# EPIC Final Report

| | |
|---|---|
| Program | Electric Program Investment Charge (EPIC) |
| Administrator | San Diego Gas & Electric Company |
| Project Number | EPIC-3, Project 3, Module 2 |
| Project Name | Application of Advanced Metering Infrastructure (AMI) Data to Advanced Utility System Operations |
| Date | December 31, 2021 |

## Attribution

This comprehensive final report documents work done in this EPIC activity. The project team for this work included the following individuals, listed alphabetically by last name.

San Diego Gas and Electric (SDG&E)

- Tom Bialek
- Jay Bick
- Fidel R. Castro
- Frank Goodman
- Chippy Impreso
- Julian Jones
- Kyle Kewley
- Song Lee
- Gina Lindsay
- Jenell McKay
- William O'Brien
- Subburaman Sankaran
- Amin Salmani
- Matt Smith
- Tyson Swetek
- Stacy Williams
- William Wood

Vendors Personnel

- Fred Behrman
- Molly Du
- Jaishankar Jayaraman
- Erik Nilssen
- Brian Sherwin
- John Stalcup
- Thad Stalcup
- Mark Sieben
- Brendan Stellarum
- Ralph Young

## Executive Summary

The objective of EPIC-3, Project 3 was to demonstrate capabilities for leveraging SDG&E's advanced metering infrastructure (AMI) system to provide actionable secondary voltage data and analysis to SDG&E staff and other prospective users. The project focus included two modules. Module 1 focused on using AMI data for a voltage sensor network, while Module 2 focused on using AMI data to identify endpoint phasing and meter-to-transformer mapping. This report describes the pre-commercial demonstration for Module 2.

This report consists of five parts. Part I contains general information about Module 2 of this EPIC project. Parts II, III and IV describe the three separate and distinct methodologies that were demonstrated in this module; where Parts II and III describe demonstrations where SDG&E worked collaboratively with external vendors, and Part IV captures the internal effort by SDG&E personnel to identifying endpoint phasing based on publicly available studies. In Part IV, no work was done on meter-to-transformer mapping. Finally, Part V contains a summary of the methodologies; findings, and recommendations; and information on technical transfer and commercialization.

*PART I – General Information*

This demonstration seeks to determine the feasibility of executing two use cases – AMI meter phasing and meter-to-transformer mapping while constraining the AMI input source to two meters per transformer. Through a competitive process, two vendors were selected to assist in executing these use cases.

SDG&E's AMI system provides abundant usage and voltage information with high resolution and accuracy. By using these data, the project sought to address several issues:

- Traditionally, the only way for a utility to correct its phasing model was through deployment of personnel into the field to verify meter-to-transformer connectivity and transformer-to-phase connectivity. This practice is both expensive and time consuming and further complicated by increased risk for employee safety.
- Utility circuits are increasingly more complex as new technology is integrated into the grid. With the complexity of the grid ever increasing, the need for accurate phase ID solutions becomes more prominent to promptly address issues of phase balancing and transformer loading.
- Circuit distribution information in utility systems has never been perfectly accurate. Where the information is inaccurate, it can lead to unbalanced phases, as well as overloading, and underutilization of transformers and other equipment.

There are three dimensions that describe the benefits associated with improving records accuracy of phase identification and meter-to-transformer mapping. These include distribution network reliability, increased safety, and reduced cost. However, any utility can improve records accuracy by manually validating conditions using field technicians, also known as field verification. Therefore, the true benefits of this program focus on the last two focus dimensions, increased safety and reduced cost.

The metrics used to determine success are the accuracy rates of prediction. This is simply measured by comparing the correct predictions to the total amount of endpoints. This implies that the true values for

phasing and meter-to-transformer are known. For this project, field verification was used as the source of the true value even though 100% accurate field verification is not assured. In fact, through the efforts of one vendor, the project team discovered field verification rates of approximately 95%. While not perfect, this is the best means of determining the "source of truth".

In order to demonstrate the algorithms, SDG&E selected two feeders: Feeders (circuits) A and B. Feeder A has 325 connected transformers and 5,173 connected meters, while Feeder B has 649 connected transformers and 2,393 connected meters. The voltage readings had a precision of 0.15 volts and were collected over the course of two years, from October 21, 2018, to October 20, 2020.

*PART II – Methodology A*

Part II captures the results of Methodology A, the first of two methodologies where SDG&E worked with an external vendor. This methodology demonstrated use of an established, data analytics platform to ingest, analyze, and evaluate end point phase identification and meter-to-transformer mapping.

In this methodology, the project team executed four distinct tasks – 1) data collection and cleansing, 2) execution of phase identification algorithms, 3) execution of meter-to-transformer mapping algorithms, and 4) evaluation of results using field verified data. During tasks 2 and 3, the vendor executed several iterations of the algorithm in order to optimize the results. Once optimized, the results were compared to field verification results in task 4. This methodology had mixed results – relatively high accuracy for phase identification with mediocre results for meter-to-transformer mapping. For phase identification, results of 98% and 97%, and for meter-to-transformer mapping, results of 82% and 79% were achieved for circuits A and B respectively.

*PART III – Methodology B*

Part III captures the results for Methodology B, the second of two methodologies where SDG&E worked with an external vendor. This methodology demonstrated use of an established, data analytics platform to ingest, analyze, evaluate, and display results for end point phase identification and meter to transformer mapping.

The project was organized into three tasks – 1) phase identification, 2) meter-to-transformer mapping, 3) field validation. In this methodology, data cleansing occurred during tasks 1 and 2. Data for four circuits, circuits A, B, C, and D were provided to this vendor; however, field verification data were only provided for circuits A and B. Therefore, no field validation was performed for circuits C and D. For phase identification, results of 83% and 92%, and for meter-to-transformer connectivity, results of 65% and 89% were achieved for circuits A and B respectively.

*PART IV – Internal Methodology*

Part IV captures the results of the internal methodology executed by SDG&E personnel and describes the demonstration based on publicly available studies. Unlike Methodology A and B, the project team focused solely on phase identification. No work was done on meter-to-transformer mapping.

The project team used an internally developed clustering algorithm based on research from publicly available sources. This methodology was executed to give the project team a baseline metric of results accuracy using simple time-series clustering. This methodology had two limitations – 1) the results output

is provided in “phase groups”, rather than the actual phase, and 2) the analysis is restricted to single-phase, line to neutral meters. Results for phase identification were 72.5% and 95.5% for circuits A and B respectively.

*PART V – Summary and Recommendations*

All methodologies agree that automatic phase identification is achievable at acceptable levels of accuracy using only two meters per transformer. Meter-to-transformer connectivity, however, proved less precise. Commercialization of any of the methodologies is not recommended given the constraint of two meters per transformer. This constraint minimizes the amount of infrastructure support needed to transfer data from meters to a centralized location for further processing and therefore reduces cost. This cost minimization constraint is a primary focus of the demonstration. Advancements in machine learning, advanced data mining, and artificial intelligence coupled with reduced data storage costs and improved network throughput have created numerous opportunities to use AMI data beyond the use case of meter reading and billing. The project team does not recommend pursuing the successful use case in this study, analytical based phase identification, as a single use case. Rather, it recommends exploring additional use cases that would benefit a wider audience. Exploring additional use cases is beyond the scope of this EPIC project.

# Table of Contents

# PART I

## 1.0 Introduction

The objective of EPIC-3, Project 3 was to demonstrate capabilities for leveraging SDG&E's advanced metering infrastructure (AMI) system to provide actionable secondary voltage data and analysis to SDG&E staff and other prospective users. The project focus included two modules. Module 1 focused on using AMI data for a voltage sensor network, while Module 2 focused on using AMI data to identify endpoint phasing and meter-to-transformer mapping. This report describes the pre-commercial demonstration for Module 2.

This report consists of five parts. Part I, this part, contains general information about Module 2 of this EPIC project. Parts II, III, and IV describe the three separate and distinct methodologies that were used in the demonstration. Parts II and III describe demonstrations where SDG&E worked collaboratively with external vendors. Part IV captures the internal effort by SDG&E personnel to identifying endpoint phasing based on publicly available studies. In this last methodology, no work was done on meter-to-transformer mapping. Finally, Part 5 contains a summary of the methodologies; findings and conclusions; and information on technical transfer and commercialization.

## 2.0 Module Objectives

This project module seeks to demonstrate and assess analytical approaches to predict phase identity and meter-to-transformer mapping. While endpoint phasing and meter-to-transformer mapping has been accomplished using analytical methods in the past, this project attempts to accomplish this using data from only two meters per transformer. The project module provides proof of concept by using algorithms that can consume SDG&E's AMI, SCADA, GIS, and other relevant data to determine phasing and meter-to-transformer connectivity on two sample circuits. Data from these two sample circuits were used in all three methodologies to help aid in consistency of results. A goal of the pre-commercial demonstration is to identify a path for SDG&E to replace or reduce the existing expensive methods of verifying phasing and meter to transformer mapping with a reliable and accurate analytical approach.

## 3.0 Issues and Policies Addressed

The AMI system provides abundant usage and voltage information with high resolution and accuracy. Recent advances in machine learning models make it possible to identify phase and predict meter-to-transformer connections through the analysis of this high-frequency data. Issues addressed during this project include:

- Traditionally, the only way for a utility to correct its phasing model was through deployment of personnel into the field to verify meter-to-transformer connectivity and transformer-to-phase connectivity, and thereby identify endpoint phasing. This practice is both expensive and time consuming and further complicated by increased risk for employee safety.
- Utility circuits are increasingly more complex as new technology is integrated into the grid. With the complexity of the grid ever increasing, the need for heightened monitoring capabilities on

utility grid equipment is growing. In particular, the need for accurate phase ID solutions becomes more prominent to promptly address issues of phase balancing and transformer loading.

- Circuit distribution information in utility systems has never been perfectly accurate. This inaccurate information can lead to unbalanced phases, as well as overloading, and underutilization of transformers and other equipment. Overloading shortens the life expectancy of distribution equipment and in the most extreme cases, presents a safety hazard. Underutilization leads to unnecessary capital expenditures on additional equipment. Imbalanced phases result in higher technical losses in transfer and increases operational costs.

The notion that voltage data can be used to solve phase ID records inaccuracy relies on the fact that voltage fluctuations on two meters have a closer correlation when they are on the same transformer as compared to when they are on separate transformers. The same principal applies to the meters on the same phase on a feeder. If this voltage data can be leveraged to create an accurate connectivity model, then all the benefits associated can be accessed at a fraction of the cost and much faster than is possible with field verifications.

## 4.0 Project Focus

The focus of the project was to test whether phase identification and meter-to-transformer can be performed accurately using existing data from AMI, SCADA and GIS data sources. More specifically, the purpose of this project is to test the performance of vendor and internally developed algorithms with a limited amount of data. For each transformer on the two feeders, two meters were chosen for voltage data collection. To mitigate the concern of overwhelming network traffic if every meter was included in the analysis, only a subset of data was used. If the phase identification and meter-to-transformer performance has an acceptable level of accuracy on such a limited dataset, this proves that data-driven solutions are viable in areas with low network bandwidth. Low network utilization could also cut the cost of data transferring, data storage, and data processing. Finally, this project will also shed light on promising directions of further research.

## 5.0 Project Scope Summary

The project scope is to discover, demonstrate, evaluate, and validate vendor and internally developed methods to automatically identify meters' phasing information and meter-to-transformer mapping.

Testing of the algorithms to provide a baseline metric of results accuracy using time series clustering methods is performed utilizing five-minute interval AMI voltage data at all service transformers for the selected feeders. For this project, SDG&E selected two feeders: Feeders (circuits) A and B. Feeder A has 325 connected transformers and 5,173 connected meters. It serves a relatively dense suburban neighborhood with a mix of overhead and underground wiring and a relatively even mix of line-line (L-L) and (L-N) phasing on the transformers. Feeder B has 649 connected transformers and 2,393 connected meters. It serves a spread-out suburban neighborhood with predominantly underground wiring and predominantly line to neutral (L-N) phasing on the transformers.

The data used for the phase ID and meter-to-transformer solutions consisted of voltage readings for two meters per transformer across the two feeders. The voltage readings had a precision of 0.15 volts and were collected over the course of two years, from October 21, 2018, to October 20, 2020.

Primary outcomes include:

- Evaluation of data analytics for phase identification and meter-to-transformer mapping
- Demonstration using SDG&E's SCADA, GIS, and AMI data to predict phase ID and meter-to-transformer mapping
- Demonstration of any additional analytical methods/applications of AMI data to enhance the electric utility's operations
- Recommendations for full-scale deployment for operational use
- Support to SDG&E in determining costs and benefits for adoption of the demonstrated methods into commercial practice

# 6.0 Benefits Analysis/Metrics

## 6.1 Initial Benefit Estimate and Value Proposition

The initial benefit estimate focused on the following core areas:

- Improved distribution network reliability
    - Allow for more accurate phase balancing
    - Improved data for asset management, especially transformer utilization
- Increased safety
    - Better identification of impacted endpoints during outages
    - Better guidance for trouble teams
    - Lower risk of injury by reducing field visits
- Reduced cost
    - Reduce the need to store exceptionally large data sets/reduce AMI related capital infrastructure expense
    - Decrease the volume of costly field visits
    - Reduce data mining and field visit requirements using readily available data – AMI, SCADA, GIS
    - Increase accuracy in forecasting – reduce/delay capital expenditure

These dimensions describe the benefits associated with improving records accuracy of phase identification and meter-to-transformer mapping. Any utility can improve records accuracy in these two metrics – phase identification and meter to transformer mapping – by manually validating conditions using field technicians. However, field verification is time consuming, costly, and represents a safety risk to field personnel. Therefore, the true benefits of this program focus on increased safety and reduced cost.

### 6.1.1 Increased Safety

Automated asset phase mapping reduces the need for manual field verification on asset phasing, thereby reducing potential SDG&E employee contact with live wires when manually identifying phase. In addition to reducing electrical hazards, reduction of field visits translates to fewer truck rolls and the risks associated with cumulative miles traveled.

Improvements to phase balancing also supports the avoidance of transformer overload failures and provides better loading data to mitigate unsafe loading conditions that could result in hazardous exposure to equipment explosions or downed wires.

### 6.1.2 Reduced Cost

*Network Storage/Reduction in AMI Capital Infrastructure*

One primary focus of this project is to determine if accurate phasing and meter-to-transformer mapping can be accomplished using only two meters per transformer. At the time of project initiation, this appeared unprecedented. Past studies and existing commercially available products use a much higher ratio of meters to transformers. By achieving a high level of accuracy using only two meters per transformer, the requirement of storing these data is drastically reduced, thus reducing the cost of network storage.

AMI capital infrastructure, specifically the back-haul network, is not designed to transfer substantial amounts of data. They are typically designed to transmit only what is needed for reading meters. By using the AMI infrastructure to carry, not only meter reading data, but also voltage data, the network capacity requirement increases dramatically. By minimizing the amount of data needed from the meter to voltage readings from only two meters per transformer, the requirement to increase back-haul capacity is minimized. Therefore, the overall cost is reduced.

*Reduction in Field Visits*

Only a subset of field visits can be eliminated by using an algorithm. Accuracy well above 95% can be achieved using field visits. Utilities *will* update their systems records (OMS, GIS, etc.) using field verified results because it is a time proven method. However, many utilities will *not* update their system records when phasing and meter-to-transformer mapping is verified using an algorithm, largely because the process is new and unproven. Therefore, the reduction in field visits will be limited to those operational use cases where phasing and meter-to-transformer mapping *can* be accomplished without field verification, that is, with the aforementioned algorithm. At this time, these operational use cases are limited to:

- Distribution load balancing
- System planning
- Outage response (meter-to-transformer mapping only)
- Model validation – specifically where system records may be inconsistent with reality and an algorithm is used in conjunction with field validation
- Distributed energy resource (DER) hosting approval

- Future analytics such as transformer utilization, system planning analysis, voltage analytics and outage management

*Using Available Data*

By using readily available data, such as those from SCADA, the MDMS (AMI data), etc., additional data mining and field visits can be eliminated. At SDG&E, data from many systems are stored in OSIsoft PI and is readily available for analysis, and therefore eliminates the need to capture and store data using other methods.

*Increase Accuracy in Forecasting*

Accurate phase information is needed to effectively plan distribution assets. Distribution planning engineers will access available records and then use that information to forecast capital expenditures. By having access to an easy method of determining phasing information, planners can more efficiently gather the information they need to make the right decisions. This in turn can result in a reduction or delay in capital expenditure. This of course assumes that the results returned from the algorithm are sufficiently accurate.

## 6.2 Initial Selection of Metrics

The metrics used to determine success are the accuracy rates of prediction. This is simply measured by comparing the correct predictions to the total amount of endpoints. Obviously, this implies that the true values for phasing and meter-to-transformer are known. For this project, field verification was used as the source of the true value even though 100% accurate field verification is not assured. In fact, through the efforts of other vendors, the project team discovered field verification rates of approximately 95%. While not perfect, this is the best means of determining the “source of truth”.

There is no industry standard for minimum level of accuracy. Further, a minimum level of accuracy depends on the operational use case relying on the data and the safety risk to employees. As noted above, using an algorithm-based phase identification method may never be acceptable when employee safety is involved. However, using this analytical method for distribution load balancing, system planning, model validation, etc., may be perfectly acceptable with each specific use case requiring a minimum value of accuracy. In general, accuracy of greater than 95% is considered acceptable.

## PART II

Part II captures the results of Methodology A, the first of two methodologies where SDG&E worked with an external vendor.

## PART II List of Illustrations

## Part II List of Tables

| Table Number | Description of Tables |
|---|---|
| Table 10 | Distribution of meters by average voltage group |
| Table 11 | Phase Identification Data Trimming |
| Table 12 | Meter-to-transformer Data Trimming |
| Table 13 | Confusion matrix that compares model prediction and utility company's records for Feeder A |
| Table 14 | Confusion matrix that compares model prediction and utility records for Feeder B |
| Table 15 | Updated confusion matrix that compares model prediction and updated ground truth for Feeder A |
| Table 16 | Updated confusion matrix that compares model prediction and updated ground truth for Feeder B |
| Table 17 | Phase identification model accuracy rate by transformer size |
| Table 18 | Phase identification model accuracy rate by number of sample months |
| Table 19 | Confusion matrix that compares 10-min model prediction against utility records for Feeder A |
| Table 20 | Confusion matrix that compares 10-min model prediction against utility company's records for Feeder B |
| Table 21 | Confusion matrix that compares 10-min model prediction against 5-min model prediction for Feeder A |
| Table 22 | Confusion matrix that compares 10-min model prediction against 5-min model prediction for Feeder B |
| Table 23 | Phase identification model accuracy rate by consistency level |
| Table 24 | Phase identification model accuracy rate by distance from hybrid index cutoff point |
| Table 25 | Meter-to-transformer accuracy rate |
| Table 26 | Meter-to-transformer accuracy rate by transformer size |
| Table 27 | Meter-to-transformer accuracy rate by number of sample months |

## Part II List of Acronyms

| Acronym | Acronym Description |
|---|---|
| AMI | Advanced Metering Infrastructure |
| DER | Distributed Energy Resources |
| EPIC | Electric Program Investment Charge |
| EV | Electric Vehicle |
| GIS | Geographical Information System |
| GMSV | Google Maps Street View |
| L-L | Line to Line (phasing) |
| L-N | Line to Neutral (phasing) |
| MDMS | Meter Data Management System |
| Phase ID | Phase Identification (meter to phase connectivity) |
| RD&D | Research, Development and Demonstration |
| SaaS | Software as a Service |
| SCADA | Supervisor Control and Data Acquisition |
| SDG&E | San Diego Gas and Electric Company |
| TD&D | Technology Demonstration and Deployment |
| VRTU | Voltage at Remote Terminal Unit |

# 1.0 Overview

Methodology A demonstrated use of an established, data analytics platform to ingest, analyze, and evaluate end point phase identification and meter-to-transformer mapping. The scope of this pre-commercial demonstration was divided into four distinct tasks:

1) Collect Data
2) Run phase identification
3) Run meter-to-transformer
4) Evaluate Accuracy with Field Verifications

## 1.1 Collect Data

The data used for both the phase identification and meter-to-transformer solutions consisted of voltage readings for two meters per transformer across the two feeders. There was also a supplemental equipment dataset containing location and address information for the meters and transformers within both feeders. The location dataset was used in the meter-to-transformer algorithm. It was also used to visualize and present the results of both phase identification and meter-to-transformer.

In addition to data collection, data transfer and clean-up had to be performed prior to running the phase identification and meter-to-transformer solutions. A fraction of meters had to be omitted from the analysis due to incomplete or corrupted data. Likewise, a fraction of time intervals also had to be excluded due to “frozen” reads during periods such as the daylight savings days in March and November. Location data also required a clean-up step as the latitude and longitude data for meters was known to be inaccurate in many cases. Where possible, accurate geolocation data was extracted from meter addresses. For most of the meters this sufficed, however a fraction of meters had addresses that could not be accurately located on a map or were located outside of the extent of the feeder. These meters were also omitted from the meter-to-transformer analysis.

## 1.2 Run Phase Identification

The phase identification solution was run on more than 1,500 meters from the two selected SDG&E feeders. The intended outcome was that each meter would be assigned to one of three phases. This assumed that, on a given feeder, meters would be predominantly connected to either L-N phases (A, B, and C) or L-L phases (AB, BC, and AC), but not both. Initial results from field verification indicated that this was not the case, and so the assumption was removed for the final run. In the final run, the phase identification solution assigned each meter to one of six possible phases (A, B, C, AB, BC, AC) for the given feeder.

Phase predictions for each meter were accompanied by various descriptive metrics including metrics that reflected the confidence of the algorithm. Phase predictions were presented in tables and on interactive maps and charts.

## 1.3 Run Meter-to-Transformer

The meter-to-transformer solution was also run on over 1,500 meters. Typically, the results from this meter-to-transformer solution falls into two categories. The first category of results is referred to as the *exception views*. Each exception view is a table containing meters or transformers for which no meter-to-transformer prediction can be made. The list of typical exception views is provided below:

1) Transformers with 0 connected meters
2) Transformers with 1 connected meter
3) Transformers with 2 connected meters
4) Transformers with poor internal correlation
5) Transformers with bad latitude/longitude data
6) Meters on transformers with 1 connected meter
7) Meters on transformers with 2 connected meters
8) Meters on transformers with poor internal correlation
9) Meters with bad latitude/longitude data

For mathematical reasons, voltage correlation analysis cannot accurately correct errors in any of these situations without introducing new errors (Error Correction Code, 2021). A more in-depth discussion of the relevant mathematics can be found in Part II, Section 4.1 Findings Discussion. The exception views are provided so that stakeholders are aware of the portions of the connectivity model that cannot be verified using voltage correlation analysis.

The second category of results are the meter-to-transformer predictions. Each meter not included in an exception view gets assigned to the transformer that the algorithm selects as the best match. In cases where the prediction of the algorithm disagrees with the prior connectivity model, the prediction is often referred to as a meter-to-transformer “suggestion”.

Due to the nature of the voltage dataset collected, every single meter should fall into Exception View 7, “Meters on transformers with 2 connected meters”. The dataset also contained transformers that had voltage readings for a single connected meter and some transformers that had voltage readings for three connected meters. Nonetheless, under normal conditions, very few predictions would be made if the meter-to-transformer solution were run on this dataset. To best answer the question, “Can voltage data for two meters per transformer be used to accurately predict and correct meter-to-transformer connectivity on a given feeder?”, the solution was reconfigured to give predictions in the case where there were two or more meters on a transformer in the prior connectivity model.

The final exception views and predictions were presented in tables and accompanied by interactive maps and charts.

