	0	100	L3	3	8	1PH
600017_0034	601136	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601137	0	0	5	0	0	100	L3	5	9	2PH
600017_0034	601138	4	0	8	33.33	0	66.67	3PH	12	31	3PH
600017_0034	601139	0	0	2	0	0	100	L3	2	11	1PH
600017_0034	601140	1	0	0	100	0	0	L1	1	3	1PH
600017_0034	601142	0	0	1	0	0	100	L3	1	1	3PH
600017_0034	601144	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601145	1	0	8	11.11	0	88.89	3PH	9	20	1PH
600017_0034	601154	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601155	0	0	4	0	0	100	L3	4	4	1PH
600017_0034	601156	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601157	0	0	1	0	0	100	L3	1	13	3PH
600017_0034	601158	0	0	1	0	0	100	L3	1	4	1PH
600017_0034	601163	0	0	4	0	0	100	L3	4	6	2PH
600017_0034	601164	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601167	0	0	2	0	0	100	L3	2	2	1PH
600017_0034	601169	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601170	0	1	0	0	100	0	L2	1	1	1PH
600017_0034	601173	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601174	0	0	2	0	0	100	L3	2	2	1PH
600017_0034	601176	0	0	1	0	0	100	L3	1	1	1PH

Confidential

600017_0034	601177	0	0	2	0	0	100	L3	2	3	1PH
600017_0034	601178	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601181	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601182	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601183	0	0	1	0	0	100	L3	1	3	1PH
600017_0034	601187	0	0	9	0	0	100	L3	9	11	1PH
600017_0034	601188	0	3	0	0	100	0	L2	3	5	1PH
600017_0034	601189	0	0	8	0	0	100	L3	8	21	1PH
600017_0034	601191	0	4	0	0	100	0	L2	4	11	1PH
600017_0034	601192	7	0	12	36.84	0	63.16	3PH	19	76	3PH
600017_0034	601194	9	0	19	32.14	0	67.86	3PH	28	99	3PH
600017_0034	601197	0	0	1	0	0	100	L3	1	4	1PH
600017_0034	601198	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601200	0	0	1	0	0	100	L3	1	4	1PH
600017_0034	601202	0	0	1	0	0	100	L3	1	1	2PH
600017_0034	601203	0	0	1	0	0	100	L3	1	6	2PH
600017_0034	601210	3	0	0	100	0	0	L1	3	6	2PH
600017_0034	601214	1	0	0	100	0	0	L1	1	1	1PH
600017_0034	601217	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601218	0	0	1	0	0	100	L3	1	4	1PH
600017_0034	601222	1	0	7	12.5	0	87.5	3PH	8	19	1PH
600017_0034	601223	1	0	0	100	0	0	L1	1	2	3PH
600017_0034	601225	0	0	2	0	0	100	L3	2	4	1PH
600017_0034	601227	0	0	1	0	0	100	L3	1	3	1PH
600017_0034	601230	1	0	0	100	0	0	L1	1	1	1PH
600017_0034	601231	0	0	1	0	0	100	L3	1	3	2PH
600017_0034	601232	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601235	0	0	5	0	0	100	L3	5	5	1PH
600017_0034	601236	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601238	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601239	0	0	2	0	0	100	L3	2	4	2PH
600017_0034	601245	2	0	0	100	0	0	L1	2	2	1PH
600017_0034	601248	0	0	1	0	0	100	L3	1	3	1PH
600017_0034	601249	0	0	1	0	0	100	L3	1	2	1PH
600017_0034	601256	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	601261	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	606300	8	0	13	38.1	0	61.9	3PH	21	53	3PH
600017_0034	606588	0	1	0	0	100	0	L2	1	1	1PH
600017_0034	606690	0	0	1	0	0	100	L3	1	1	1PH
600017_0034	606922	0	0	1	0	0	100	L3	1	1	1PH
770003_0001	760932	0	0	1	0	0	100	L3	1	3	3PH
770003_0001	760935	1	0	1	50	0	50	3PH	2	7	1PH
770003_0001	760939	0	0	1	0	0	100	L3	1	4	1PH
770003_0001	760940	0	0	3	0	0	100	L3	3	4	1PH
770003_0001	760941	2	1	0	66.67	33.33	0	3PH	3	6	1PH
770003_0001	760946	1	0	2	33.33	0	66.67	3PH	3	18	3PH
770003_0001	760954	0	0	2	0	0	100	L3	2	3	1PH
770003_0001	760958	1	0	0	100	0	0	L1	1	6	3PH

Confidential

770003_0001	760960	0	0	1	0	0	100	L3	1	5	1PH
770003_0001	760961	0	0	3	0	0	100	L3	3	8	1PH
770003_0001	760963	1	0	0	100	0	0	L1	1	1	1PH
770003_0001	760965	2	0	0	100	0	0	L1	2	2	1PH
770003_0001	760978	3	0	0	100	0	0	L1	3	4	3PH
770003_0001	760989	0	0	2	0	0	100	L3	2	6	1PH
770003_0001	762512	1	0	0	100	0	0	L1	1	1	1PH
770003_0001	762518	0	1	0	0	100	0	L2	1	3	1PH
770003_0001	762519	4	2	1	57.14	28.57	14.29	3PH	7	9	3PH
770003_0001	762526	0	1	0	0	100	0	L2	1	3	1PH
770003_0001	762527	0	2	1	0	66.67	33.33	3PH	3	10	1PH
770003_0001	762725	0	1	0	0	100	0	L2	1	8	3PH
770003_0001	762727	0	0	1	0	0	100	L3	1	10	1PH
770003_0001	762737	1	2	2	20	40	40	3PH	5	8	3PH
770003_0001	762803	9	0	0	100	0	0	L1	9	15	1PH
770003_0001	762806	0	0	3	0	0	100	L3	3	6	1PH
770003_0001	762808	0	0	1	0	0	100	L3	1	3	1PH
770003_0001	763122	13	7	3	56.52	30.43	13.04	3PH	23	53	3PH
770003_0001	763151	2	0	0	100	0	0	L1	2	2	1PH
770003_0001	763173	0	0	1	0	0	100	L3	1	2	1PH
770003_0001	763188	1	0	0	100	0	0	L1	1	5	3PH
770003_0001	763695	0	3	0	0	100	0	L2	3	8	1PH
770003_0001	774822	6	8	5	31.58	42.11	26.32	3PH	19	43	3PH
770003_0001	776454	9	10	8	33.33	37.04	29.63	3PH	27	60	3PH
770003_0001	776719	12	10	10	37.5	31.25	31.25	3PH	32	68	3PH
770003_0001	776900	14	13	14	34.15	31.71	34.15	3PH	41	86	3PH
770003_0001	777216	14	15	18	29.79	31.91	38.3	3PH	47	100	3PH
770003_0001	777217	15	21	14	30	42	28	3PH	50	99	3PH
770003_0001	777218	26	18	7	50.98	35.29	13.73	3PH	51	114	3PH
770003_0001	777219	21	19	18	36.21	32.76	31.03	3PH	58	121	3PH
770003_0009	744033	11	4	0	73.33	26.67	0	3PH	15	35	3PH
770003_0009	744143	1	0	0	100	0	0	L1	1	1	1PH
770003_0009	744342	1	0	0	100	0	0	L1	1	6	3PH
770003_0009	744391	3	2	0	60	40	0	3PH	5	16	3PH
770003_0009	744392	2	1	0	66.67	33.33	0	3PH	3	7	1PH
770003_0009	744403	1	0	0	100	0	0	L1	1	1	3PH
770003_0009	745216	2	1	0	66.67	33.33	0	3PH	3	4	3PH
770003_0009	745414	0	0	1	0	0	100	L3	1	9	1PH
770003_0009	745670	6	0	0	100	0	0	L1	6	11	1PH
770003_0009	745680	2	0	0	100	0	0	L1	2	6	1PH
770003_0009	745686	13	0	2	86.67	0	13.33	3PH	15	32	3PH
770003_0009	745694	1	0	0	100	0	0	L1	1	4	1PH
770003_0009	745712	0	1	0	0	100	0	L2	1	2	1PH
770003_0009	762509	0	0	2	0	0	100	L3	2	9	1PH
770003_0009	762510	0	0	2	0	0	100	L3	2	7	1PH
770003_0009	762511	1	0	0	100	0	0	L1	1	2	3PH
770003_0009	762520	1	0	0	100	0	0	L1	1	6	3PH
770003_0009	762522	0	0	2	0	0	100	L3	2	9	1PH