## 1.4 Evaluate Accuracy with Field Verifications

To evaluate the success of the demonstrations of meter-to-transformer and phase identification, the accuracy of predictions from each solution had to be measured. To evaluate accuracy, a 100% accurate source of truth must be established to measure against. Once the source of truth was established, it was contrasted with the phase identification and meter-to-transformer predictions in a confusion matrix (Tyagi, 2021), a tool for predictive analysis in machine learning. The confusion matrix provides information about how a machine classifier has performed, matching suitably classified examples corresponding to misclassified examples.

The formula for accuracy is:

$$\frac{Correct\ Predictions}{Total\ Predictions} = Accuracy$$

The source of truth selected for this demonstration was field verification. This field verification was performed in advance of running phase and meter-to-transformer identification, but the field verification results were not disclosed until after the initial runs.

After the initial runs were conducted, time was allotted to adjust or reconfigure the phase identification and meter-to-transformer solutions, if it was believed that a significantly better accuracy could be achieved. The phase identification and meter-to-transformer predictions after adjustment were used to calculate the final accuracy values.

While analyzing the results from this demonstration, it was discovered that the field verifications were not 100% accurate, as was believed initially. This had a significant impact on the interpretation of the final accuracy values.

# 2.0 Methodology Approach

## 2.1 Initial Selection of Metrics

The metrics used to determine success or failure for both phase and meter-to-transformer identification are the accuracy rates. There is no industry standard cutoff for success and failure accuracy rates. Higher accuracy rates are generally preferred, and rates of maximum trust are required for issues involving safety.

As mentioned above, many utilities will not update their system records when phasing and meter-to-transformer mapping are verified using an algorithm. In this case, meter-to-transformer and phase identification models can still help reduce field visits and increase utility company's records quality, but with only limited help. When model prediction has a higher accuracy rate than the expected accuracy of utility records, model results should be used to replace the utility records.

Here is an example that, although oversimplified, provides some insight into the selection of model prediction vs field verification. Assume that for one feeder the utility company has only 50% confidence in their records, and that the utility is capable of field results with a modest 95% accuracy rate. Also, assume that the machine learning model has a modest 90% accuracy rate. In this situation, one would expect that

out of the 1,000 meters on the feeder, the utility records are wrong for 500, and the model predictions are wrong for 100, all randomly distributed as in Table 1 below.

*Table 1. Example One*

| | False | True | Total |
|---|---|---|---|
| **False** | 50 | 50 | 100 |
| **True** | 450 | 450 | 900 |
| **Total** | 500 | 500 | 1,000 |

Without using the model, the utility could field verify every meter on the feeder and improve the overall accuracy of the connectivity model to 95%. However, by using the machine learning model they could achieve an even higher accuracy with almost half the work. Also assume that when both records and predictions are wrong, the chance that they are the same is one third. In the context of meter-to-transformer this would mean that a meter was incorrectly found on the same transformer by both the existing utility records and the machine learning algorithm. In the context of phase identification this would mean the meter was incorrectly found on the same phase.

By using the algorithm, the utility would not have to visit all 1,000 meters, but could prioritize to visit all the unmatched meters, which is 450 + 50 + 2/3 * 50 = 533 and with 95% accuracy rate, get correct records for 506 meters.

Therefore, after a round of field verifications, the records would still be incorrect for 17 of the meters in the top-left cell of Table 1, for which the model prediction matches utility's wrong records, and the 27 meters for which field verification was inaccurate. The updated accuracy rate on this feeder would then be 96%. That is higher than the field verification accuracy rate, with only 533 field verifications as opposed to 1,000.

The utility could iterate this process and visit the unmatched meters for the second round and increase the accuracy rate even more. In this case, the model helps setting target on the right meters for field visits, but since model accuracy is lower than the field verification accuracy rate, field verifications are still required to achieve maximum accuracy.

If the model has a higher accuracy than field verification, such as 99%, the story changes.

*Table 2. Example Two*

| | False | True | Total |
|---|---|---|---|
| **False** | 5 | 5 | 10 |
| **True** | 495 | 495 | 990 |
| **Total** | 500 | 500 | 1,000 |

Following the same logic, the utility would have to visit only 503 meters if they wished to field verify the discrepancies. Compared to example one, the number of meters visited decreased by only 6%. This is

caused by the assumption the utility has 50% confidence in their existing connectivity model. In such a case, even with a 100% correct back-office solution, to field validate the unmatched meters the utility would have to send out workers to 500 meters.

In this example field verification would get correct records for 478 meters, a 95% accuracy rate. The utility would end up with wrong records for two meters out of the top left cell of the matrix above, and 25 meters for which field verification fails to provide a correct answer. Overall, the accuracy rate for utility's records is 97%, higher than the field verification accuracy rate, but lower than the model prediction accuracy rate. Including the field verifications actually *reduced* the overall accuracy of the system. When the accuracy of a model is significantly greater than the accuracy of field verifications, the need for truck rolls is removed entirely. It makes the most sense to use model prediction as the single method to update the records.

## 2.2 Description of Pre-Commercial Demonstration

The purpose of this project is to assess analytical approaches to phase identification and meter-to-transformer mapping to enhance utility system operations and thereby improve the customer experience, in terms of reliability, safety and costs.

### 2.2.1 Use Case Description

The two use cases that were executed in this project were phase identification and meter-to-transformer mapping. Both use cases were executed using voltage correlation analysis.

*Phase Identification Use Case*

The use case of phase identification involves making predictions and corrections for the meter to phase connectivity within a feeder. On a feeder, electricity is typically distributed using three powered lines. Each line has a different phase of alternating current, each separated from the other two by 120°. Often these three phases are labeled A, B, and C. In between the powered distribution lines and residential electric meters, transformers are used to reduce voltages to safe levels. There are many ways to wire transformers between the power distribution lines. The result is the low voltage wires coming from a single-phase transformer can transmit electricity in one of six possible phases (A, B, C, AB, BC, AC), depending on the wiring configuration of the transformer. These phases are split into two groups. The L-N phases occur when the transformer is wired between a powered distribution line and a neutral line. They conduct electricity with a phase corresponding to the phase of the powered line. The L-L phases occur when the transformer is wired between two powered distribution lines. They conduct electricity with a phase corresponding to the difference between the two powered distribution lines. Technically speaking each phase also has an inverse phase (i.e. -A, -B, -C, etc.), but in this project there is no need to distinguish between phase A and phase -A for example, because a load on either phase will place a load on the phase A distribution line. Utilities typically keep track of the transformer to phase connectivity, because all the meters connected to a single-phase transformer share the same phase. For this use case, however, meter-to-phase connectivity is predicted. The primary reason for this is the meter-to-transformer connectivity is also in question. Accurate meter-to-phase connectivity is sufficient for use in phase balancing. Meter-to-phase connectivity can also be used to cross-validate meter-to-transformer

connectivity. Another reason that meter-to-phase connectivity is predicted, and not transformer-to-phase connectivity, is that voltages are not metered on the transformers.

### *Meter-to-Transformer Use Case*

The use case of meter-to-transformer mapping involves making predictions and corrections for the meter-to-transformer connectivity within a utility territory. Transformers typically have anywhere from one to dozens of connected meters. Each transformer typically serves a parcel of land within which it resides. Accurate meter-to-transformer connections are plotted on a map represented by lines connecting meters to the appropriate transformers and displayed as a collection of starburst patterns with very few crossing lines. Meter-to-transformer connectivity errors often manifest on the borders between two transformer "domains". For example, imagine two transformers serving meters on the same suburban street. The arrangement of meter-to-transformer connections might look something like the graphic in Figure 1.

*Figure 1. Typical meter-to-transformer connectivity on a suburban street*

In this example the meters most likely connected to the wrong transformer are the four in the middle. This is because they lie on the border between two transformers. If a tree falls and some of the wiring on this street must be redone, it is quite possible they could be rewired to a different transformer. Errors like these are difficult to find. Distance is not a good metric to use because both transformers are close enough to be possible. Street addresses are also insufficient for the same reason. In cases like these, voltage correlation analysis shines as a means of uncovering the correct connectivity. By combining voltage reads from the meters, a robust estimate for transformer voltage is constructed for each transformer. After this, the voltage correlation between each meter and each transformer is calculated. Meters tend to correlate to the transformer they are connected to. There are, however, mathematical limitations to this technique. To create a robust estimate for transformer voltage a sufficient number of connected meters is necessary. With three connected meters it is possible to correct a single error without introducing more errors into the connectivity model. With more meters it becomes possible to fix two or more mistakes on a transformer. At two meters per transformer, it is possible to detect the presence of a single error, but it is not possible to correct that error without introducing more errors into the system. Error correction with voltage data for only two meters per transformer was attempted in this

project. Transformers with fewer than two meters were listed in exception views, along with meters and transformers that had bad location data.

### 2.2.2 Software Requirements

The meter-to-transformer and phase identification solutions demonstrated are commercially available proprietary software, each with patents pending.

### 2.2.3 Supporting SDG&E Infrastructure and Data Requirements

Table 3 below compares data needed for the analysis and demonstration purposes, side-by-side with the data provided by SDG&E.

For the analysis purpose, the most important data is meter voltage interval data, which is essential for both phase identification and meter-to-transformer use cases. SDG&E provided two-year voltage data with 0.15 volts precision on five-minute intervals. For most of the meters in the sample, the voltage data covers the entire period. In cases where voltage data is not available, such as during an outage, or during the one hour lost because of the daylight-saving time change, the voltage data appears “frozen”. The system automatically interpolates the gap by constructing a linear line between the start and end of the period. When plotted, the interpolated gap period is shown as an artificial straight line, in the middle of curvy and random ups and downs, as if it is frozen. More details on the correction of the frozen periods are discussed in Section 2.3.

Metadata provides information on location, connectivity, and more. It first serves as a list of sample meters. For each feeder in the study, the metadata lists the meters under the feeder and limits the scope of analysis. SDG&E’s list is of high quality and very clean. SDG&E also added some “not real” meters into the study, which might mimic the potential of labeling meters under the wrong feeder. More details on excluding possible mislabeled meters are discussed in Section 2.3.

Metadata also provides latitude and longitude for each meter, transformer, and feeder. The information is important for meter-to-transformer and visualization purposes. SDG&E’s latitude and longitude data works for transformers but is not very accurate for the meters. For some meters, the latitude and longitude are the same as the linked transformers. For most of the meters, the problem is solved by converting meter addresses from metadata to GIS data. However, for a small proportion of meters, the latitude and longitude seem questionable. This is more critical in the meter-to-transformer use case and is discussed in Section 2.3.

*Table 3. Compare Data Necessary and Data Provided*

| Purpose | Data Necessary | SDG&E Data Description | Data Quality |
|---|---|---|---|
| Analysis | One year AMI data that provides meter voltages at interval level | Two year five-min voltage data | Excellent |
| Analysis | Metadata showing which meters are under the experiment circuits | Metadata for the whole two circuits under study | Excellent |
| Analysis & Demo on Map | GIS data of each meter, and transformers | Latitude and Longitude for each meter and transformers | Good for transformers. Questionable for meters |
| Analysis & Demo on Map | GIS data of each meter, and transformers | Customer Information System (CIS) data that includes address for each meter | Good for meters |

### 2.2.4 Updated Metrics

The primary metric for this demonstration was accuracy. The accuracy calculation assumed that field verification was a 100% accurate means of assessing meter-to-transformer and meter to phase connectivity. However, while analyzing the results it was discovered that field verification accuracies for both phase identification and meter-to-transformer were found at rates lower than 100%. Field verification accuracy could only be tested using GMSV on sections of the feeder with overhead wiring. This process was time consuming and so it was only completed thoroughly for the meter-to-transformer and phase identification examples where there was a discrepancy between the voltage correlation result and the field verification result. There were 35 cases where phase identification results from voltage correlation analysis disagreed with the field verification results and were available for analysis by GMSV due to the presence of overhead wiring. Thirty-four of these were on Feeder A and one was on Feeder B. In every instance, the field verification was incorrect, and the result from voltage correlation analysis was either correct, or likely but unverifiable. GMSV analysis suggested the utility field verification accuracy for transformer to phase connectivity was 95% on Feeder A. For meter-to-transformer a similar GMSV analysis was conducted and in several instances the findings of the field verification result were incorrect.

Because the field verification results could no longer be assumed 100% accurate, conclusions about accuracy were made using the following assumptions about ground truth:

1) In cases where the field verifications agreed with the voltage correlation analysis, the result was assumed correct.
2) In cases where the GMSV analysis was conducted and presented to the utility, the results from GMSV analysis were assumed correct.
3) In all other cases the utility field verifications were assumed correct only for the purposes of calculating a lower bound for the voltage correlation accuracy.

For this reason, the final accuracies for phase identification are presented as greater than 98% for Feeder A and greater than 97% for Feeder B.

### 2.2.5 Execution of Demonstrations

The demonstrations of phase identification and meter-to-transformer were carried out in accordance with the original schedule.

For phase identification, the solution was modified after the initial results came back to correct the assumption that each feeder had only three phases. The final result used the modified solution which assumed that both feeders had meters on all six phases.

For meter-to-transformer, the original solution was modified prior to providing initial results to give predictions in the case where there were two meters per transformer. The initial results from the meter-to-transformer were also accepted as the final result.

### 2.2.6 Use Case Execution

*Phase Identification Use Case*

The purpose of phase identification is to identify each meter's phase configuration. However, since the meter voltages cannot be compared to a known line voltage of each phase, a more accurate name for the algorithm is phase clustering. The algorithm clusters all the meters on a feeder into groups, and each group is labeled as a separate phase.

The clustering is based on voltage correlations. When electricity is consumed at some point on the grid, the voltage starts to fluctuate, and meters of the same phase move in similar direction and similar level, and therefore their voltage correlation is higher. Once voltage correlations are calculated, then the next task is to cluster the meters into their phases.

This solution begins the clustering process by finding "kernels" for different phases. Kernels are groups of meters that are strongly representative of each phase. When kernels are defined, the other meters' phases are identified by comparing the correlation with kernels. Agglomerative Hierarchical Clustering was selected to group the meters. There are many clustering algorithms readily available in many computer programming languages. Agglomerative Hierarchical Clustering Algorithm was selected because it provides a useful means for selecting kernels.

After an initial three kernels are constructed, a metric called the Hybrid Index is used to separate L-L meters and L-N meters. L-N phases are easier to distinguish than L-L because voltage differences are more pronounced between L-N phases. When there is a voltage change on phase A, both AB and AC are affected. The correlation's tendency is to squeeze together, and tangle with one another. One unique aspect of this solution is the separation of L-L meters and L-N meters, which makes it easier to cluster the L-L meters.

The last step in the phase identification solution used in this project was to iteratively identify kernels and rerun the correlations with those updated kernels as a starting point. With each iteration, the correlations become more accurate, and so are the kernels.

The algorithm is plotted in Figure 2 and Figure 3 below, where Figure 2 plots the steps for the analysis on a monthly basis, and Figure 3 begins with a summarization of all sample months and iterates using the summarization as a starting point.

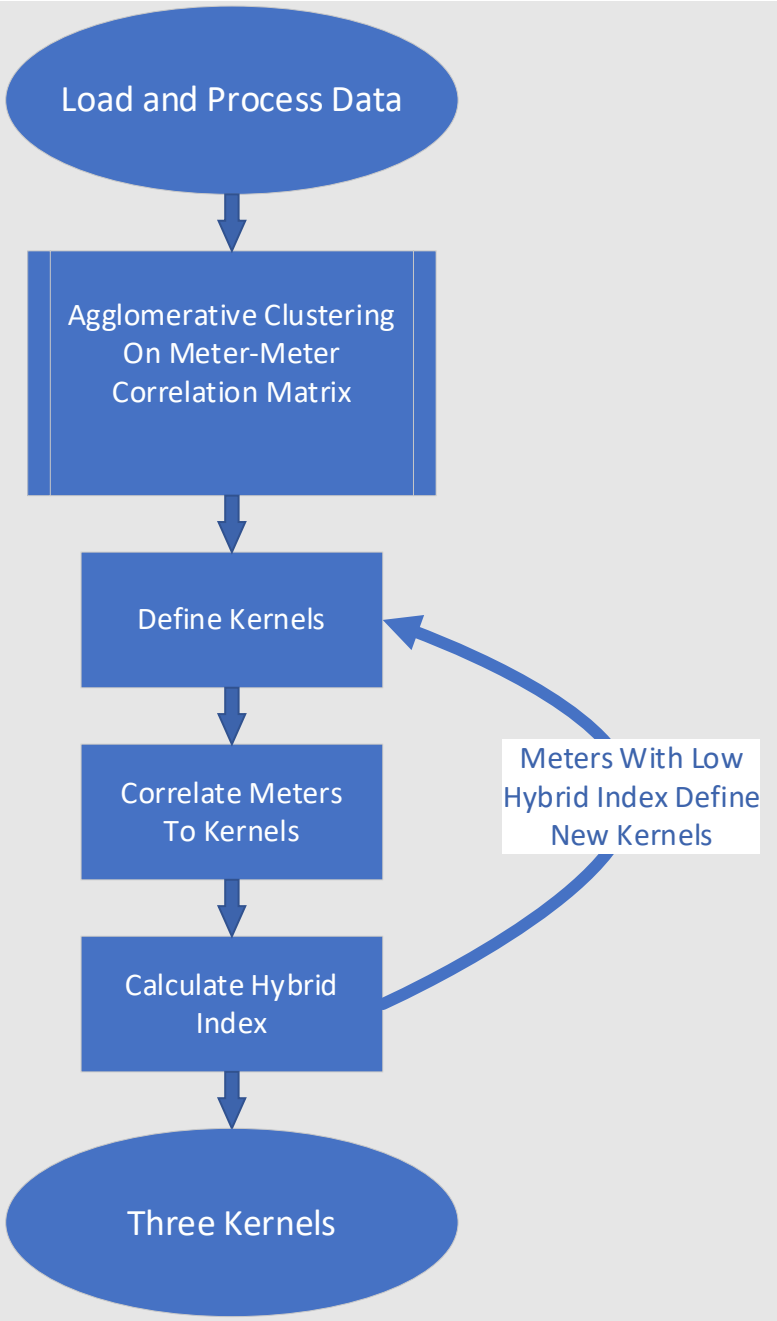



*Figure 2. Phase Identification Algorithm – Step 1 Flow Chart*

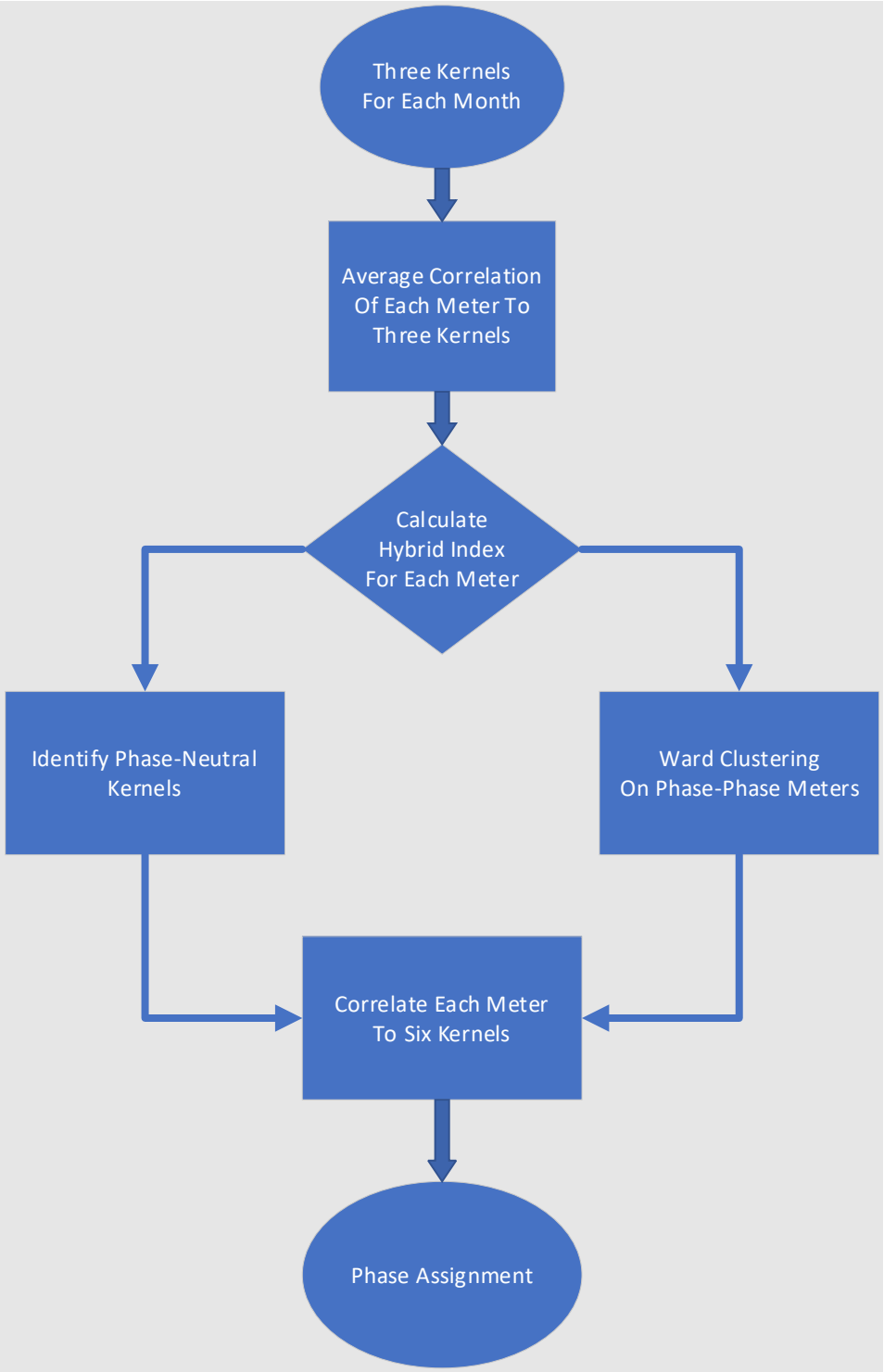



*Figure 3. Phase Identification Algorithm – Step 2 Flow Chart*

*Meter-to-Transformer Use Case*

The meter-to-transformer algorithm is performed meter by meter, transformer by transformer on the entire utility territory at once. The reason this does not lead to an extremely slow calculation is that the first step in the meter-to-transformer algorithm limits the number of potential transformers for each meter to the 15 closest ones. It also requires than any potential transformer is within 500 meters. These two numbers are default settings which can be adjustable by the utility if desired. For this demonstration the default values were used.

The second step in the solution is estimating a voltage-time series for each transformer. These estimates are created using the voltage readings for each connected meter in the existing connectivity model. If no existing connectivity model exists, an initial model is constructed by assigning each meter to the nearest transformer. For these reasons, accurate equipment location data is required. The algorithm assumes the model had greater than 50% accuracy, and that errors are uniformly distributed across the territory. In previous work, the vendor found that constructing an initial connectivity model by assigning each meter to the nearest transformer leads to accuracies of approximately 65%. For this project an initial connectivity model was provided with an estimated accuracy near 100%. Estimates for transformer voltages are made using a robust measure for central tendency, so that in the presence of errors on less than 50% of the connected meters, the estimate for transformer voltage is still appropriate. The robustness of the estimate breaks down completely with only two meters connected to the transformer from which to make the estimate. There is no measure of central tendency that is robust to errors when a dataset consists of only two datapoints.

The final step involves correlating the voltage time series of each meter to the estimated voltage of the nearest 15 transformers within 500 meters. The highest correlation between the voltage time series data for the meters and estimated voltage of the transformers is the basis for prediction of meter-to-transformer connectivity.

## 2.3 Data Analysis

### 2.3.1 Data Acquisition

As depicted in Figure 4. Data Transferring Timeline, the data transforming process had several issues, and thanks to a quick response from the SDG&E team, all issues were solved right away.

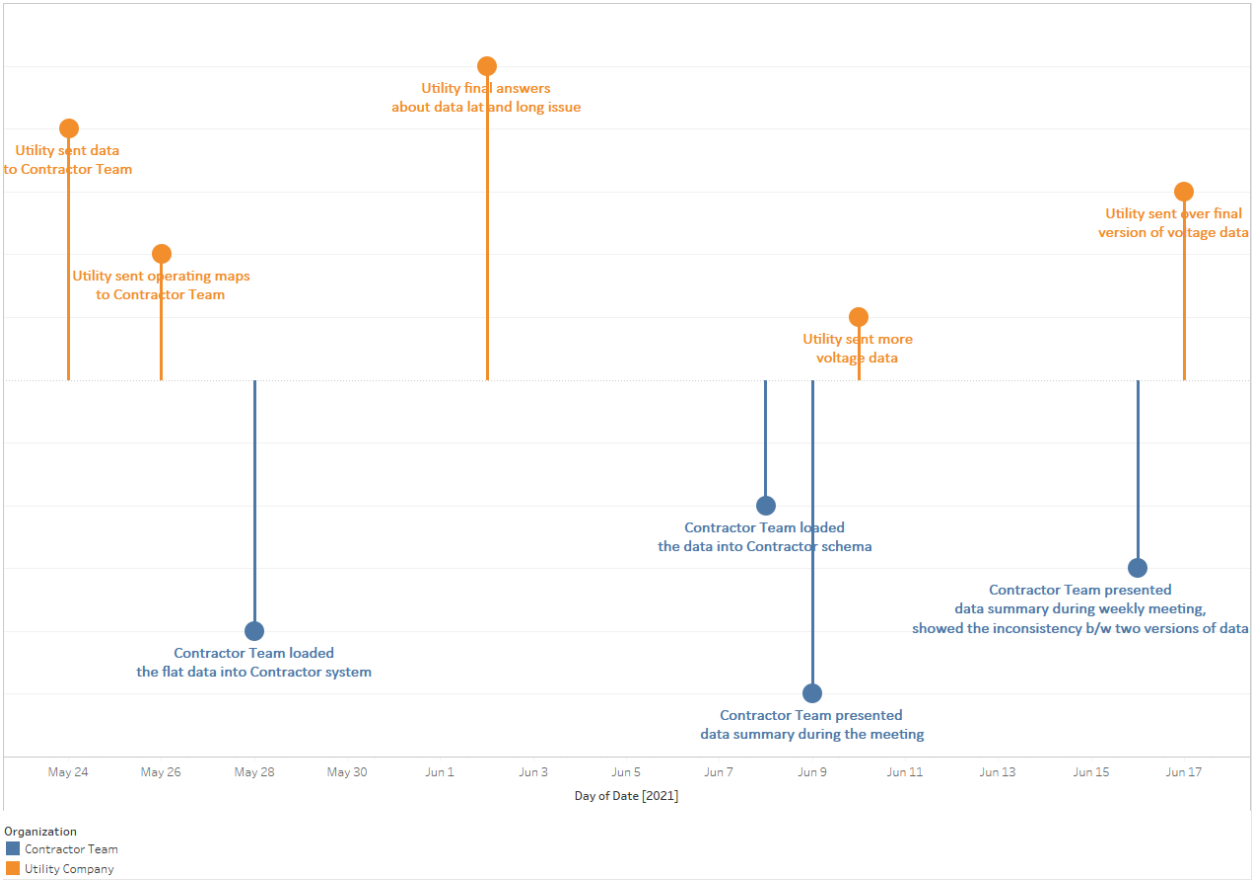



*Figure 4. Data Transferring Timeline*

The data from SDG&E includes:

- Bus voltage data. The bus voltage data contains the voltage level at the substation level, with irregular timestamp.
- Circuit data. The circuit data includes MW and MVAR readings at the circuit breaker, and the current reading by phase, with irregular timestamp.
- Switch SCADA devices. The switch data includes MW and MVAR at the switch level, along with current for three phases plus neutral phase, with irregular timestamp.
- Voltage at Remote Terminal Unit (VRTU) SCADA devices. VRTU data has voltage readings for each of the three phases, at the circuit level, with irregular timestamp.
- Tap position. This is the tap position data with irregular timestamp.
- Meter load data. Meter load data is the hourly kWh for each meter. It also provides tariff rate information.

- Transformer load data. Transformer load data is the hourly kWh consumption at the transformer level.
- Meter voltage data. Meter voltage data has five-minute voltage readings at the meter level, along with maximum voltage and minimum voltage across the five-minute interval.
- Metadata. Metadata is at the meter level, including:
    - GIS information, such as latitude, longitude, meter address, and zip code.
    - The transformer information that the meter is connected to, such as transformer ID, transformer latitude, transformer longitude, transformer KVA rating, and the circuit the transformer is connected to.
    - Meter connectivity date information, such as the date the meter was installed and removed.

The data needed in phase identification and meter-to-transformer algorithms are mainly in meter voltage data and metadata. Therefore, there was no need to clean up the duplication and missing data issues in the other data sources.

### 2.3.2 Introduction to the Two Feeders Under Study

For this analysis, SDG&E provided data for two feeders, A and B. Figure 5 and 6 below show maps of the two feeders. The red plus signs represent transformers, and the blue triangles represent meters. The size of the meters is proportional to the amount of data the meter has, which is a proxy of how much each meter contributes to the analysis.

SDG&E provided voltage data for only a subset of the meters. If a transformer has only one meter, the meter is selected; if a transformer has two or more meters, usually two meters are selected. For some larger transformers, where each links to 10, 20, or even more meters, it is also possible that three or four meters are selected. Out of the 974 transformers, 92 have more than two meters selected. The meters not selected are shown in the maps as small blue triangles. Concentrated small blue triangles indicate that area has many large transformers, each link to many meters, and a lot of the meters are not selected into this analysis.

The lines show which transformer connects to each meter. The very long lines that connect to off the chart points or across the whole map are due to bad latitude and longitude information.

Feeder A has 5,173 meters, much larger than B, but has only 325 transformers. On average, each transformer has 15.9 meters. The largest transformer is linked to 190 meters. In this area, phase balancing and meter-to-transformer accuracy are highly valuable. For example, if one transformer has 100 meters, and 10% of the connected customers have an EV, then when they all charge at a default charging time, the transformer will be under an extreme burden. Yet, with correct meter-to-transformer information the situation is avoidable by connecting EV rate meters to different transformers.

Feeder B has 649 transformers and 2,393 meters. On average, each transformer is linked to 3.7 meters. It can be seen from the map that the bigger transformers, where each linked to 10, 20, or up to 30 plus meters, are mostly concentrated at the southeast corner of the map. There are also some transformers at the top part of the map, or in the middle, that are linked to five to 10 meters. Most of the transformers that are spread out on the map are linked to less than five meters. There are 217 transformers that link to

only one meter, and 143 to two. For this area, driving from one location to another takes longer, checking one location verifies just one or two meters, and most of the power lines in this feeder area are underground. For these reasons, it is costly to do a field check on the transformer phase and/or meter-to-transformer, and more costly to check for technical problems during an outage period.

### 2.3.3 Data Description, Data Cleaning, and Data Trimming

*Partial Data*

As mentioned above, while the metadata included the whole frame of the two feeders, meter voltage data was provided for only a proportion of the feeder.

As shown in Table 4, Feeder B has 649 transformers and 2,393 meters. On average, each transformer is linked to 3.7 meters. Figure 6 shows for Feeder B, most of the transformers that are spread out on the map are linked to less than five meters.

Out of these 2,393 meters, 12 meters are not included in the metadata table. Since five-minute voltage data is available for these 12 meters, covering all 731 days of the two-year study period, these 12 meters are included in the sample, and assigned to a virtual transformer. These 12 meters were included in phase identification analysis but excluded from meter-to-transformer because their latitude and longitude information were missing.

*Table 4. Basic information for Feeders A and B*

| | A | B |
|---|---|---|
| # Meters | 5,173 | 2,393 |
| # Meters with Voltage Data | 695 | 1,031 |
| # Transformers | 325 | 649 |
| Avg # Meters per Transformer | 15.9 | 3.7 |

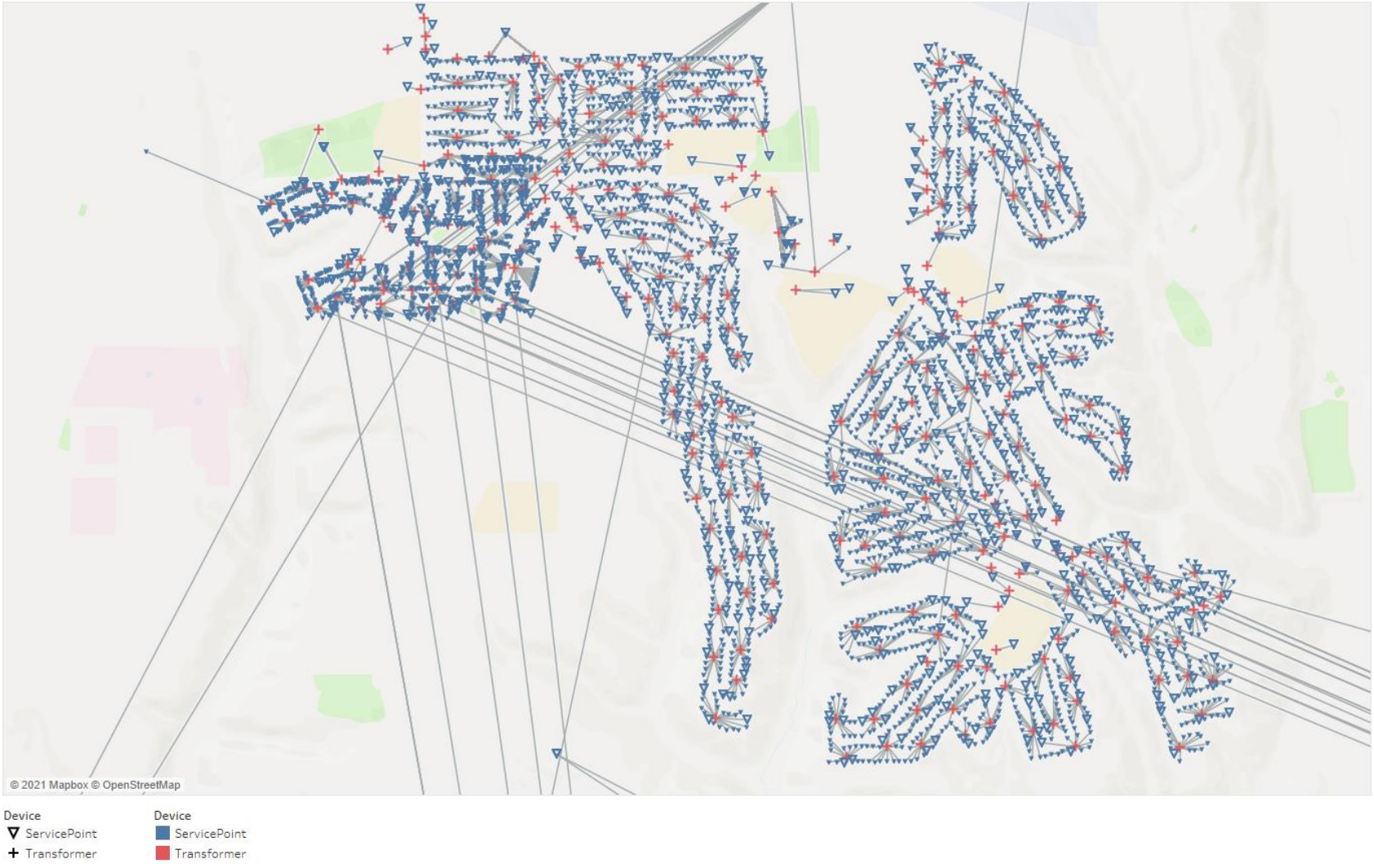



*Figure 5. Feeder A on Map*

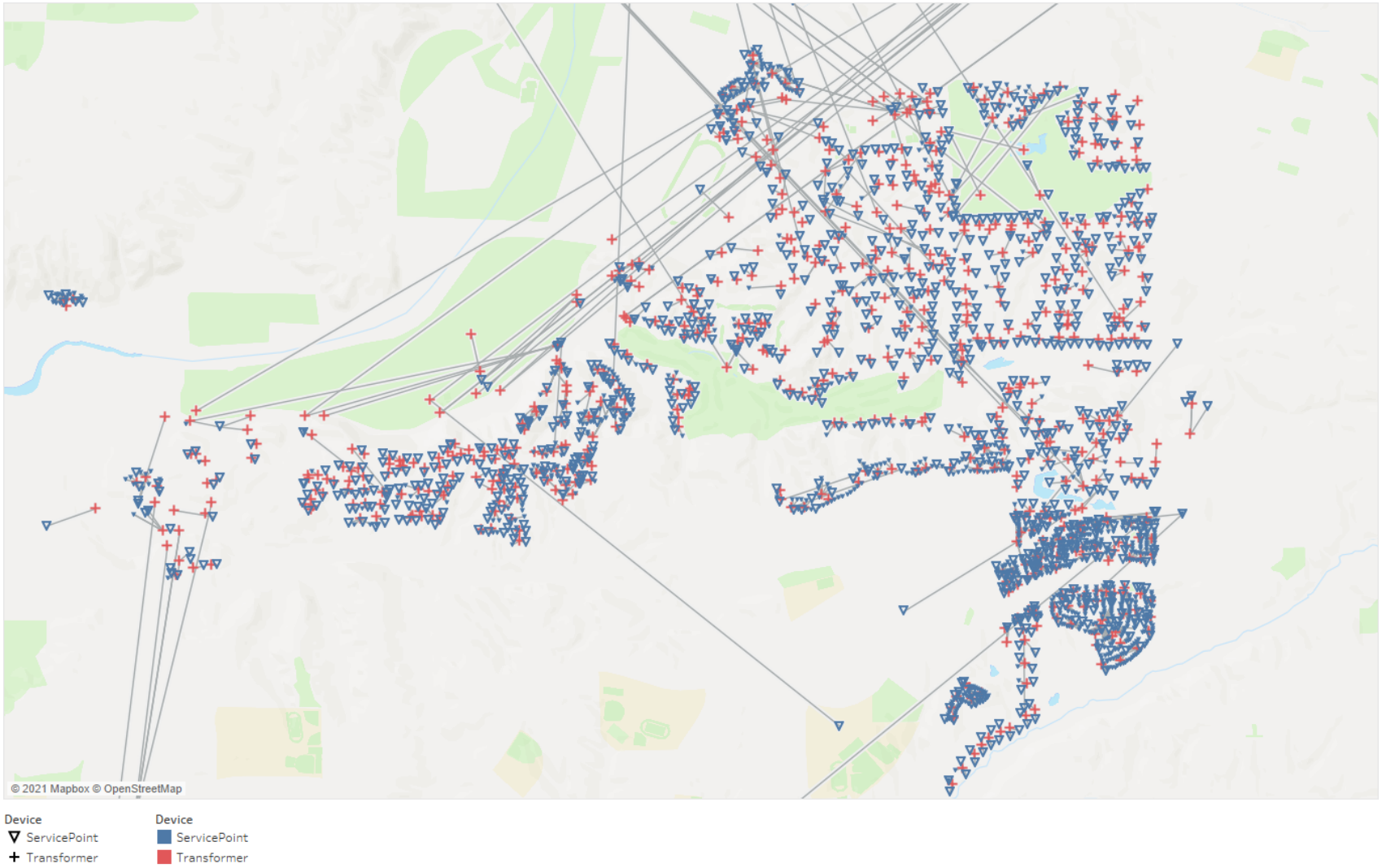



*Figure 6. Feeder B on Map*

Table 5 below lists detailed transformer distribution by size. More specifically, it shows the number of transformers by how many meters are linked to it. The 650$^{th}$ transformer, the virtual one, was excluded from this table along with the 12 meters "linked" to it. On column one, 217 transformers each link to only one meter, and 143 to two. That accounts for 55% of the transformers in Feeder B.

For this group of transformers, almost all meters linked to them were selected for the study, and five-minute meter level voltage data were available. The information is found in Table 6. Table 6 shows number of transformers by size in each row, and number of meters selected for the study for which voltage data is available in columns. The virtual transformer was excluded from this table. On Feeder B, for transformers with only one meter linked to them (as shown in row one), 209 out of the 217 transformers had one meter selected, for which voltage data was available. There were eight transformers, whose only meter was not selected, and hence was excluded from the analysis. Among the 143 transformers with two meters connected to them (row two in Table 6), there were 112 transformers for which all two meters were selected, 30 had one meter selected, and only one had zero meters selected. The larger transformers, with more than two meters linked to them, only had a proportion of meters selected. Out of the 289 transformers in this category, 222 had two meters selected into the study, or 77%; 46 transformers, or 16% had one meters selected, and 21, or 7% had more than two meters selected.

Due to the smaller transformer size in Feeder B, in terms of number of meters linked, a larger proportion of meters was selected into the study. Overall, 1,019 out of 2,381 meters were selected, which accounted for 43% of the meters in this feeder.

*Table 5. Distribution of Transformer by Size*

| # Meters | Feeder A | Feeder B | Total |
|---|---|---|---|
| 1 | 23 | 217 | 240 |
| 2 | 15 | 143 | 158 |
| 3 | 6 | 85 | 91 |
| 4 | 5 | 52 | 57 |
| 5 | 6 | 26 | 32 |
| (5, 10) | 56 | 80 | 136 |
| (10, 20) | 158 | 41 | 199 |
| (20, 30) | 26 | 3 | 29 |
| 30+ | 30 | 2 | 32 |
| **Total** | **325** | **649** | **974** |

*Table 6. Feeder B: Number of Transformers by Size and # Meters with Voltage Data*

| # Meters | Zero | One | Two | Three | Four | Total |
|---|---|---|---|---|---|---|
| 1 | 8 | 209 | 0 | 0 | 0 | 217 |
| 2 | 1 | 30 | 112 | 0 | 0 | 143 |
| 3 | 0 | 2 | 78 | 5 | 0 | 85 |
| 4 | 0 | 8 | 38 | 6 | 0 | 52 |
| 5 | 0 | 2 | 22 | 2 | 0 | 26 |
| (5, 10) | 0 | 19 | 59 | 2 | 0 | 80 |
| (10, 200 | 0 | 12 | 23 | 3 | 3 | 41 |
| (20, 30) | 0 | 2 | 1 | 0 | 0 | 3 |
| 30+ | 0 | 1 | 1 | 0 | 0 | 2 |
| **Total** | **9** | **285** | **334** | **18** | **3** | **649** |

Feeder A was very different than Feeder B. As shown in Table 4, Feeder A had 5,173 meters, many more meters than Feeder B, but had only 325 transformers, just half as many as the number of transformers on Feeder B. On average, each transformer on Feeder A had 15.9 meters. The largest transformer was linked to 190 meters.

Like Feeder B, there were seven meters on Feeder A that were not in the metadata table and were combined onto a virtual transformer. Again, these seven meters were included in phase identification analysis, but excluded from meter-to-transformer, because their latitude and longitude information were missing.

Table 5 above lists the number of transformers by transformer size group. The virtual transformer was excluded from this table. Recall that more than half of the transformers on Feeder B had one to two meters. Here in Feeder A, however, the group with the highest number of transformers was 10 to 20. There were 158 transformers that were linked to 10 to 20 meters, accounting for 49% of the total number of transformers. Fifty-five, or 17% of the transformers, had less than or equal to five meters, compared to 81% in the case of Feeder B. Fifty-six transformers, another 17% of the total transformers were linked to 20 or more meters, compared to 7.1% for Feeder B.

Table 5 clearly shows that Feeder A 's transformers were much larger in terms of the number of meters linked. Even though Feeder A had only one half as many transformers as Feeder B, it had more than two times as many meters compared to Feeder B. As shown in Table 6, for Feeder B, SDG&E tried to select two meters per transformer whenever possible, with slight adjustments here and there. Therefore, for Feeder A, the sample for analysis was expected to be smaller, in terms of absolute number of meters, and the proportion of the total meters.

Table 7 below is in the same format as Table 6. It shows the distribution of transformers by size, or how many meters linked, and number of meters selected into the analysis. On Feeder A, since most of the transformers had two meters or more, there were not as many transformers with only one meter

selected. Most of the transformers, 214 out of 325, or 66%, had two meters selected into the analysis. Forty transformers each had only one meter selected, including 23 that were linked to just one meter. There were more transformers with more than two meters selected. Sixty-four had three meters, and seven had four meters, accounting for 22% of the transformers, much more than on Feeder B where only 21 had more than two meters selected.

*Table 7. Feeder A: Number of Transformers by Size and # Meters with Voltage Data*

| # Meters | One | Two | Three | Four | Total |
|---|---|---|---|---|---|
| 1 | 23 | 0 | 0 | 0 | 23 |
| 2 | 8 | 7 | 0 | 0 | 15 |
| 3 | 3 | 3 | 0 | 0 | 6 |
| 4 | 0 | 5 | 0 | 0 | 5 |
| 5 | 1 | 5 | 0 | 0 | 6 |
| (5, 10) | 0 | 42 | 14 | 0 | 56 |
| (10, 20) | 1 | 107 | 43 | 7 | 158 |
| (20, 30) | 2 | 20 | 4 | 0 | 26 |
| 30+ | 2 | 25 | 3 | 0 | 30 |
| **Total** | **40** | **214** | **64** | **7** | **325** |

Comparing the two feeders in the analysis, Feeder B had fewer meters, but more transformers, and more meters were included in the analysis. Such a sample design does not affect the phase identification much because phase configuration is mainly at the transformer level and since all transformers are covered, all phases have good representation in the sample.

For the meter-to-transformer task, however, the situation was different. On average, there were 2.1 meters on each transformer for Feeder A and 1.6 meters on each transformer for Feeder B. Chances are, some of the transformers' selected meters were not a very good representation for the transformer, which may have impacted the results. Also, the meter-to-transformer model uses the median of the meter's voltage as proxy for the transformer's voltage. With only one or two meters on each transformer, finding reasonable proxy was a challenge. Furthermore, the quality of geo information of the meters further complicated the situation.

The effect of sample design on the model performance is discussed in more detail when introducing the algorithm for each task.

### *Latitude and Longitude Coordinates*

While latitude and longitude coordinates are not included in the phase identification algorithm, they are crucial for the meter-to-transformer model. To reduce the line loss from electricity transferring, and to reduce the length of service wire, a meter is always connected to the closest transformer whenever possible. When the length of the service wire from one meter to the closest transformer is not available,

the geo distance based on latitude and longitude coordinates is used as a proxy. The statistics show that more than 50% of the time, a meter is connected to the transformer with the closest geo coordinates.

As mentioned previously, SDG&E's metadata uses transformers' latitude and longitude as geographic coordinates for the linked meters, which made it impossible to calculate the distance between meters and transformers. Addresses were converted from customer information addresses to latitude and longitude for each meter. Most of the meters had reasonable latitude and longitude coordinates after geo conversion with a few remaining suspicious.

To ensure the validity of the latitude and longitude coordinates, and avoid introducing unnecessary noise into the analysis, the meter-to-transformer analysis included only the meters whose geo coordinates were recognizable by Google Map. Table 8 below shows the number of meters with valid geo information by feeder. Overall, 400 meters' geo coordinates were not recognizable, and among the meters with voltage data, 161 meters were in this category. Hence, these meters were excluded from the meter-to-transformer analysis.