Confidential

770003_0009	762523	5	1	6	41.67	8.33	50	3PH	12	52	3PH
770003_0009	762524	21	16	13	42	32	26	3PH	50	103	3PH
770003_0009	762525	0	1	0	0	100	0	L2	1	5	1PH
770003_0009	762581	5	6	13	20.83	25	54.17	3PH	24	38	3PH
770003_0009	762588	1	0	0	100	0	0	L1	1	6	1PH
770003_0009	762589	0	0	1	0	0	100	L3	1	3	1PH
770003_0009	762591	6	0	0	100	0	0	L1	6	10	1PH
770003_0009	762592	1	0	0	100	0	0	L1	1	7	1PH
770003_0009	762593	2	0	0	100	0	0	L1	2	6	1PH
770003_0009	762596	5	4	3	41.67	33.33	25	3PH	12	27	3PH
770003_0009	762597	0	4	0	0	100	0	L2	4	6	1PH
770003_0009	762598	0	1	0	0	100	0	L2	1	2	1PH
770003_0009	762605	1	0	0	100	0	0	L1	1	1	1PH
770003_0009	762606	1	1	0	50	50	0	3PH	2	3	1PH
770003_0009	762608	0	1	0	0	100	0	L2	1	1	1PH
770003_0009	762610	2	0	0	100	0	0	L1	2	6	1PH
770003_0009	762611	3	1	1	60	20	20	3PH	5	18	3PH
770003_0009	762612	1	1	0	50	50	0	3PH	2	6	3PH
770003_0009	762613	5	0	0	100	0	0	L1	5	29	3PH
770003_0009	762614	0	0	1	0	0	100	L3	1	2	1PH
770003_0009	762616	1	0	0	100	0	0	L1	1	4	1PH
770003_0009	762617	1	0	0	100	0	0	L1	1	7	1PH
770003_0009	762618	7	0	0	100	0	0	L1	7	18	1PH
770003_0009	762723	0	0	7	0	0	100	L3	7	13	1PH
770003_0009	762731	0	0	1	0	0	100	L3	1	2	1PH
770003_0009	762732	0	0	1	0	0	100	L3	1	5	3PH
770003_0009	762733	0	0	7	0	0	100	L3	7	14	1PH
770003_0009	762734	0	0	3	0	0	100	L3	3	6	1PH
770003_0009	762735	0	0	8	0	0	100	L3	8	21	1PH
770003_0009	762741	1	0	1	50	0	50	3PH	2	5	3PH
770003_0009	762742	3	5	0	37.5	62.5	0	3PH	8	17	1PH
770003_0009	762743	12	6	8	46.15	23.08	30.77	3PH	26	40	3PH
770003_0009	762744	1	0	0	100	0	0	L1	1	1	1PH
770003_0009	762745	8	0	1	88.89	0	11.11	3PH	9	13	1PH
770003_0009	762746	0	1	0	0	100	0	L2	1	2	1PH
770003_0009	762799	0	0	1	0	0	100	L3	1	2	3PH
770003_0009	762800	0	0	1	0	0	100	L3	1	2	1PH
770003_0009	762804	0	0	4	0	0	100	L3	4	8	1PH
770003_0009	762807	0	0	1	0	0	100	L3	1	9	3PH
770003_0009	762809	1	0	0	100	0	0	L1	1	2	1PH
770003_0009	762810	0	0	1	0	0	100	L3	1	2	1PH
770003_0009	762811	2	0	0	100	0	0	L1	2	2	1PH
770003_0009	762813	1	0	0	100	0	0	L1	1	7	1PH
770003_0009	762814	11	11	12	32.35	32.35	35.29	3PH	34	90	3PH
770003_0009	762815	0	0	1	0	0	100	L3	1	5	1PH
770003_0009	762817	2	0	2	50	0	50	3PH	4	23	3PH
770003_0009	762818	0	3	2	0	60	40	3PH	5	22	3PH
770003_0009	762826	38	16	13	56.72	23.88	19.4	3PH	67	95	3PH

Confidential

770003_0009	762933	6	10	13	20.69	34.48	44.83	3PH	29	66	3PH
770003_0009	762940	0	0	1	0	0	100	L3	1	4	1PH
770003_0009	763121	5	7	1	38.46	53.85	7.69	3PH	13	30	3PH
770003_0009	763238	2	0	0	100	0	0	L1	2	6	1PH
770003_0009	763243	1	0	0	100	0	0	L1	1	7	1PH
770003_0009	763244	0	0	1	0	0	100	L3	1	1	1PH
770003_0009	763250	3	0	0	100	0	0	L1	3	6	1PH
770003_0009	763256	0	0	1	0	0	100	L3	1	2	1PH
770003_0009	763261	10	0	2	83.33	0	16.67	3PH	12	22	1PH
770003_0009	763263	0	1	2	0	33.33	66.67	3PH	3	13	3PH
770003_0009	763335	3	1	3	42.86	14.29	42.86	3PH	7	14	3PH
770003_0009	763348	0	1	1	0	50	50	3PH	2	4	3PH
770003_0009	763351	0	0	1	0	0	100	L3	1	1	1PH
770003_0009	763355	1	0	0	100	0	0	L1	1	6	3PH
770003_0009	763357	0	0	1	0	0	100	L3	1	5	1PH
770003_0009	763359	2	2	2	33.33	33.33	33.33	3PH	6	13	3PH
770003_0009	763360	1	0	2	33.33	0	66.67	3PH	3	15	1PH
770003_0009	763365	1	0	0	100	0	0	L1	1	4	3PH
770003_0009	763366	0	0	4	0	0	100	L3	4	5	1PH
770003_0009	763369	3	0	0	100	0	0	L1	3	15	1PH
770003_0009	763370	0	1	0	0	100	0	L2	1	1	1PH
770003_0009	763371	0	1	0	0	100	0	L2	1	5	3PH
770003_0009	763372	2	0	0	100	0	0	L1	2	5	1PH
770003_0009	763375	0	0	1	0	0	100	L3	1	1	1PH
770003_0009	763376	0	2	1	0	66.67	33.33	3PH	3	7	3PH
770003_0009	763379	4	0	0	100	0	0	L1	4	5	1PH
770003_0009	763381	1	0	0	100	0	0	L1	1	1	1PH
770003_0009	763382	1	0	0	100	0	0	L1	1	3	1PH
770003_0009	763383	0	0	1	0	0	100	L3	1	3	1PH
770003_0009	763386	0	0	1	0	0	100	L3	1	5	1PH
770003_0009	763387	1	0	0	100	0	0	L1	1	3	1PH
770003_0009	763389	0	0	2	0	0	100	L3	2	5	1PH
770003_0009	763471	10	7	11	35.71	25	39.29	3PH	28	48	3PH
770003_0009	763472	1	0	2	33.33	0	66.67	3PH	3	5	3PH
770003_0009	763483	14	11	14	35.9	28.21	35.9	3PH	39	87	3PH
770003_0009	763676	3	0	1	75	0	25	3PH	4	6	1PH
770003_0009	774592	8	9	7	33.33	37.5	29.17	3PH	24	49	3PH
770003_0009	776655	64	52	44	40	32.5	27.5	3PH	160	299	3PH
770003_0009	776718	44	38	47	34.11	29.46	36.43	3PH	129	313	3PH
770003_0009	776899	20	19	13	38.46	36.54	25	3PH	52	105	3PH
770003_0009	776978	24	17	17	41.38	29.31	29.31	3PH	58	127	3PH
770003_0009	777180	7	12	8	25.93	44.44	29.63	3PH	27	56	3PH
770003_0009	777214	24	25	16	36.92	38.46	24.62	3PH	65	127	3PH
770003_0009	777215	13	12	13	34.21	31.58	34.21	3PH	38	73	3PH

Confidential

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

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

Confidential

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.