*Table 8. Latitude and Longitude Validity*

| | A | B | Total |
|---|---|---|---|
| # Meters | 5,173 | 2,393 | 7,566 |
| # Meters with Geo Info | 4,960 | 2,206 | 7,166 |
| # Meters with Voltage Data | 695 | 1,031 | 1,726 |
| # Meters with Voltage Data and Geo Info | 651 | 914 | 1,565 |

### *Study Period and Time Range of Data*

For the selected meters in the two feeders under analysis, SDG&E provided five-minute interval voltage data that covered a two-years period, beginning October 21, 2018, to October 20, 2020. For more than 90% of the meters, the interval data covered 731 days of the study period. Table 9 below summarizes the number of meters by data completeness. Overall, 1,610 out of 1,726 meters had 731 days of data which is 93% of the sample.

Twelve meters have all data missing. This could be attributed to meter removal. Since this only accounts for less than 0.1% of the sample, no further investigation was warranted. Among the meters with only partial data available, most, or 76 of them, have roughly half a year of data. For all 76 meters, the data covers the last part of the analysis period. For 73 of the meters, the data ranges from April 23, 2020, to October 20, 2020. For the remaining three, the data begins in May or June, and all data ends on October 20, 2020. There are 60 meters on Feeder B and 16 meters on Feeder A.

The 10 meters with one to two years of data are on Feeder B. The data for these meters include random start and end dates with gaps for a few meters.

Eighteen meters have data gaps of less than one week. These meters were grouped as having complete data but are singled out to emphasize the fact that more than 93% of the meters have voltage data that perfectly covers the whole study period with no gap.

*Table 9. Time Range of the Voltage Data*

| | A | B | Total |
|---|---|---|---|
| No Valid Data | 1 | 11 | 12 |
| Up to Half Year of Data | 16 | 60 | 76 |
| One to Two Years of Data | 0 | 10 | 10 |
| Up to One Week Gap | 2 | 16 | 18 |
| Whole Study Period | 676 | 934 | 1,610 |
| **Total** | **695** | **1,031** | **1,726** |

Figure 7 below plots the number of meters changing over time. The upper section of the chart shows the number of meters on Feeder B, and the lower section shows the number of meters on Feeder A. The plot shows the same trend as described above where most of the meters have complete data across the whole period. The most obvious change occurred toward the last half of 2020 where seven, or 4.4% of the meters, were added into the sample. Since the plot for the number of transformers looked the same, the chart below is sufficient. Toward the last half of 2020, 4.2% of the transformers were added into the sample.

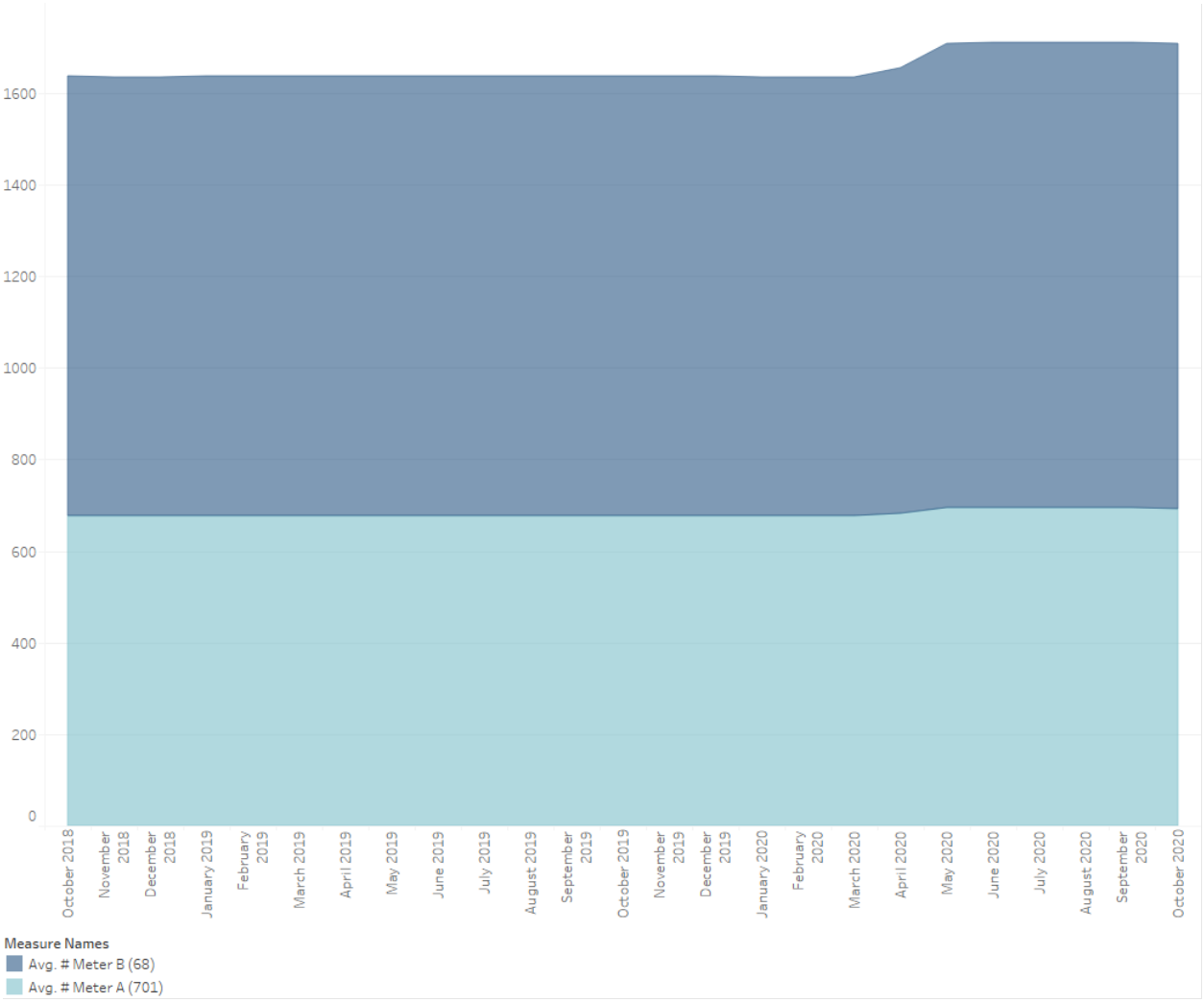



*Figure 7. Number of Meters with Voltage Data by Time*

## *Average Voltage and Data Interval*

Figure 8 below shows the distribution of average voltage by meter. The chart on the top left is a traditional histogram plot of the 1,714 meters with any associated voltage data. Most of the meters are in the range of 240 volts, a small proportion of meters are in the range of 120 volts, and another small proportion of meters, 720 volts.

The overwhelming proportion of 240-volt meters skews the distribution of the other groups. Therefore, the same histogram is plotted using log form y-axis, to flatten the 240-volt group which is two to three magnitudes higher than the other groups.

Similar charts are plotted on the second row of Figure 8. It is clear in the bottom two charts that Feeder A and Feeder B both have a few meters in a voltage range that are unexpected.

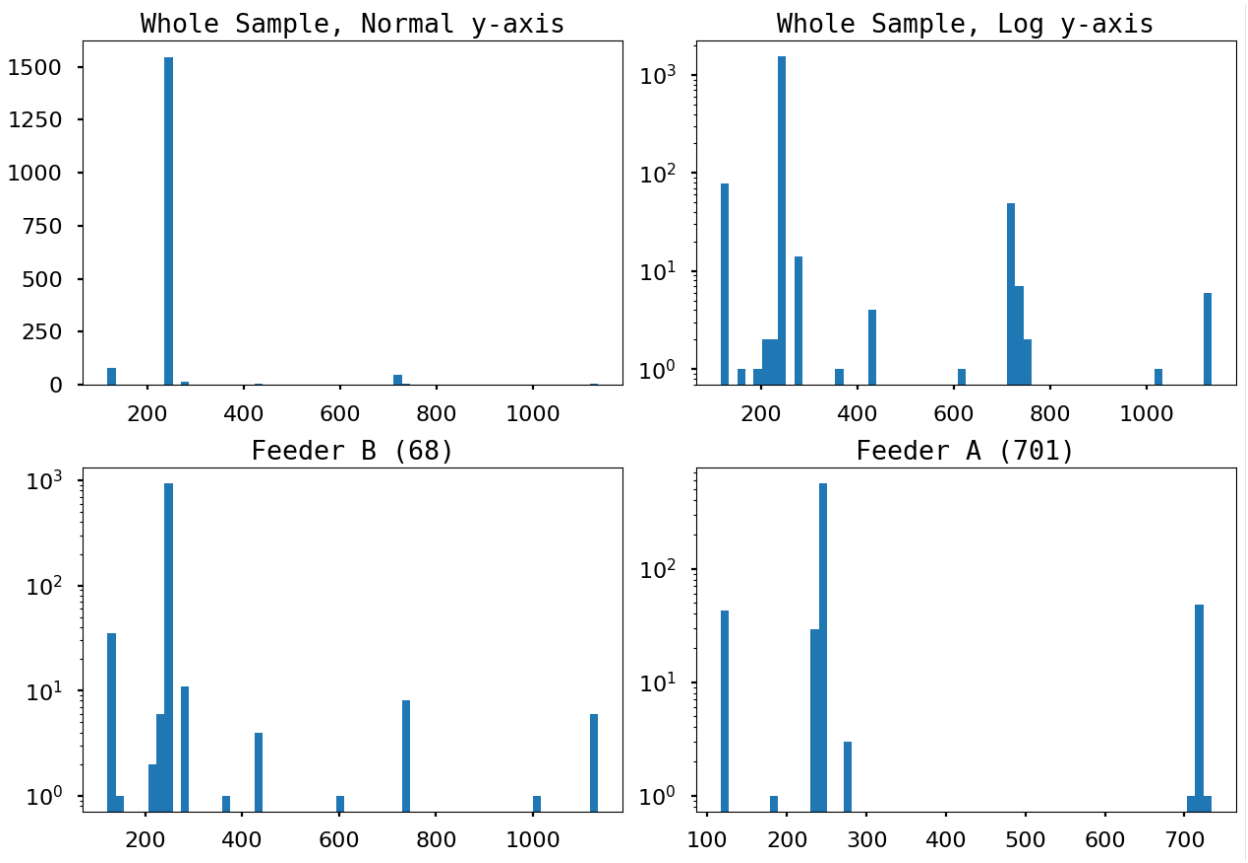



*Figure 8. Average Voltage Histograms*

It appears that the meters are sending data at different frequencies. For example, if a meter's data is not at five-minute intervals, but 15-minute instead, its interval voltage is then about three times the value of a five-minute interval meter. Therefore, if a meter's average hourly voltage should be 240 volts, with 15-minute interval data, the average can be 720 volts. In cases where a meter's data frequency changed in the middle of the study period, the observed average voltage is somewhere between 240 and 720 volts.

These meters were excluded from the analysis because they had voltages missing for most of the timestamps, making the correlation not comparable with the correlations calculated between two pairs of meters with complete data. Since the analysis was performed at a monthly level, if a meter changed data frequency during the early part of the analysis period, it might still be included in the analysis in the later part of the period.

Another possible explanation is when some meters' voltage dropped to zero at some point, due to a data issue. This was the case for one meter in the sample whose voltage was about 240 volts, until it dropped to zero on May 1, 2020. Therefore, this meter was excluded from the analysis after May 1, 2020, but was included in the analysis period prior to the voltage decline.

Table 10 below lists the distribution of meters by average voltage groups. Feeder A has 52 meters that will potentially be excluded from the analysis, and Feeder B has 24. Overall, 69 meters might be excluded, accounting for 4% of the meters where the voltage data is available.

*Table 10. Distribution of Meters by Average Voltage Group*

| Avg Volt Group | A | B | Total |
|---|---|---|---|
| Below 120 | 1 | 0 | 1 |
| 120 | 42 | 35 | 77 |
| 120 to 240 | 1 | 3 | 4 |
| 240 | 597 | 950 | 1,547 |
| 277 | 3 | 11 | 14 |
| 277 to 720 | 0 | 6 | 6 |
| 720 | 50 | 8 | 58 |
| Above 720 | 0 | 7 | 7 |
| **Total** | **694** | **1,020** | **1,714** |

### *Frozen Period*

As mentioned in an earlier section, the voltage data may have some "frozen" periods, where the voltage is missing, but interpolated and represented as a linear line between the start and end points. Figure 9 and Figure 10 provide two examples of frozen periods. The first frozen period, as illustrated in Figure 9, starts on November 3, 2018, at 8:00 AM, and ends on November 4, 2018, at 7:55 AM (labeled using the gray vertical band). November 3, 2018, was the end of the daylight saving days in 2018. SDG&E stops reading voltage data during time changes; thus, the system draws a linear line between 241.49 volts, and 240.90 volts, the voltages at the two ends of the frozen period.

The upper portion of Figure 9 shows the voltage, with the lower portion showing delta voltage, or $\Delta Volt_t = Volt_t - Volt_{t-1}$. During the frozen period, voltage is plotted as a linear line. Hence, delta voltage appears to be a constant, and in this case $\Delta Volt = -0.001$.

These data points can adversely affect the correlation coefficient matrix. Consider the case where two meters are negatively correlated, and while one has voltage going up, the other's voltage is going down. One frozen period starts at the time where two meters' voltages are similar, even though one was in the middle of going up and the other going down; and ends at the time, coincidently, that the two meters' voltages are similar again. During the frozen period, the two meters' voltages appear to be similar, going in the same exact direction, and at the same exact level. Therefore, the correlation of the two meters is no longer negative, and in fact might increase significantly.

Of course, the frozen period may also cause the correlation between two meters to decrease. The effect is random, like noise in signals. Therefore, the data cleaning process is necessary to delete these periods, erase the noise, and emphasize the signals.

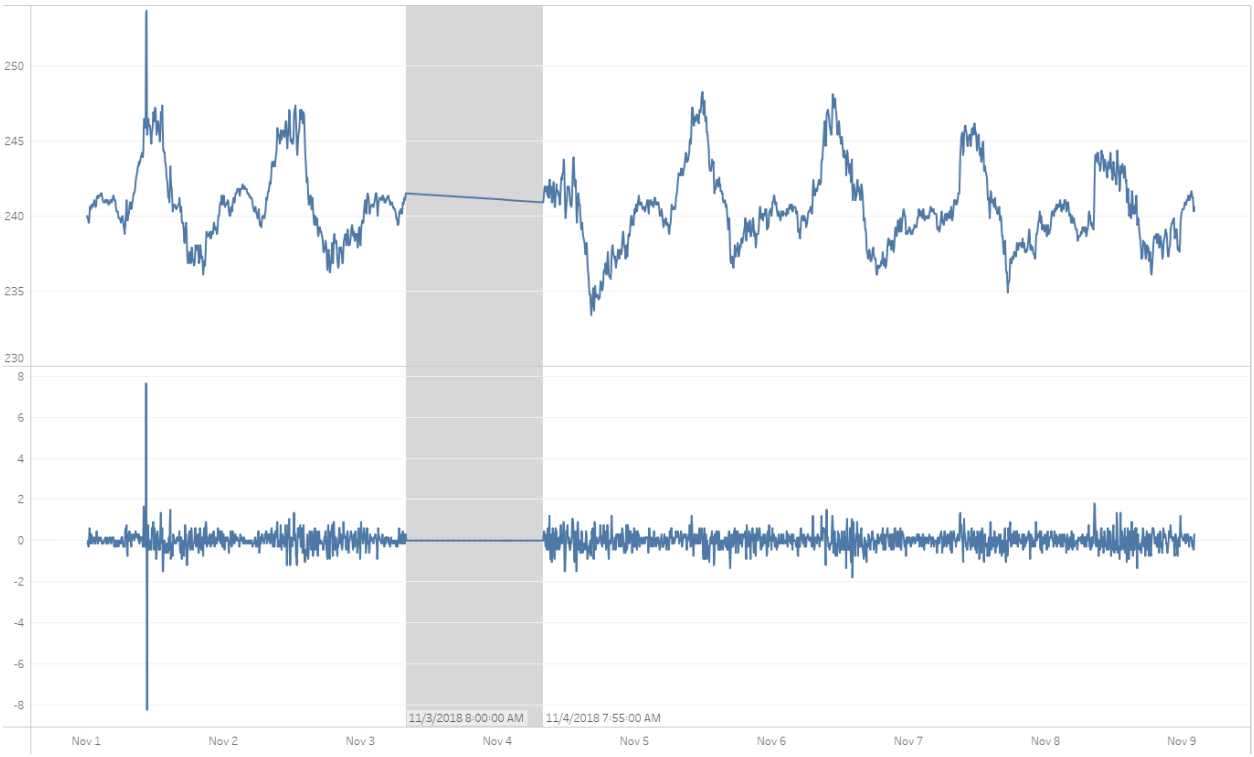



*Figure 9. Frozen Example ONE*

Figure 10 shows two frozen periods that are not due to a daylight-saving day time change. The time range for the frozen period on the left is from November 21, 2018, 8:00 AM to November 22, 2018, 7:55 AM. November 22, 2018, was Thanksgiving Day. This case serves as evidence that the frozen periods occur other than during daylight saving time changes.

The list below includes the dates with frozen data lasting longer than 24 hours:

- November 3 - 4, 2018 (daylight saving)
- November 21 - 22, 2018
- January 3 - 4, 2019
- January 18 - 19, 2019
- March 9 - 10, 2019 (daylight saving)
- November 2 - 3, 2019 (daylight saving)
- November 14 - 23, 2019
- March 7 - 8, 2020 (daylight saving)
- March 20 - 21, 2020
- March 27 - 28, 2020
- June 20 - 21, 2020
- July 10 - 11, 2020
- August 29 - 30, 2020

Figure 10 also shows another frozen period from November 23, 2018, 1:50 AM to 3:20 AM. The meter's voltage remains unchanged at 120 volts for one and a half hours. Examining if this is really a frozen period, the upper part of Figure 10 appears to have low resolution with voltage changes shown in steps rather than smooth lines. The lower part of Figure 10 confirms the voltage changes are in units of 0.15 volts, and $\Delta Volt$ takes the values of 0.15, 0.3 and 0.45, all multiplies of 0.15.

This is common in utilities' voltage data. In some cases, there is data changing in one voltage unit. This will also introduce measurement errors into the analysis, and the measurement error appears as white noise too. Since there are many meters with data like this, the analysis will not keep the data as-is and will not tackle the issue with minor effects. If the voltage remains unchanged for a period longer than one hour, the problem is treated as a frozen period, and the data will be dropped. SDG&E provides two years of data, enough for the analysis, irrespective of some intermittent data points.

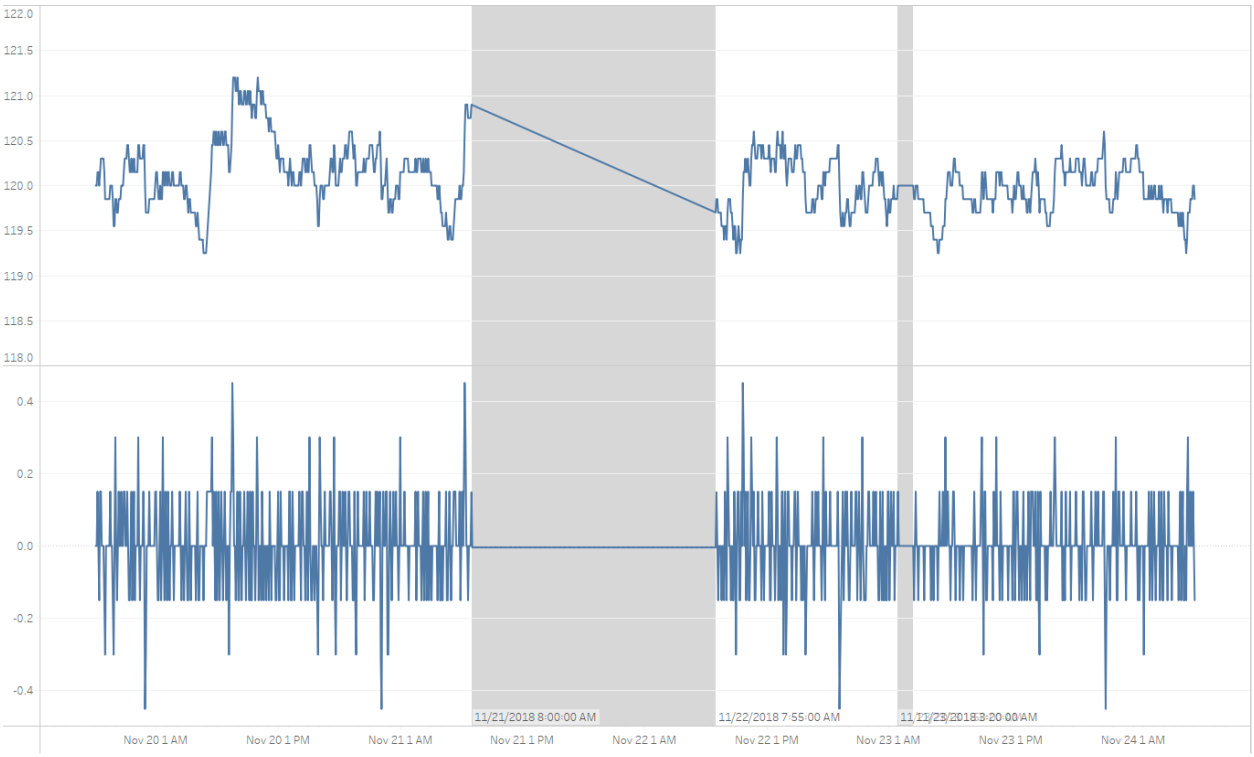



*Figure 10. Frozen Example TWO*

Figure 12 and Figure 11 below plot the data loss due to a frozen period, for Feeder A and Feeder B respectively. The top line, shown in aqua, is the number of meters in the raw data. This is calculated by taking the average number of meters with raw data for each interval to the monthly level, showing the same data as shown in Figure 7. The second line in green is the number of meters after excluding those with an abnormal voltage mean, as discussed previously.

The third line in red is the number of meters after deleting the frozen periods, where $\Delta Volt$ stays constant. The data volume dropped significantly in November 2018, January, March, and November of

2019, and again in March, June, July, and August of 2020. This aligns with the dates where frozen data lasts longer than 24 hours, as listed above. Overall, 3% of the data on Feeder A is dropped due to frozen period, and 4% for Feeder B.

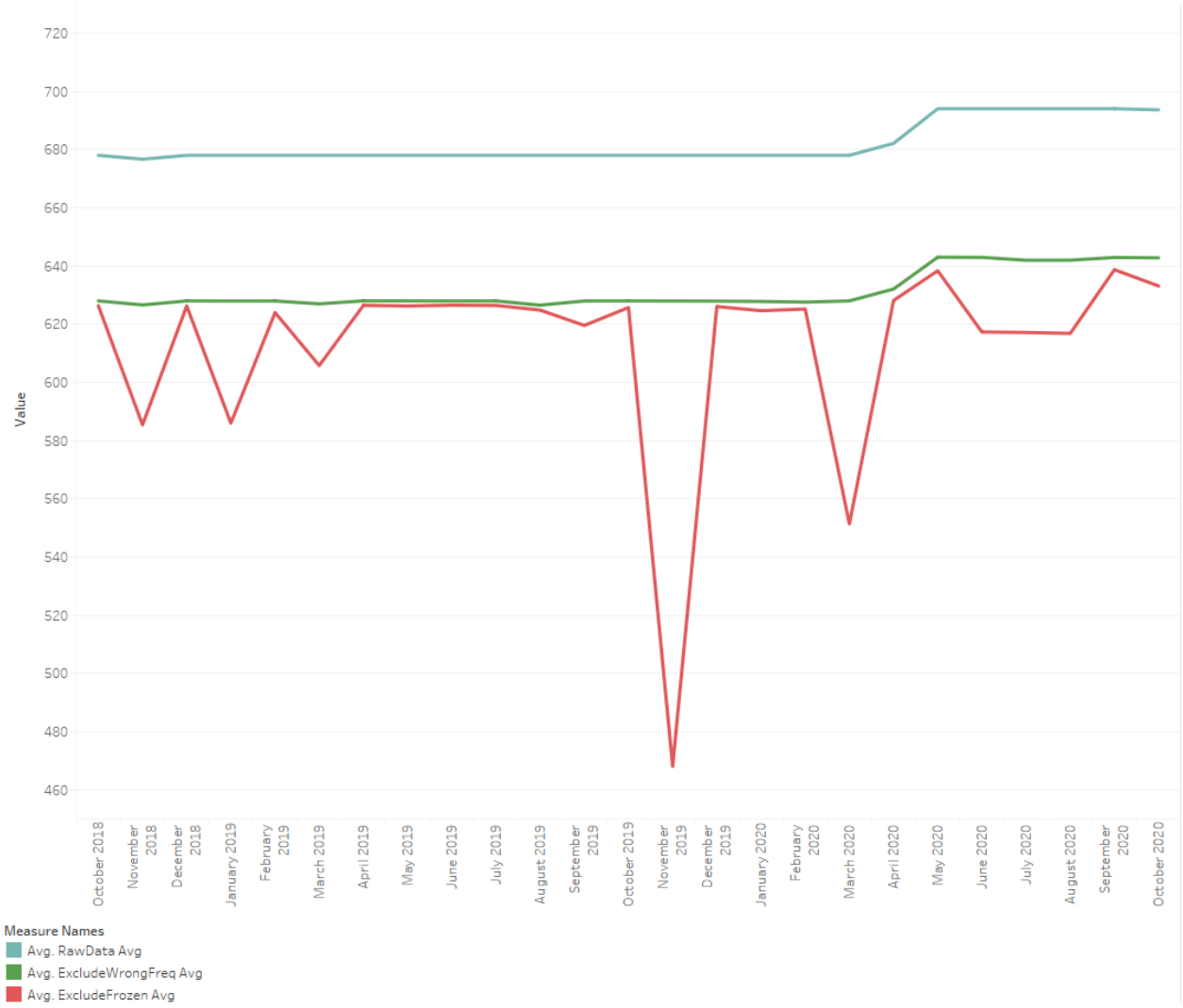



*Figure 11. Time Series of Frozen Period – Feeder A*


*Figure 12. Time Series of Frozen Period – Feeder B*

## *Jumps or Big Changes*

Another data issue to consider is the jump on November 2, 2018, shown in Figure 12. Voltage volatility is usually caused by consumers' activities on the grid. When consumption goes up, voltage goes down, and vice versa. Meters on the same transformer tend to have voltages moving in the same direction, and with similar scale. Meters of the same phase tend to have similar voltage movements as well. This is fundamental to the phase identification and meter-to-transformer algorithm. Utility activities can have a larger impact on the voltage than consumption activities do. Utility activities usually impact the whole feeder, with some significant operations that cause significantly larger impact to voltage volatility than activities attributed to customer consumption. This may adversely affect the correlation coefficient matrix dramatically.

Figure 13 below shows three meters' voltage from August 19, 2019, to August 21, 2019. There are two big jumps during this period. In the morning of 19th, at 9:00 AM, the voltage jumped up, and at about

9:00 PM, the voltage jumped down. Without the jump, the correlation between the meter shown in red and the other two meters are very low, about 20%; but the jumps increase the correlation to a 40% level.

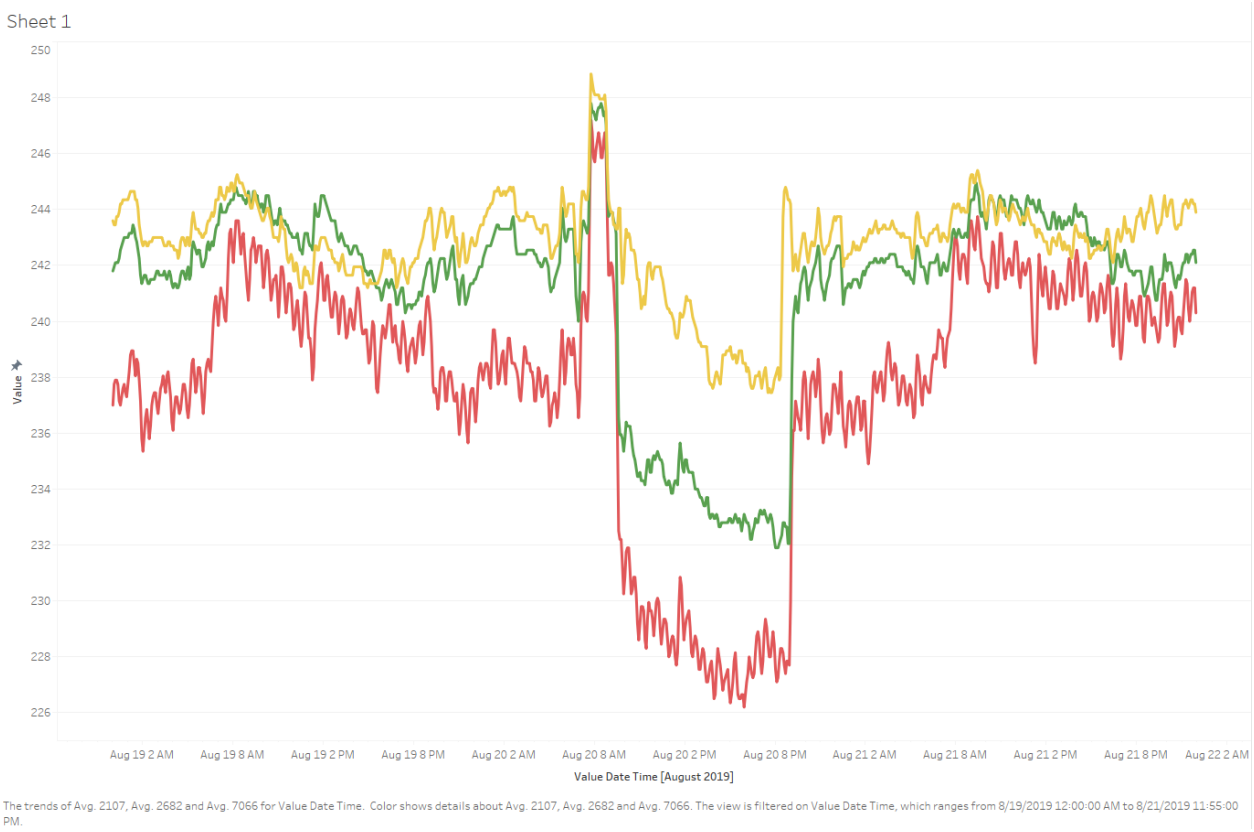



*Figure 13. Example for Jumps*

According to SDG&E, the utility's activities on the grid can change the voltages up to two or three times per day. There are 288 five-minute intervals each day. These activities can cause big jumps in up to 1% of the intervals. Therefore, the top one percentile jumps are dropped from the analysis. Such data cleaning steps are very likely to eliminate useful voltage volatility that is due to consumption activities, and hence cause loss of valuable information that contributes to the correlation among meters. Fortunately, due to SDG&E's abundant data sources, high quality data could be utilized for this analysis.

### 2.3.4 Data trimming

*Phase Identification*

Figure 14 and Figure 15 below plot the average amount of data for each sample month, by feeder, after each step of data trimming. From the raw data, the observations are excluded from the analysis because of the reasons listed below. Some of these data trimming steps drop intervals but not meters, some of the steps drop meters, and some do both.

If a meter loses some intervals, it may still have a prediction for the sample month. Even if the meter is dropped from the sample month, since the analysis is done monthly, it may still have results from the other months. These steps were used to cleanse/trim the data.

1) The meters have no valid interval voltage data. This step removes meters from all the sample months, and there is no prediction for these meters. The average number of meters is plotted in red for each sample month, in Figure 14 and Figure 15 for feeders A and B, respectively.
2) The records come in with wrong intervals, or the meter's average voltage is out of normal boundary. The "normal" boundary includes 1) 120 +/- 5%, 2) 240 +/- 5%, and 3) 277 +/- 5%. This step removes meters from sample months. The average number of meters is plotted in orange.
3) The records are from a period when voltage is frozen. This step drops intervals only. If a meter loses too many intervals, it may not have enough data, and hence is excluded from the sample month. The average number of meters is plotted in yellow.
4) The records have spikes that exceed the threshold, either up or down. This step drops intervals only. Again, if a meter loses too many intervals, it may get deleted from the sample month. The average number of meters is plotted in green.

For each sample month, the prediction is generated for all the meters that remain in the sample. The average number of meters is plotted in blue, which almost always coincides with the sample after Step 3, and thus the blue line is hidden behind the green line.

For both feeders, November 2019 is dropped from the analysis, mainly because the frozen period is too long, and hence none of the meters have enough data for the month. Again, thanks to the abundance of data provided by SDG&E, the analysis team can choose data quality over quantity and do not have to lower the criteria.

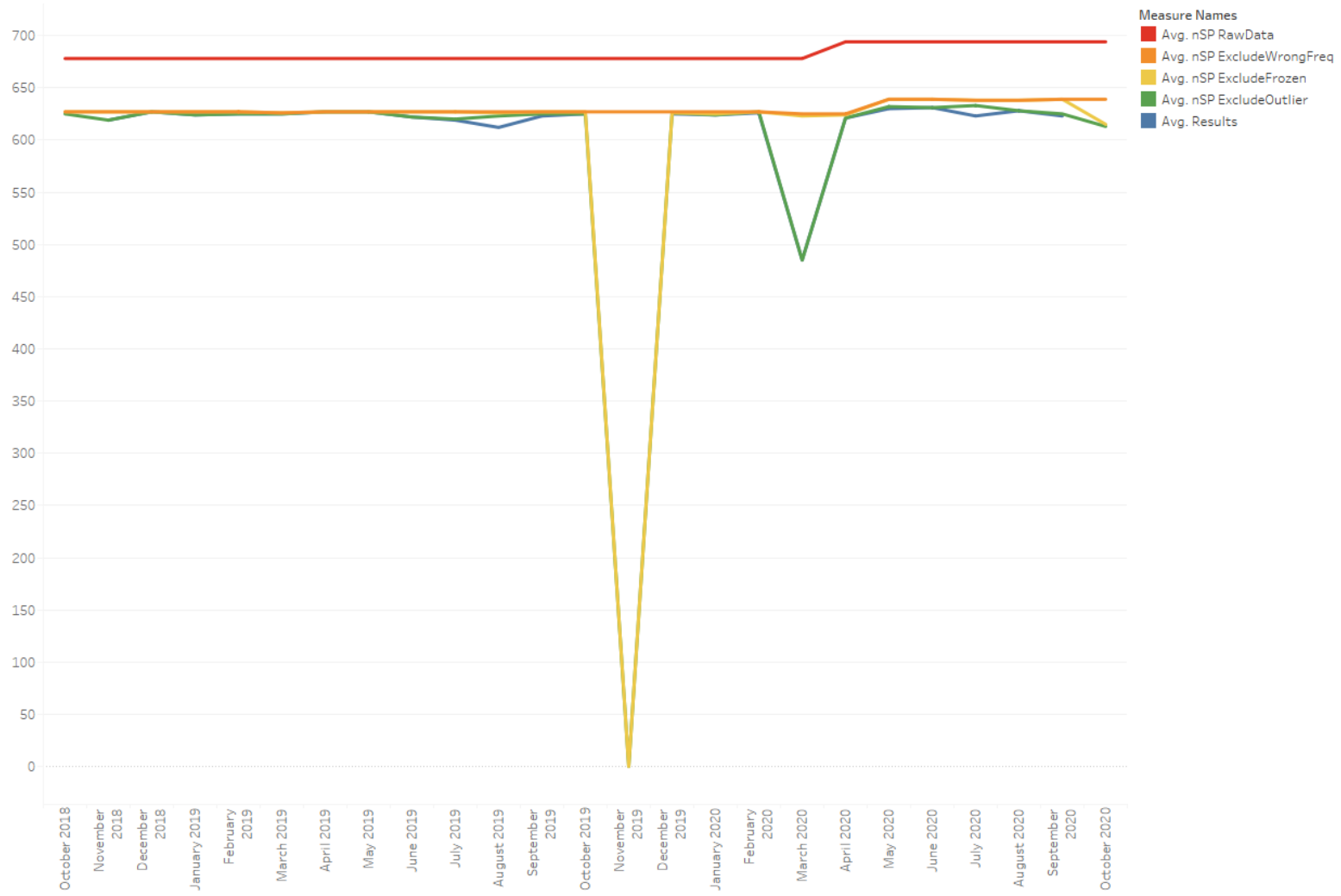



*Figure 14. Data Trimming for Phase identification – Feeder A*

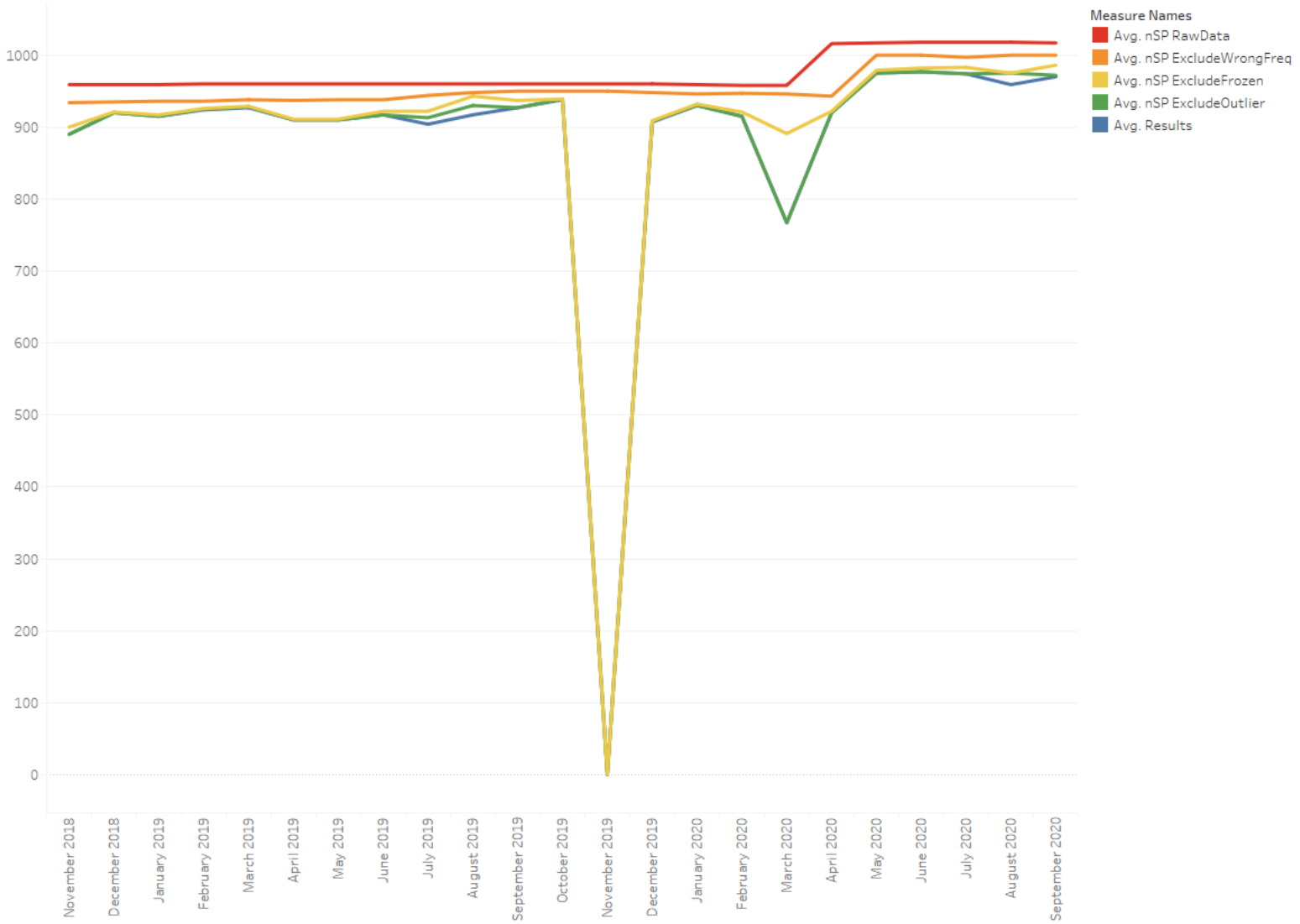



*Figure 15. Data Trimming for Phase identification – Feeder B*

Table 11 lists the average number of meters across all sample months after each step of trimming, along with the percentage drop compared to the raw data. The last row, "Monthly Results", contains the number of meters that have a prediction from each sample month. As explained above, one meter might be excluded from one month but still has predictions from other sample months. Therefore, the number of meters with any results is higher than the average sample size after some steps of data trimming. The last row of Table 11 shows that, on average, the sample months have results for 1,474 meters. However, the whole analysis yields predictions for 1,632 meters in total. The results are discussed in Section 3.1.

*Table 11. Phase Identification Data Trimming*

| | A | | B | | Total | |
|---|---|---|---|---|---|---|
| | # Meters | % Decrease | # Meters | % Decrease | # Meters | % Decrease |
| Raw Data | 682 | | 975 | | 1,657 | |
| Exclude Wrong Freq | 629 | 8% | 955 | 2% | 1,584 | 4% |
| Exclude Frozen Period | 602 | 12% | 894 | 8% | 1,497 | 10% |
| Exclude Outliers | 593 | 13% | 885 | 9% | 1,477 | 11% |
| **Monthly Results** | **591** | **13%** | **883** | **9%** | **1,474** | **11%** |

### *Meter-to-Transformer*

Figure 16 and Figure 17 below are similar charts to the phase identification charts, that plot time series of the number of meters after each step of data trimming. The legend follows a rainbow spectrum from red to purple and are plotted in the same order of data trimming steps. The last time series plotted in pink is the number of meters with a prediction from the analysis. The data trimming steps are listed in order below.

1. Exclude meters with no valid voltage data. The number of meters is plotted in red.
2. Exclude meters with no valid latitude or longitude coordinates. The number of meters is plotted in orange.
3. Exclude transformers with only one meter link to it. The number of meters is plotted in yellow.
4. Exclude meters whose voltages come in wrong intervals. The number of meters is plotted in green.
5. Exclude the records when voltage is frozen for a long period of time. The number of meters is plotted in aqua.
6. Exclude the records jumping up or down that exceeds threshold. The number of meters is plotted in blue.
7. At this step, check and make sure that all transformers have at least two meters with valid data. If not, drop the whole transformer. The number of meters is plotted in purple.
8. The number of meters for which the analysis provides a meter-to-transformer prediction. The number of meters is plotted in pink.

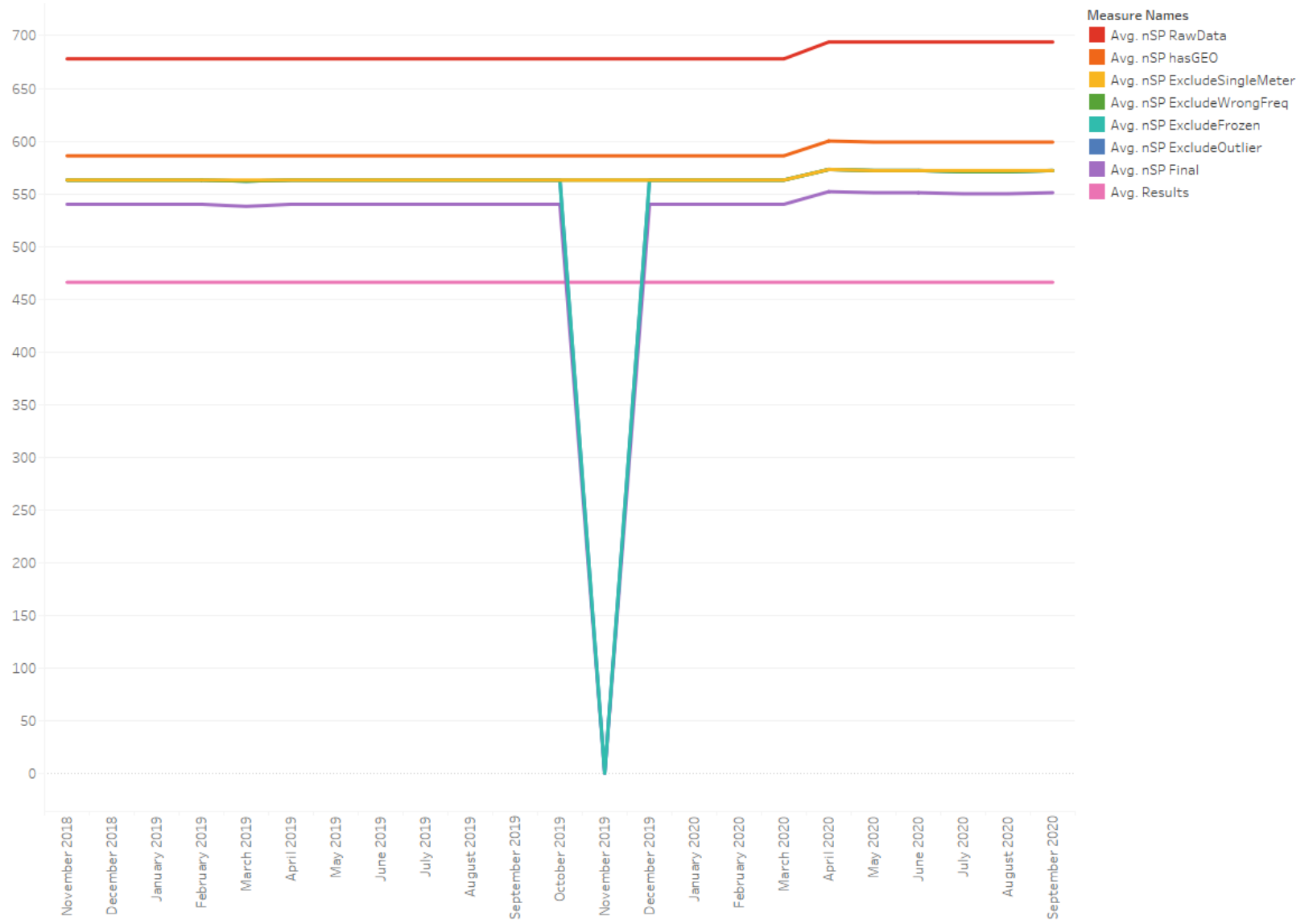



*Figure 16. Data Trimming for meter-to-transformer – Feeder A*

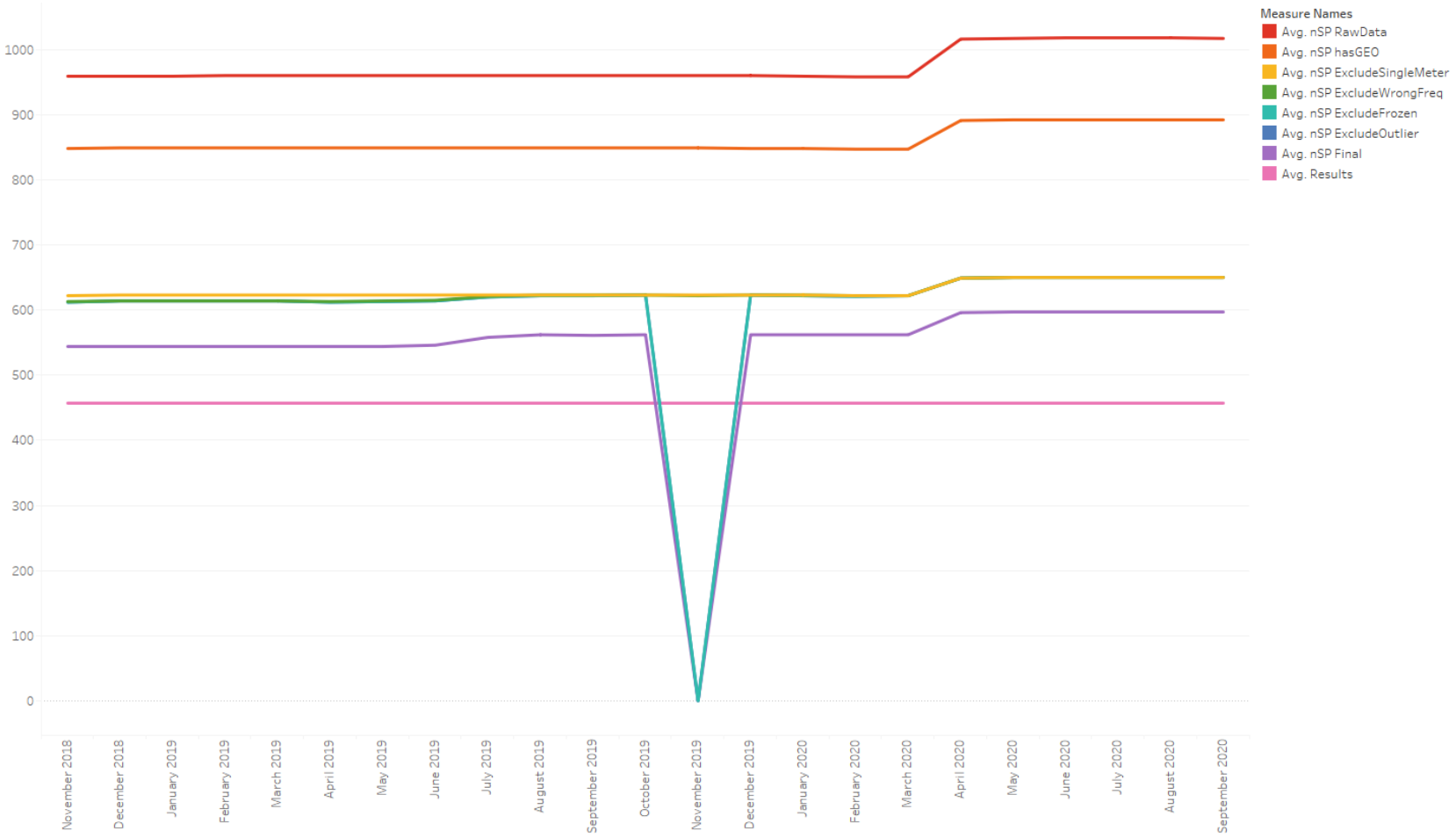



*Figure 17. Data Trimming for meter-to-transformer – Feeder B*

Similarly, as in Table 11, Table 12 lists the average number of meters across all sample months after each step of trimming for the meter-to-transformer analysis. Many more meters are dropped from the analysis, mainly because those meters are on transformers for which only one meter's voltage data is valid. When there is no information on transformer voltages, the meter-to-transformer algorithm uses all the other meters that are linked to the transformer as a proxy for the transformer. If the given meter is the only meter on a transformer, the algorithm is nonfunctional.