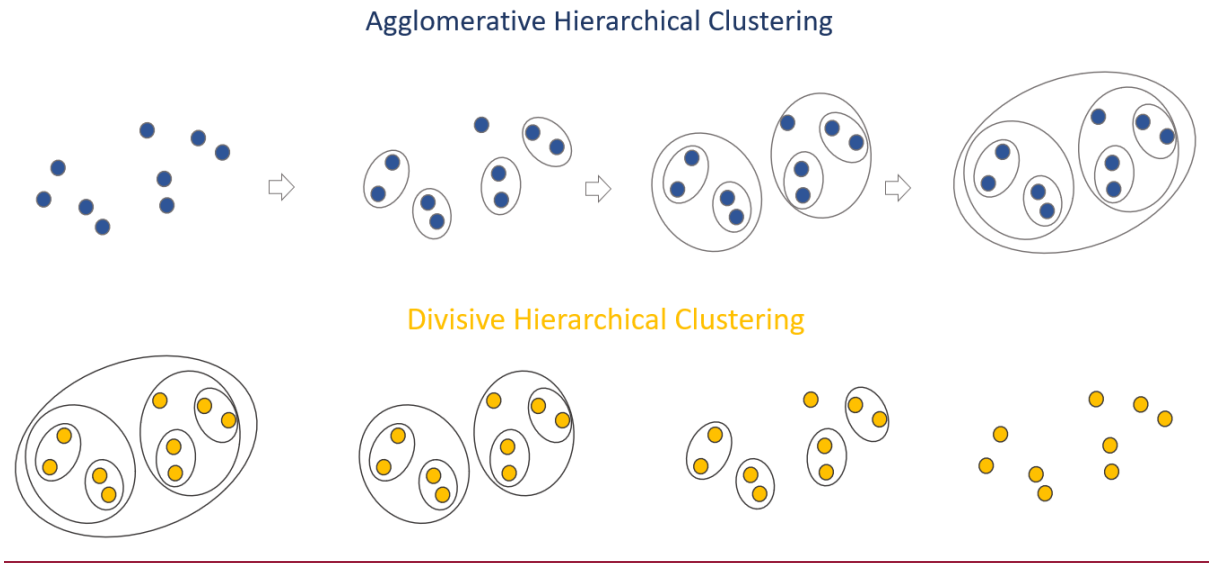


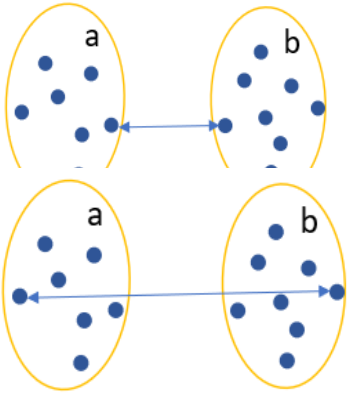
Figure 15: Agglomerative vs Divisive Hierarchical Clustering

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.



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

Complete Linkage

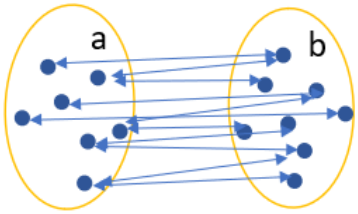
Confidential

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.

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

Average Linkage

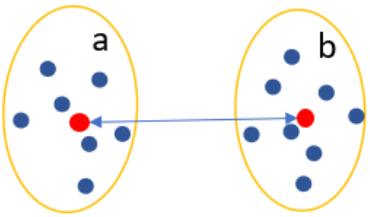
In average linkage method, the distance between two clusters is defined as the average of all pairwise distances across the two clusters.



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

Centroid Linkage

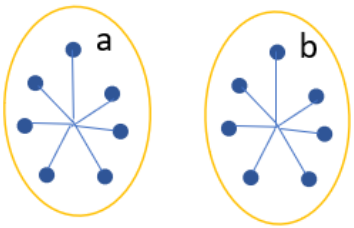
In the centroid linkage method, the distance between two clusters is the distance between the two mean vectors of the clusters. At each stage, the two clusters with the smallest centroid distance are merged.



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

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.



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

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.

Confidential

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.

Confidential

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.

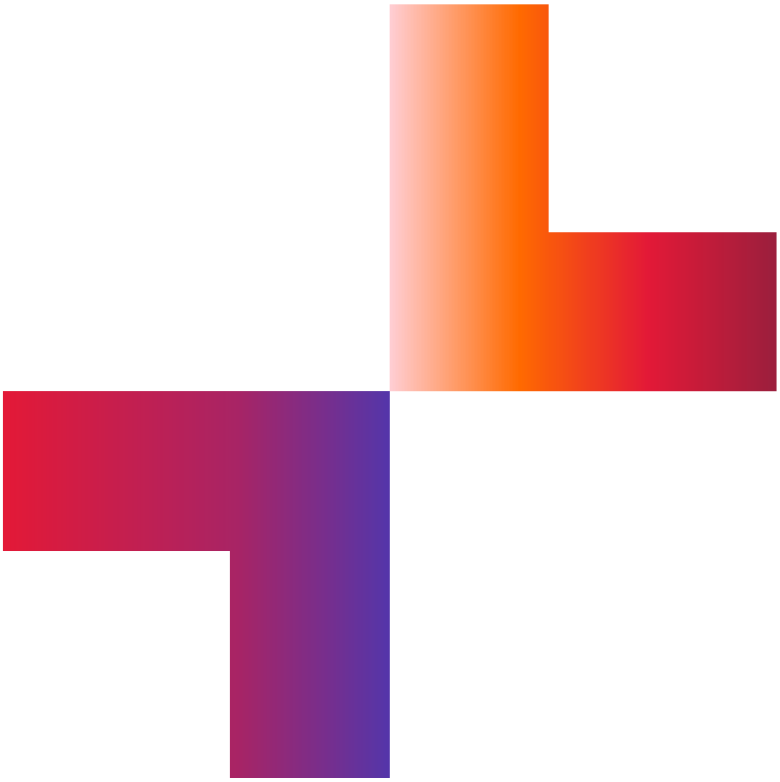
Fowlkes-Mallows Index (FMI) is a geometric mean of pairwise precision and recall, using True Positive (TP), False Positive (FP) and False Negative (FN).

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:

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

Confidential



cgi.com

Confidential