Also, the meter-to-transformer algorithm works better for transformers with more meters, and not as well for the transformer with only two meters. Imagine a transformer with only two meters, if one meter is wrong, the algorithm can determine if the two meters do not belong to the same transformer but cannot tell which meter is wrong and which is correct.

More discussion on how number of meters on a transformer affect the meter-to-transformer algorithm is provided in the execution of the meter-to-transformer use case described in Section 2.2.6.

Overall, the meter-to-transformer analysis provides a prediction for 923 meters, about 56% of the meters with voltage data.

*Table 12. Meter-to-Transformer Data Trimming*

| | A | | B | | Total | |
|---|---|---|---|---|---|---|
| | # Meters | % Decrease | # Meters | % Decrease | # Meters | % Decrease |
| Raw Data | 682 | | 976 | | 1,658 | |
| Exclude Invalid GEO | 590 | 14% | 861 | 12% | 1,450 | 13% |
| Exclude Single Meter | 566 | 17% | 630 | 35% | 1,196 | 28% |
| Exclude Wrong Freq | 565 | 17% | 627 | 36% | 1,192 | 28% |
| Exclude Frozen Period | 542 | 21% | 601 | 38% | 1,143 | 31% |
| Exclude Outliers | 542 | 21% | 601 | 38% | 1,143 | 31% |
| Final Sample | 521 | 24% | 542 | 44% | 1,063 | 36% |
| **Results** | **466** | **32%** | **494** | **49%** | **960** | **42%** |

# 3.0 Results

The purpose of this module was to apply the phase identification model and meter-to-transformer model on SDG&E's data and test the prediction accuracy. While the prediction accuracy was discussed thoroughly in the previous section, there are some additional metrics worth discussion.

## 3.1 Results Discussion

### 3.1.1 Phase Identification Prediction Accuracy

Given a model prediction, the next step required is the evaluation against the ground truth for verification of the accuracy scores. As the ground truth is not available, SDG&E's labels are used as a proxy initially. Table 13 and Table 14 below are the confusion matrices comparing SDG&E's labels and the

model predictions, for Feeder A and B, respectively. The matching percentage, "% Match" in the tables is defined as the number of meters where model prediction matches SDG&E's label over the number of meters where SDG&E's label is available.

In the confusion matrices below, the row heads are SDG&E's labels, and the column heads are model predicted phases. There are some meters labeled as "No Access", "Undermined", etc., in SDG&E's metadata. Those meters are combined as "No Info" in the table and are hence excluded when calculating matching percentages. There are some meters labeled as "ABC" and are also excluded from the "% Match" calculation.

The bold numbers on the diagonal of each matrix are the number of meters where model prediction matches SDG&E's labels, and the off-diagonal numbers are unmatched meters. The "% Match" column provides the percentage of matched meters over total meters for a given SDG&E phase group. The number at the most bottom right of each table, is the overall matching rate for the feeder.

For Feeder A, most of the meters are configured as L-L. While the overall matching rate is 92%, the matching rate for the 264 L-N meters is 96.4%, and for the 312 L-L meters, the rate is 88.1%.

*Table 13. Confusion Matrix – Feeder A*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | **64** | | 1 | | | | **65** | **98%** |
| B | | **98** | 2 | | 2 | | **102** | **96%** |
| C | 1 | 2 | **93** | 1 | | | **97** | **96%** |
| AB | | | | **67** | 1 | 13 | **81** | **83%** |
| BC | | | 1 | 1 | **137** | 1 | **140** | **98%** |
| AC | | 1 | | 17 | 2 | **71** | **91** | **78%** |
| ABC | | 1 | 1 | 25 | 10 | 5 | **42** | |
| No Info | 2 | | 6 | 2 | 1 | 8 | **19** | |
| Total | **67** | **102** | **104** | **113** | **153** | **98** | **637** | **92%** |

For Feeder B, the matching rate is 97%, with 831 matched meters, and 29 unmatched cases. On Feeder B, most of the meters are configured as L-N, and among the 832 L-N phase meters, the matching rate is 96.63%, and for L-L, 96.59%, not much difference.

*Table 14. Confusion Matrix – Feeder B*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | 259 | 1 | 1 | | 1 | 6 | 268 | 97% |
| B | 8 | 296 | 4 | | 2 | 2 | 312 | 95% |
| C | 1 | 2 | 249 | | | | 252 | 99% |
| AB | | | | 16 | 1 | | 17 | 94% |
| BC | | | | | 1 | | 1 | 100% |
| AC | | | | | | 10 | 10 | 100% |
| ABC | | 1 | | 7 | 16 | 22 | 46 | |
| No Info | 24 | 30 | 18 | 1 | 1 | 15 | 89 | |
| Total | 292 | 330 | 272 | 24 | 22 | 55 | 995 | 97% |

The two 90%+ matching rates prove the model performs well and confirms SDG&E's records are of high quality. When the two sets of records agree with each other, correct meter labeling potential increases significantly. On the other hand, when the two sets of records do not agree with each other, it might be due to wrong prediction from the model *or* to an error in SDG&E's records.

There are 75 meters for which the model prediction does not agree with SDG&E's records, 46 on Feeder A, and 29 on Feeder B. For 72 out of these 75 meters, the model has very consistent prediction across all sample months available. Using GMSV, a virtual field verification was performed for these 72 meters to sort out the phase from overhead power lines wherever possible.

Feeder A has more than half of the meters on overhead powerlines, while the meters on Feeder B are mostly underground. Therefore, the virtual field verification applied mainly to Feeder A. There is only one meter on Feeder B that was checked on GMSV. Out of the 72 meters on which the virtual field verification is attempted, GMSV gave conclusive results for 35 of them and among these 35 meters, the model predictions are correct. That is, the field verification results were incorrect. Therefore, if assuming the model prediction is wrong for all the other 40 meters, when comparing to this version of ground truth, the model accuracy rate for Feeder A increases to 98%, and for Feeder B, the rate does not change much, remaining at 97%.

Figure 18 and Figure 19 below highlight the meters where virtual field verification succeeded on Feeder A and Feeder B, respectively.

Feeder A has many meters on overhead powerline, and hence many opportunities to do virtual field verification. On the map below, five groups of meters are circled and labeled from upper left to lower right, as group two to six.

For the following description, fictitious street names are substituted to preserve anonymity.

For Group 2, there are five meters; the transformers for these meters are traced from Elm St. to Oak Rd, where another transformer is connected to the same lines. Both the model prediction and SDG&E label agree that the transformer on Oak Rd. is Phase AB, and hence Group 2 should be the same.

Group 3 has 21 meters. The information needed to prove the phase for these meters is deducted from the meters whose phases are proven true, as shown below.

- The transformer on Fir Dr. to the North of Fir Dr. – Birch Ave. connection is proven as AC.
    - The transformers on Birch Ave. between Pine Road and Fir Dr. are of the same phase, AC.
    - Transformers on Willow Place are of a different phase, and hence are not AC.
- The transformers on Ash Dr. to the SE of Ash Dr. – Willow Place connection are proven AB.
    - The transformers on Cedar Dr. are different than the transformers on Ash Dr., and hence are not AB.

The other groups are all deducted following similar logic.

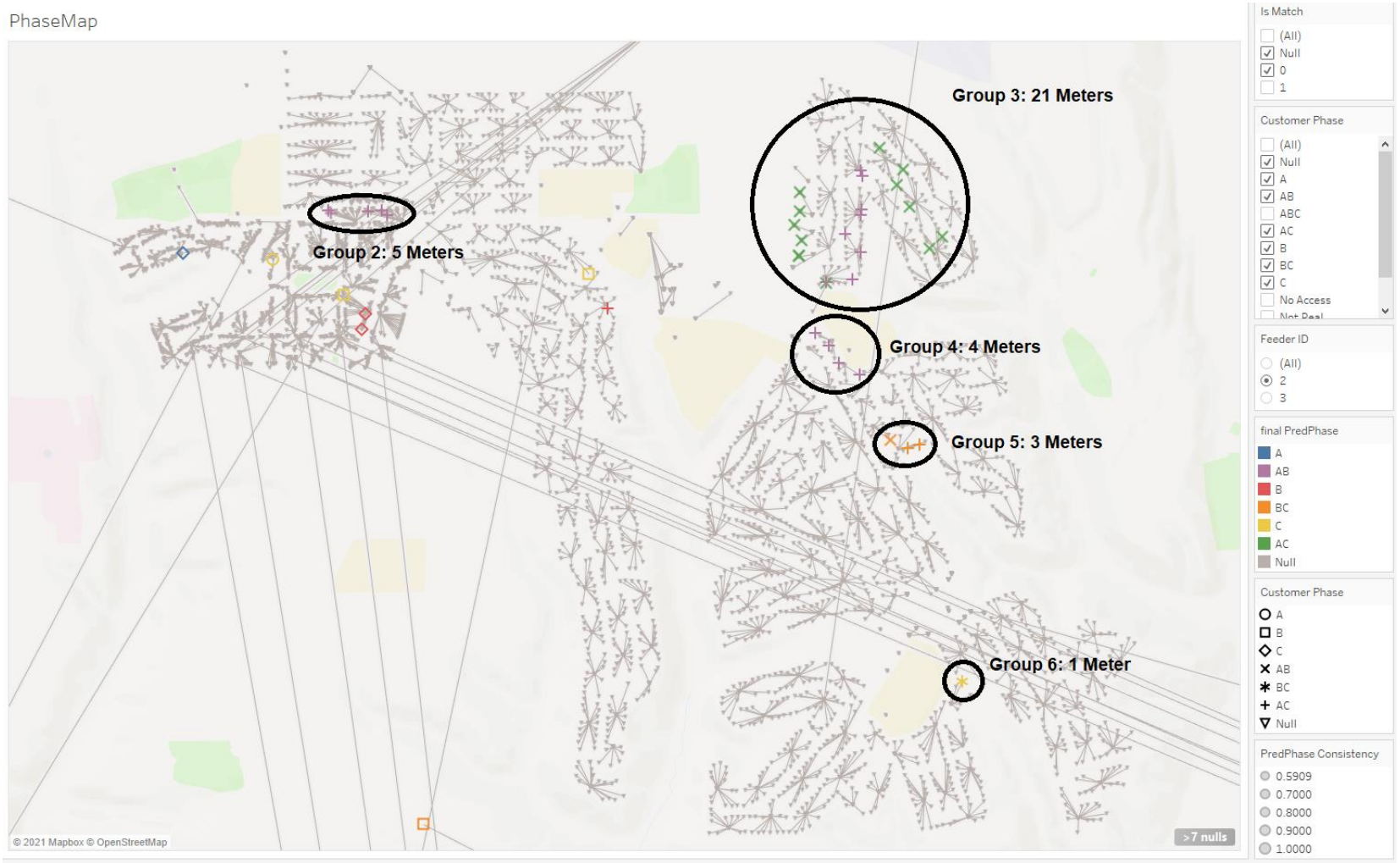



*Figure 18. Virtual Field Verification on Feeder A*

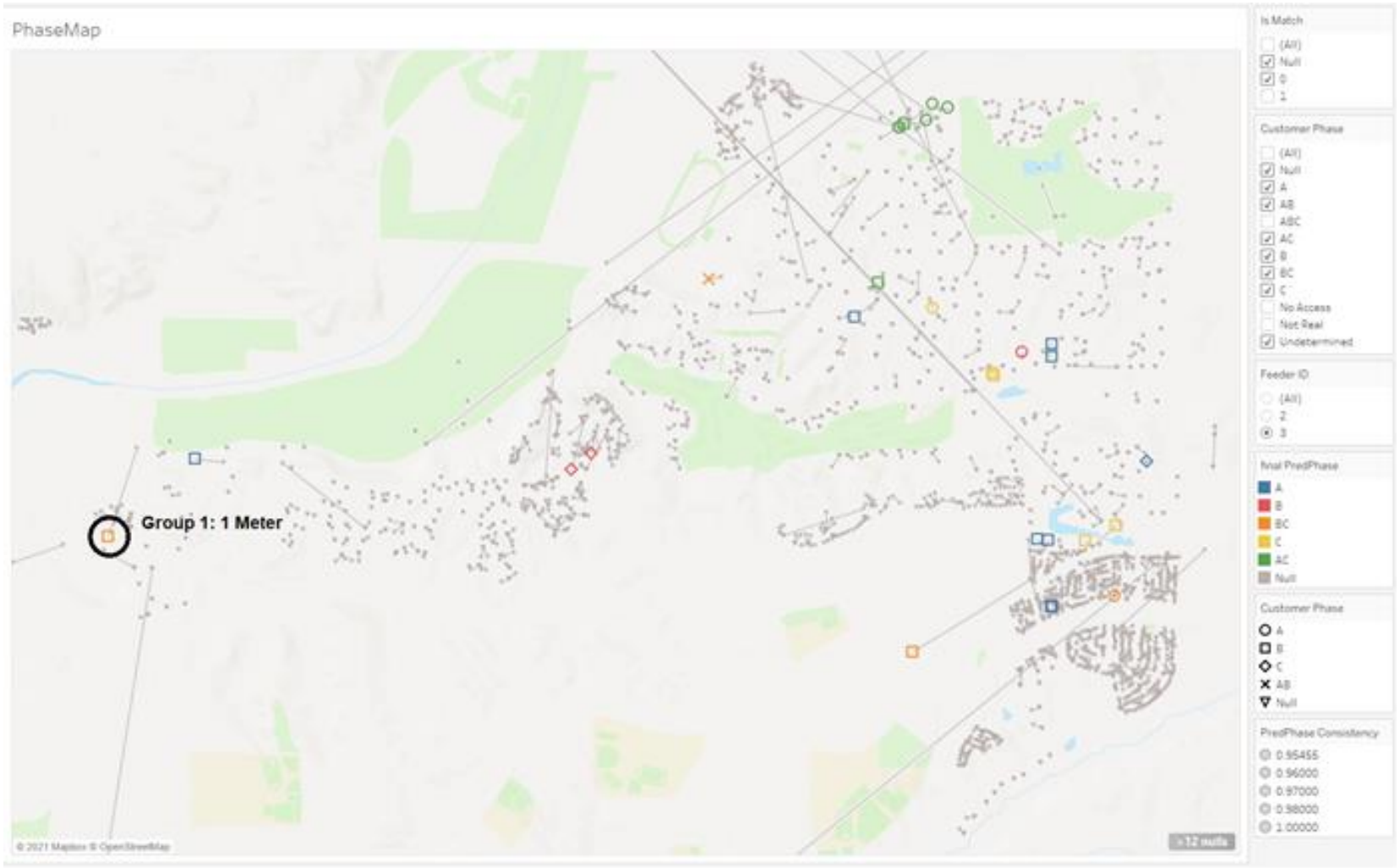



*Figure 19. Virtual Field Verification on Feeder B*

For Feeder B, the meter is on the left of the map. As shown in the legend on the right-hand side, SDG&E labels it with a square, or Phase B, but the model predicts it as orange, Phase BC. From GMSV, the meter should be L-L, not L-N. Figure 20 below illustrates the map of unmatched meters for Feeder B.

After virtual field verification, the number of unmatched meters decreases to 40, and the accuracy rates improve to 97% and 98%. The two figures below, Figure 21, and Figure 22 show these 40 meters on the map. The intention of these maps is to provide a direct and intuitive view of the model prediction, emphasizing where the model prediction does not match this version of ground truth.

In both figures, the shape of the icons is defined by ground truth. Three closed shapes, circle, square, and diamond, represent the three L-N phases, A, B, and C, respectively. Three radiant shapes include a plus sign, multiplication sign, and a star. These represent the three L-L phases, AB, AC, and BC. Phase ABC is represented with triangles, and the meters with no information are plotted as upside-down triangles.

The icon color is defined by the model prediction. Phase A, B, and C use three primary colors blue, red and yellow; and phase AB, AC, and BC use purple, green, and orange. If the prediction does not match the ground truth, the meter is emphasized with a bigger icon.

On Feeder A, there are several unmatched meters on the outskirts of the map, and hence two maps are included. The first one shows the whole picture, especially the unmatched meters that are scattered outside of the sizable cluster of meters. The second one zooms in and shows unmatched meters at the center of this feeder.

On the map, it seems many unmatched meters appear as small blocks. In fact, 34 out of these 40 meters are linked to a transformer for which two meters' voltage data is provided for the analysis. Eighteen meters, or nine pairs, have the phase prediction of one meter in agreement with the other meter on the same transformer, and both different than ground truth. Such a "coincidence" adds more confidence to the model.

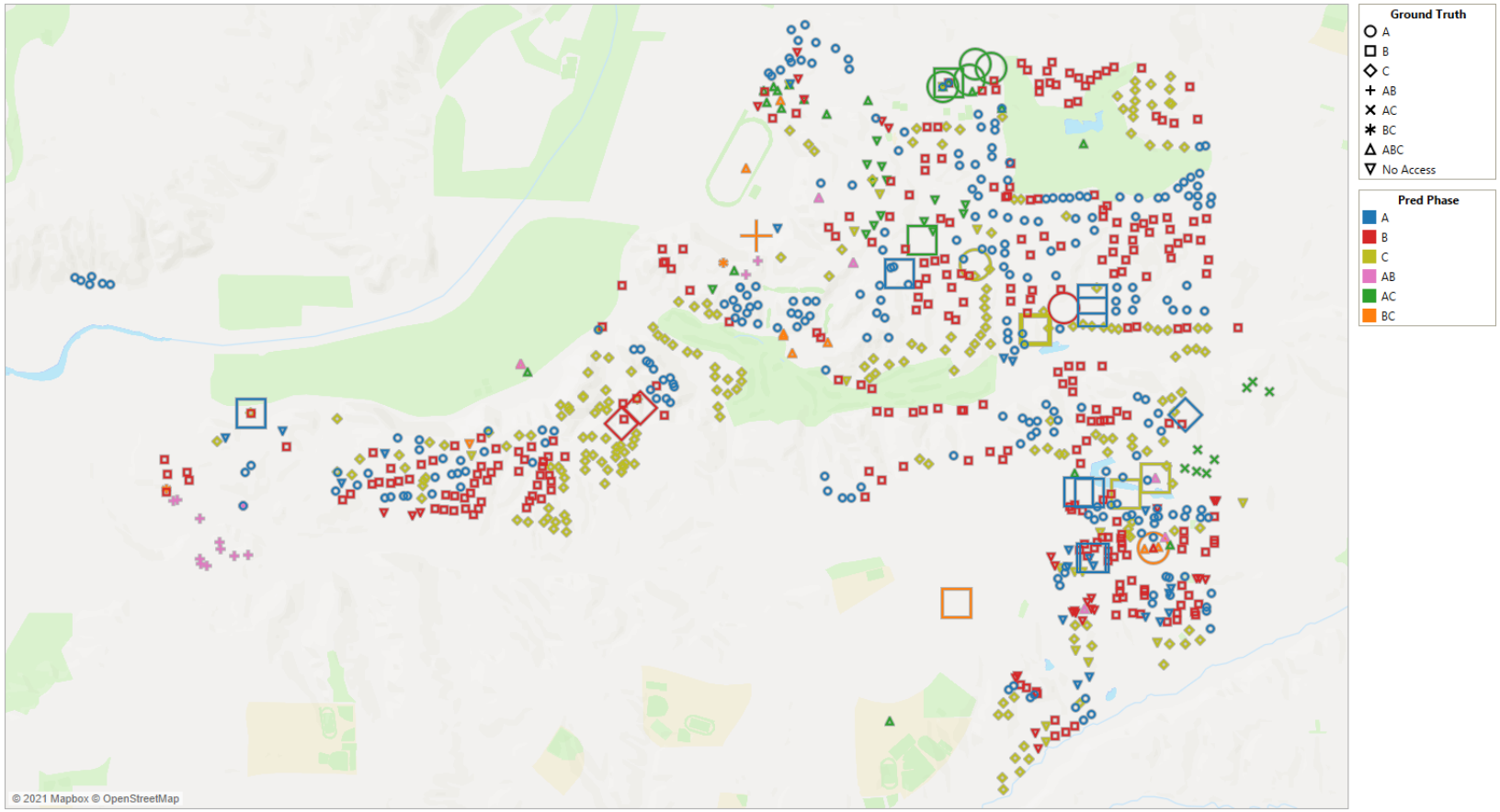



*Figure 20. Map for Unmatched Meters - Feeder B*

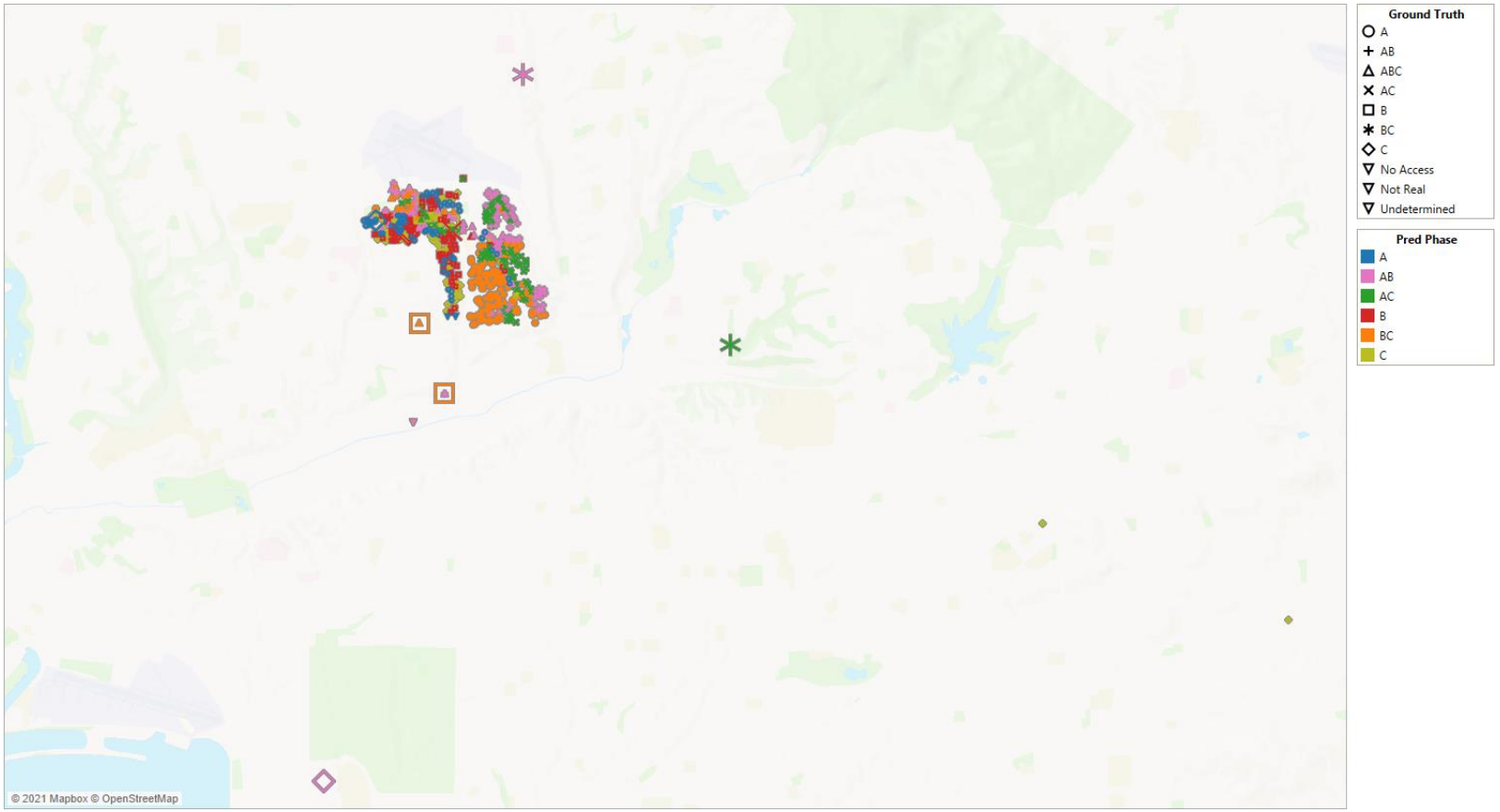



*Figure 21. Map for Unmatched Meters - Feeder A Broad View*

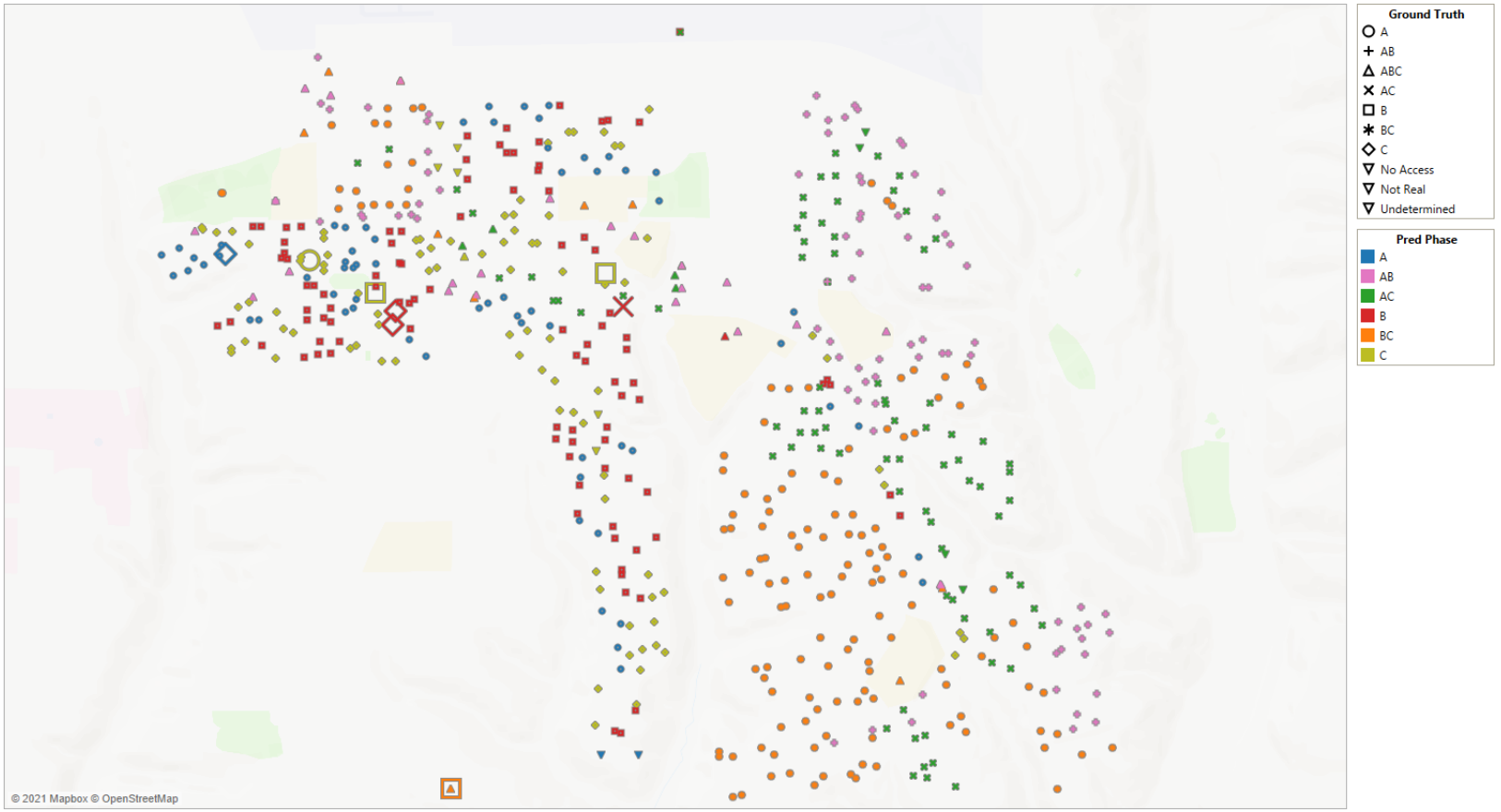



*Figure 22. Map for Unmatched Meters - Feeder A Focus View*

The other way to view these unmatched meters is through observation of where they reside within each of the clusters. The clusters are constructed using correlations with each kernel. Figure 23 and Figure 24 below project each meter's correlations with three L-N kernels onto a two-dimensional plane showing the correlation-location of the meters in clusters. Since these two plots are projections from a three-dimensional image, the x-axis scale and y-axis scale are not meaningful. They can be labeled as 0.5, 5, or 50, without any genuine change.

Again, in these two figures, shapes are defined by ground truth, colors are defined by model prediction, and size emphasizes if the prediction does not match ground truth.

Most meters' prediction matches the ground truth, and the figures have blue (prediction = "A") circle (ground truth = "A"), red (prediction = "B") square (ground truth = "B"), etc., but there are also some bigger icons.

For example, on Feeder B, Figure 24, there are many blue squares, green circles, red diamonds, and yellow squares. These meters, however, are located at the center or close to the center of each cluster. Take the blue squares as an example, it is more likely that they are in a blue circle cluster than in a red square cluster. The big orange square in the middle of Figure 24 is problematic. In fact, some sample months predict it as "AC", but more than 90% of the sample months yield prediction of "BC" as plotted in the figure. Figure 23 also shows several problematic examples on Feeder A. There is an orange square, a purple diamond, and a purple star in the middle of the plot. It seems the orange square has a higher possibility to be in an orange star cluster than in a red square cluster. Similarly, so is the purple diamond. It is far from the yellow diamond cluster at the bottom. The purple star, on the other hand, is not far from the orange star cluster, and close to the purple plus sign cluster. In fact, the prediction for this meter is not conclusive at all. About 60% of the sample months predict this meter as "AB", but the other 40% of the sample months do not agree. This meter is the purple star located on top of Figure 21, which seems to have bad latitude and longitude information, and therefore is not suitable for virtual field verification.

If the model is successful, excluding the meters located at the center of each figure, the other meters are much more likely to be in the predicted phase, as they are located closer to the center of the predicted clusters.

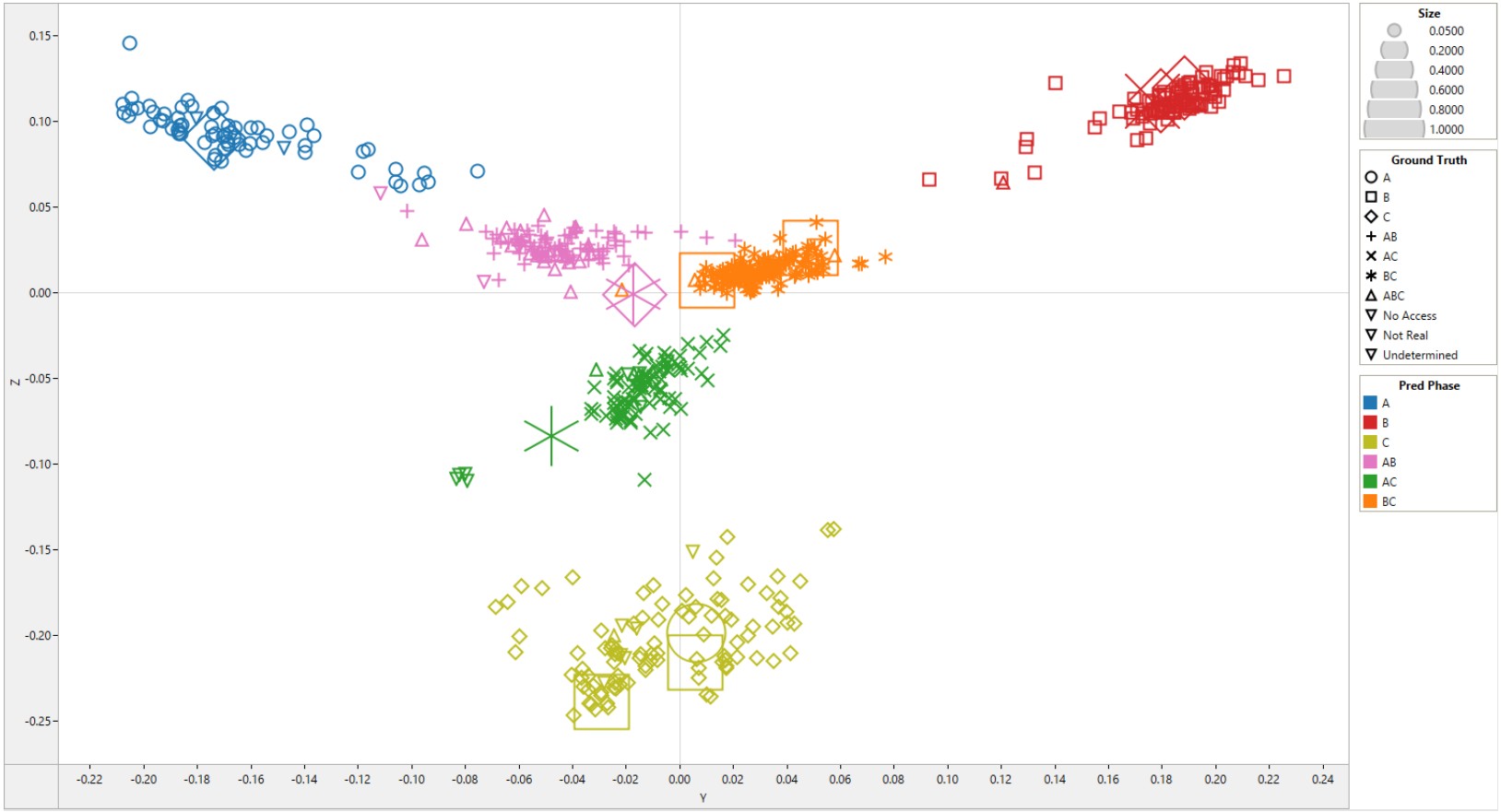



*Figure 23. Correlation Clusters Plot – Feeder A*

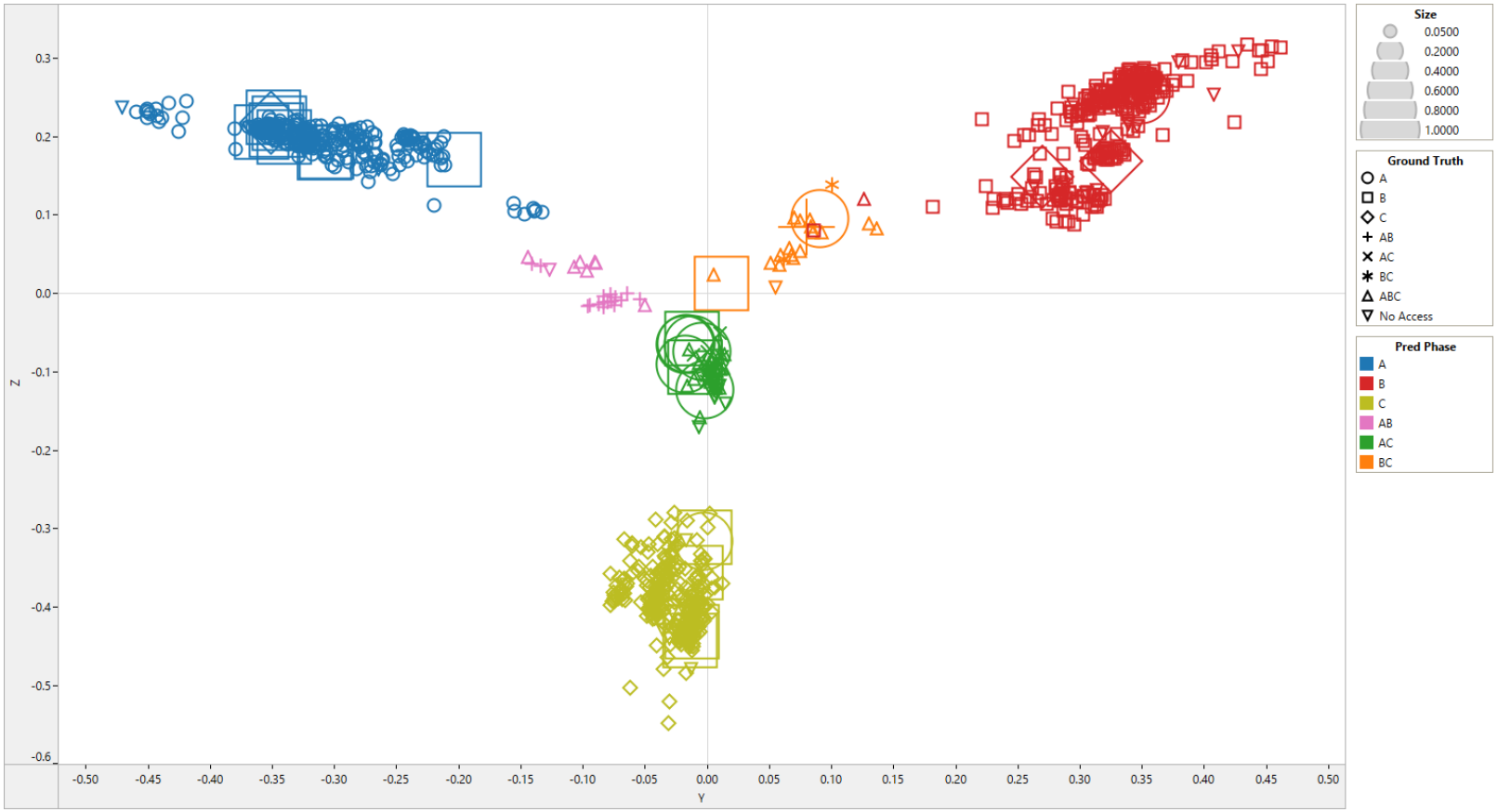



*Figure 24. Correlation Clusters Plot – Feeder B*

While the map and correlation plots add more confidence to the model prediction, without validation, an assumption for correct predictions can't be made. Therefore, SDG&E's label remains as the ground truth for these 40 meters.

Table 15 and Table 16 below are the updated confusion matrices for Feeder A and B, respectively. There are more L-N meters that do not match. Possible explanations include, 1) SDG&E's L-N meters are usually underground and cannot be virtually checked and 2) the model prediction does not perform as well for L-N meters.

*Table 15. Updated Confusion Matrix – Feeder A*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| **A** | **64** | | **1** | | | | **65** | **98%** |
| **B** | | **98** | **2** | | **2** | | **102** | **96%** |
| **C** | **1** | **2** | **94** | **1** | | | **98** | **96%** |
| **AB** | | | | **84** | | | **84** | **100%** |
| **BC** | | | | **1** | **140** | **1** | **142** | **99%** |
| **AC** | | **1** | | | | **84** | **85** | **99%** |
| **Total** | **65** | **101** | **97** | **86** | **142** | **85** | **576** | **98%** |

*Table 16. Updated Confusion Matrix – Feeder B*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| **A** | **259** | 1 | 1 | | 1 | 6 | **268** | **97%** |
| **B** | 8 | **296** | 4 | | 1 | 2 | **311** | **95%** |
| **C** | 1 | 2 | **249** | | | | **252** | **99%** |
| **AB** | | | | **16** | 1 | | **17** | **94%** |
| **BC** | | | | | **2** | | **2** | **100%** |
| **AC** | | | | | | **10** | **10** | **100%** |
| **Total** | **268** | **299** | **254** | **16** | **5** | **18** | **860** | **97%** |

In conclusion, the accuracy rate is between 97% and 98%. It is possible the real accuracy rate is even higher. It would be valuable to cross-validate the 40 meters where GMSV analysis could not be performed. Also valuable would be to have the ground truth about which phase is being metered for the other 196 meters that are labeled "ABC" or "No Info".

### 3.1.2 Phase Identification Results Discussion Part I: Effects from Data

*Number of meters that "represent" each transformer*

Several factors in the data source might affect the accuracy level. For example, as mentioned in the data description, the analysis does not use the voltage data for the whole frame, the data is only available for the selected one or two meters on each transformer. If one or two meters are enough to identify the transformers' phase, it will save data transferring bandwidth and data storage.

Will this setting affect the accuracy rate? Table 17 compares the accuracy rates for transformers with different numbers of associated meters per transformer with voltage data provided. As explained above, the accuracy rates are lower for L-N meters than L-L meters. Therefore, the comparison is shown in each category separately. For the L-L category, the accuracy rate is slightly higher when data is available for more meters on the transformer. When the number of meters with data increases from one to three, the accuracy rate increase from 95% to 100%. However, this can be explained by the fact that Feeder A has a greater number of large transformers than Feeder B, and Feeder A has a higher accuracy rate. When focusing on Feeder A, the accuracy rates are not vastly different, and for Feeder B, the sample is not big enough to draw any conclusion.

As for the L-N meter group, the results show no pattern. For both feeders, the accuracy rates are the lowest for the middle group, where transformers have two meters with voltage data.

Therefore, the conclusion is, there is no evidence the phase identification algorithm works better for transformers with more data.

*Table 17. Accuracy Rate by Transformer Size*

| | | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|---|
| **Configuration Type** | **# Meters per Transformer** | **# Meters** | **% Accurate** | **# Meters** | **% Accurate** | **# Meters** | **% Accurate** |
| L-N | 1 | 5 | 100% | 227 | 98% | 232 | 98% |
| L-N | 2 | 226 | 96% | 572 | 96% | 798 | 96% |
| L-N | 3 | 33 | 97% | 33 | 100% | 66 | 98% |
| L-L | 1 | 9 | 100% | 12 | 92% | 21 | 95% |
| L-L | 2 | 252 | 99% | 16 | 100% | 268 | 99% |
| L-L | 3 | 51 | 100% | 0 | | 51 | 100% |
| **Total** | | **576** | **98%** | **860** | **97%** | **1,436** | **97%** |

*Length of time series data*

Another factor is the number of months of valid data provided for each meter. If a longer period of data can significantly increase the accuracy rate, then it is worth incorporating longer time series into the analysis. Table 18 compares the accuracy rates by the number of sample months. "Full Data" means the meter includes all 22 months in the analysis; "Almost Full" means 20 or 21 months of data; and the other two categories are self-explanatory. The meters with less than half a year of data are all predicted

correctly and there is not much difference among the other three categories' accuracy rates. Based on this analysis, we cannot draw any conclusive result for this dimension.

*Table 18. Accuracy Rate by Number of Sample Months*

| | | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|---|
| **Configuration Type** | **# Sample Months** | **# Meters** | **% Accurate** | **# Meters** | **% Accurate** | **# Meters** | **% Accurate** |
| L-N | Full Data | 168 | 96% | 508 | 96% | 676 | 96% |
| L-N | Almost Full | 83 | 98% | 207 | 97% | 290 | 97% |
| L-N | More than half year | 11 | 91% | 74 | 97% | 85 | 96% |
| L-N | Less than half year | 2 | 100% | 43 | 100% | 45 | 100% |
| L-L | Full Data | 245 | 99% | 21 | 100% | 266 | 99% |
| L-L | Almost Full | 53 | 100% | 6 | 83% | 59 | 98% |
| L-L | More than half year | 7 | 100% | 1 | 100% | 8 | 100% |
| L-L | Less than half year | 7 | 100% | 0 | | 7 | 100% |
| **Total** | | **576** | **98%** | **860** | **97%** | **1,436** | **97%** |

*Reading frequency or interval length*

The data frequency or the interval length is another factor that affects the results greatly. If the data comes in 10-minute intervals, with the rest unchanged, the data size decreases by half. Therefore, the data transferring and storing costs go down, and data processing is faster. On the other hand, however, when the voltages are averaged across 10-minute intervals rather than five-minute intervals, some of the phase specific signals may be averaged away and blended into the white noise on the grid. Additionally, as the phase signature movements are taken away little by little, the correlations are harder and harder to cluster.

Table 19 through Table 22 provide comparisons of the model prediction using 10-minute interval voltage data with ground truth and five-minute interval voltage data, for Feeders A and B. The 10-minute interval model uses the same set of parameters as the five-minute interval model, with no adjustment. This provides a better comparison between the two data settings, since all the differences are due to different interval lengths.

However, even though the parameters used to trim data are all the same, the results are different. For example, the frozen period is defined to have at least 12 consecutive intervals where the voltage remains linear or has no change. In a five-minute interval data setting, 12 intervals equate to one hour, and in a 10-minute interval data setting, 12 intervals equate to two hours. This means fewer data points are trimmed off due to the frozen period. When less data is trimmed off, more meters are included. In Table 19, the total number of meters is 997, two more than in Table 15.

On the other hand, when fewer data are trimmed off, more noise is kept in the model, which means it is more difficult to find correlation patterns, and fewer clear clusters.

For Feeder B, the two sets of results agree by 99.1%, but for Feeder A, there is some degree of confusion between phase AB and BC, as highlighted in red in Table 20 and Table 22.

*Table 19. Confusion Matrix – Feeder A: comparing 10-min model prediction with ground truth*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | 64 | | 1 | | | | 65 | 98% |
| B | | 98 | 2 | 1 | 1 | | 102 | 96% |
| C | 1 | 2 | 93 | 1 | | | 97 | 96% |
| AB | | | | 63 | 5 | 13 | 81 | 78% |
| BC | | | 1 | 61 | 77 | 1 | 140 | 55% |
| AC | | 1 | | 19 | | 71 | 91 | 78% |
| ABC | 3 | 1 | 1 | 26 | 6 | 5 | 42 | |
| No Info | 3 | | 6 | 6 | | 4 | 19 | |
| Total | 71 | 102 | 104 | 177 | 89 | 94 | 637 | 81% |

*Table 20. Confusion Matrix – Feeder B: comparing 10-min model prediction with ground truth*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | 253 | 1 | 1 | 6 | 1 | 6 | 268 | 94% |
| B | 8 | 295 | 4 | | 2 | 2 | 311 | 95% |
| C | 1 | 2 | 251 | | | | 254 | 99% |
| AB | | | | 16 | 1 | | 17 | 94% |
| BC | | | | | 2 | | 2 | 100% |
| AC | | | | | | 10 | 10 | 100% |
| ABC | | | | 8 | 16 | 22 | 46 | |
| No Info | 24 | 30 | 18 | 1 | 1 | 15 | 89 | |
| Total | 286 | 328 | 274 | 31 | 23 | 55 | 997 | 96% |

*Table 21. Confusion Matrix – Feeder A: comparing 10-min model prediction with five-min model prediction*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | 67 | | | | | | 67 | 100% |
| B | | 102 | | | | | 102 | 100% |
| C | | | 104 | | | | 104 | 100% |
| AB | 1 | | | 107 | 5 | | 113 | 95% |
| BC | | | | 70 | 83 | | 153 | 54% |
| AC | | | | 4 | | 94 | 98 | 96% |
| Total | 68 | 102 | 104 | 181 | 88 | 94 | 637 | 87% |

*Table 22. Confusion Matrix – Feeder B: comparing 10-min model prediction with five-min model prediction*

| | A | B | C | AB | BC | AC | Total | % Match |
|---|---|---|---|---|---|---|---|---|
| A | 286 | | | 6 | | | 292 | 98% |
| B | | 328 | | | 2 | | 330 | 99% |
| C | | | 272 | | | | 272 | 100% |
| AB | | | | 24 | | | 24 | 100% |
| BC | | | | 1 | 21 | | 22 | 95% |
| AC | | | | | | 55 | 55 | 100% |
| Total | 286 | 328 | 272 | 31 | 23 | 55 | 995 | 99% |

Correlation cluster plots shed more light on understanding the difference between the two sets of model predictions. Figure 25 and Figure 26 below compare the correlation plots for Feeder A and B, respectively. For all four panels in the two figures, shape is defined by the five-minute interval model and color is defined by the 10-minute interval model, as shown in the legend section on the bottom of each figure. The unmatched meters are emphasized with larger icons.

For each figure, the left panel plots the projection of three-dimensional correlations from the five-minute model, and the right panel for the 10-minute model. The panels look similar for both feeders. It seems strange that such similar correlation plots generate different cluster results. For example, in Figure 25, the two orange squares look so out of place, and in Figure 26, it is obvious that the big chunk of purple stars near the center should be orange.

Referring to Figure 3, the algorithm flowchart diagram indicates the average correlations are calculated twice. The first calculation is the step provided in the top rectangle and the second calculation is the step provided in the bottom rectangle. The correlations from the first calculation define the kernels, and the correlations out of the second are plotted in Figure 25 and Figure 26 below. If kernels are defined using

this second set of correlations, given the similarity between the left and right panels of each figure, the phase prediction will be similar for the five-minute model and 10-minute model accordingly. In Figure 23 and Figure 24, all colors are clustered properly, meaning that the kernels out of this algorithm outlined in Figure 3 converge with the kernel feed into this step. But on the right panels of Figure 25 and Figure 26, some colors are mixed, indicating that the model is not fully converged. One way to fix this issue is to manually adjust the data trimming parameters so that the white noise decreases, and the correlations show more pattern from each phase. Therefore, the cluster step is easier and yields better results. Another way is to loop it one more round until the process produces a converged prediction.

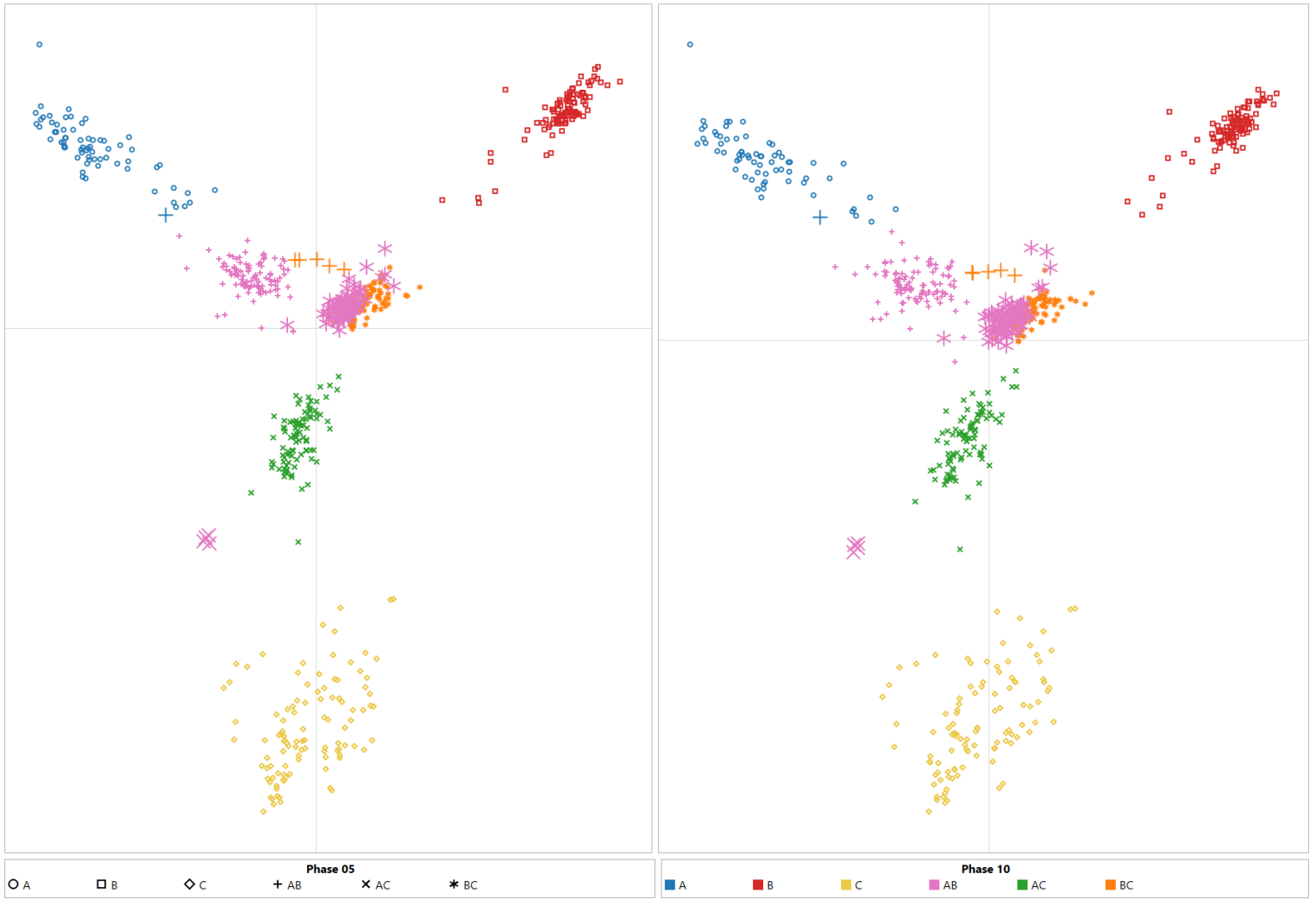



*Figure 25. Correlation Clusters Plot – Feeder A: comparing between 5-min interval and 10-min interval model*

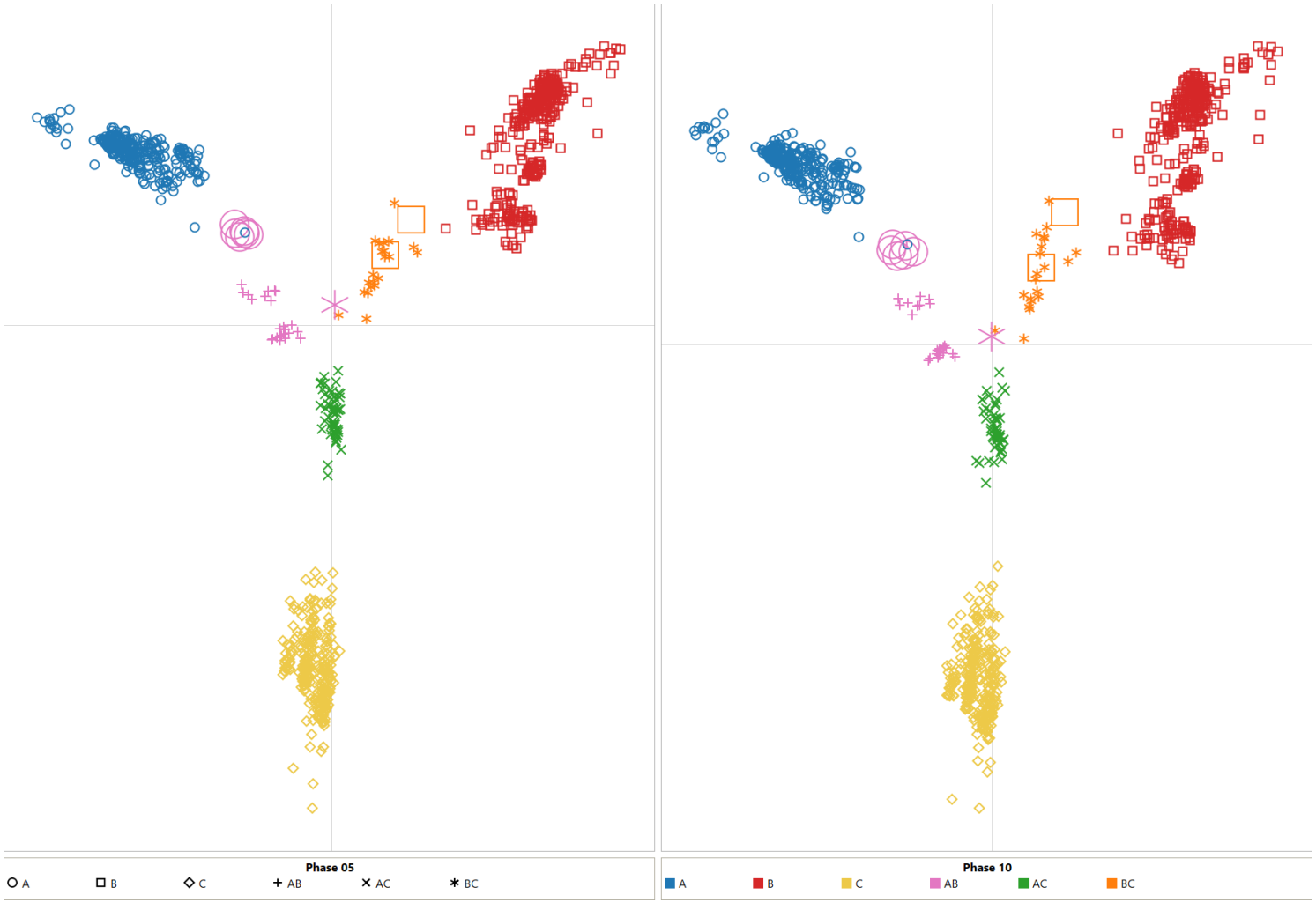



*Figure 26. Correlation Clusters Plot – Feeder B: comparing between 5-min interval model and 10-min interval model*

### 3.1.3 Phase Identification Results Discussion Part II: Model Statistics

*Consistency rate*

Consistency rate is another statistic worth mentioning. These metric measures how consistent the sample month predictions are. As explained in the phase identification algorithm, the final prediction is a summary of all sample month results. The consistency rate is defined as the number of months where the monthly prediction is consistent with the final results over the total number of months. It is expected that accurate predictions are more likely with a higher consistency rate.

Table 23 below compares the accuracy rates across different consistency levels. Feeder A has some meters with some sample month results that are inconsistent with the final prediction, and the L-L meter group shows decreasing accuracy rates when the consistency level drops. For Feeder B, almost all meters have 100% consistent predictions. Such results show strong confidence in the model, but at the same time contribute little to no value to understanding the relationship between consistency level and accuracy rate.

*Table 23. Accuracy Rate by Consistency Level*

| | | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|---|
| Configuration Type | Consistency Level | # Meters | % Accurate | # Meters | % Accurate | # Meters | % Accurate |
| L-N | 100% | 188 | 97% | 829 | 97% | 1,017 | 97% |
| L-N | 90% and up | 74 | 99% | 2 | 50% | 76 | 97% |
| L-N | Less than 90% | 2 | 0% | 1 | 100% | 3 | 33% |
| L-L | 100% | 264 | 100% | 27 | 96% | 291 | 100% |
| L-L | 90% and up | 43 | 95% | 1 | 100% | 44 | 95% |
| L-L | Less than 90% | 5 | 80% | 0 | | 5 | 80% |
| **Total** | | **576** | **98%** | **860** | **97%** | **1,436** | **97%** |

*Hybrid index*

Another important statistical output is the hybrid index. The hybrid index is used to separate L-N phases from L-L phases. The meters with a lower hybrid index are L-N, and the meter with a higher hybrid index are L-L. The Agglomerative Cluster method is used to decide on the cutoff point. If the model works well, the cutoff point should be obvious.

However, there are still meters closer to the cutoff points than the other meters. Those are the meters that pose some challenge to the model. Therefore, it is worth comparing the accuracy rates between the meters closer to the cutoff point and the meters farther away.

Table 24 provides such comparison, and the numbers look interesting. While Feeder B's accuracy rate increases as the distance from the cutoff point increases, Feeder A shows the opposite trend. For the L-N group on Feeder A, the accuracy rate is 100% for all the meters that are close to the cutoff points, and the rate drops to 98% and then 94% for meters farther away.

*Table 24. Accuracy Rate by Distance from Hybrid Index Cut-Off Point*

| | | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|---|
| Configuration Type | # Distance from Cutoff | # Meters | % Accurate | # Meters | % Accurate | # Meters | % Accurate |
| L-N | Very Close | 5 | 100% | 7 | 86% | 12 | 92% |
| L-N | Close | 18 | 100% | 3 | 33% | 21 | 90% |
| L-N | Far | 123 | 98% | 45 | 87% | 168 | 95% |
| L-N | Very Far | 118 | 94% | 777 | 98% | 895 | 97% |
| L-L | Very Close | 4 | 100% | 0 | | 4 | 100% |
| L-L | Close | 13 | 92% | 7 | 86% | 20 | 90% |
| L-L | Far | 183 | 99% | 5 | 100% | 188 | 99% |
| L-L | Very Far | 112 | 100% | 16 | 100% | 128 | 100% |
| **Total** | | **576** | **98%** | **860** | **97%** | **1,436** | **97%** |

### 3.1.4 Meter-to-Transformer Prediction Accuracy

As discussed in Section 2.3.4 Data Trimming, the meter-to-transformer algorithm trims off meters in the same way as the phase identification algorithm. Additionally, it also trims off meters with wrong latitude and longitude information, and the transformers for which valid voltage data is available for only one meter on the transformer.

The latitude and longitude information are important in meter-to-transformer algorithm because a meter is always connected to one of the closest transformers, if possible, to reduce the length of service wire and hence to reduce the energy loss. Therefore, the meter-to-transformer algorithm only searches for N (N is a parameter fed into the algorithm that can be adjusted, and usually takes the value of 10, 15, or 20) closest transformer for each meter as an initial mapping. Without valid latitude and longitude coordinates, the meter-to-transformer algorithm has no starting point.

The meter-to-transformer algorithm cannot deal with transformers with only one meter either. With no information on transformers' voltages, the algorithm must summarize the voltages across all meters on the transformer as a proxy for the transformer's voltage. If a meter is the only one on a given transformer, it is always 100% correlated to itself, and the algorithm does not work.

Table 25 below summarizes the matched rate for the meter-to-transformer model. Overall, 81% of the model predictions match SDG&E's records, and the other 19% show discrepancies. The rates look similar across the two feeders. On Feeder A, the match rate is 82%, and on Feeder B, it is 79%, slightly lower.

*Table 25. Meter-to-Transformer Accuracy Rate*

| | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|
| | # Meters | % Match | # Meters | % Match | # Meters | % Match |
| Matched | 384 | 82% | 390 | 79% | 774 | 81% |
| Unmatched | 82 | 18% | 104 | 21% | 186 | 19% |
| **Total** | **466** | | **494** | | **960** | |

Figure 27 and Figure 28 on the next two pages provide visualizations of the meter-to-transformer predictions on a map. The circles represent meters, and the stars represent transformers. The lines connecting stars and circles represent the imaginary power lines. If the line is green, the meter-to-transformer connectivity matches SDG&E's records. If, on the other hand, the line is red, it means the model suggests that the meter should be connected to the transformer on the other side of the red line. If the line is grey, it means that the voltage data is not provided or is not adequate to draw a conclusion.


*Figure 27. Meter-to-Transformer Prediction on Map – Feeder A*


*Figure 28. Meter-to-Transformer Prediction on Map – Feeder B*

### 3.1.5 Meter-to-Transformer Results Discussion

As discussed in Section 3.1.2, there are some data issues that may affect the model performance. Section 3.1.2 discusses three issues: number of sample months, transformer size, and interval length or reading frequency. While the meter-to-transformer model was not tried using 10-minute interval data, this section discusses the other two factors.

*Number of meters that "represent" each transformer*

Table 26 below compares accuracy rate by transformer size. It shows how the accuracy rate increases as the number of meters on the transformer increases, from about 60% to 95% on transformers with three meters.

Transformer size is an especially crucial factor that affects the meter-to-transformer model dramatically. Since transformer meters are not measured by most utilities, transformer voltage is not available. The detour is to summarize the transformer's other meters' voltages as a proxy for transformer voltage. The meter-to-transformer model correlates each meter with all its nearby transformers' proxy voltages and assigns it to the transformer with the highest correlation.

Therefore, if there is only one meter, M, on a transformer, and when it is M's "turn" to apply correlation against each nearby transformer, its "home" transformer, the one that it is currently on, has no voltage, because there are no other meters on it.

In case of a transformer with two meters, the model is not very stable either. If either meter is mislabeled, the model prediction will not work for the other, because the model prediction for the second meter relies on the first one as a proxy for the transformer.

However, when the number of meters increases, one or two mislabeled meters does not affect the results as much, since the transformer's voltage is a summary of several meters, and the mistake is mitigated by the correct meters.

*Table 26. Accuracy Rate by Transformer Size*

| Transformer Size | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|
| | # Meter | % Match | # Meter | % Match | # Meter | % Match |
| 1 | 12 | 67% | 33 | 61% | 45 | 62% |
| 2 | 394 | 81% | 446 | 80% | 840 | 80% |
| 3 | 60 | 93% | 15 | 100% | 75 | 95% |
| **Total** | **466** | **82%** | **494** | **79%** | **960** | **81%** |

*Length of time series data*

Table 27 below compares the accuracy rate by number of sample months, for Feeder A, and Feeder B separately. The results look counter-intuitive. As the number of sample months increases, the accuracy rate declines, from 85% to 79%. However, the decline is not a big one and not statistically significant. The standard deviation of the 85% accuracy rate for "Other" group is 0.057, or 5.7%, and hence 79% is just one standard deviation away.

*Table 27. Accuracy Rate by Number of Sample Months*

| # Sample Months | Feeder A | | Feeder B | | Total | |
|---|---|---|---|---|---|---|
| | # Meter | % Match | # Meter | % Match | # Meter | % Match |
| Full | 341 | 81% | 310 | 76% | 651 | 79% |
| Almost Full | 116 | 86% | 154 | 84% | 270 | 85% |
| Other | 9 | 89% | 30 | 83% | 39 | 85% |
| **Total** | **466** | **82%** | **494** | **79%** | **960** | **81%** |

## 3.2 Updated Benefits Analysis

Initial discussions of the benefits for meter-to-transformer and phase identification hinged on the assumption that a utility can get 100% accurate field verification results if they are willing to put in the person-hours. In that context, the value of back-office solutions to meter-to-transformer and phase identification were tied to the tradeoff between time and money saved by avoiding truck rolls, and the reduced accuracy of the data-driven solution. While analyzing the phase identification results on Feeder A, however, it happened that the accuracy of the voltage correlation analysis (>98%) was greater than the accuracy of the field verifications (95%). The implications of this unexpected result are subtle, but profound. While it was estimated by SDG&E that typical field verifications also yield accuracies greater than 98%, the notion that a data-driven solution is necessarily less reliable than a manual inspection must be called into question. Moreover, use cases involving safety were mostly excluded from the initial phase identification benefits discussion on the grounds that field verifications are the most reliable source of truth. This report does not suggest that voltage correlation analysis should replace field verifications when safety is a concern. However, the results of this project indicate that a legitimate added benefit comes from the corroboration of field verifications against the output of a voltage-correlation based phase identification solution. If the phase of a meter as determined in a field verification matches the voltage correlation result, the expected accuracy is greater than 0.9996%. On the other hand, in the unlikely event that the field verification and back-office solution disagree on the phase, this prompts a more thorough reexamination in the field, potentially averting a safety issue. The other more obvious takeaway from the high phase identification accuracy is that any prospective reduction in cost and time should suffice as a reason to prefer the back-office solution over the field verifications for use cases that do not involve safety.

The results for meter-to-transformer did not significantly affect the initial benefits analysis.

# 4.0 Findings

A voltage correlation-based phase identification solution exists which achieves accuracies in the range of field verification accuracy (~98%). This accuracy comes from voltage data with a resolution of 0.15V, and an interval of five minutes collected for two years between October 10, 2018, and October 20, 2020. The voltage data were only provided for two meters per transformer within each feeder, and the solution was tested on two feeders (Feeder A and Feeder B), with different phase compositions. Feeder A had a relatively even distribution of all six possible phases (A, B, C, AB, BC, and AC). Feeder B had predominantly L-N phasing (A, B, and C).

A voltage-correlation-based meter-to-transformer solution exists, which achieved 80% accuracy when supplied with voltages for only two meters per transformer on Feeder A and Feeder B. Several accurate corrections to the field verified connectivity model were suggested, but for each correct suggestion, at least one incorrect suggestion was also generated. In addition, there were several suggestions that were both incorrect and unrelated to a "good" suggestion.

Field verification accuracies for both phase identification and meter-to-transformer were found to be lower than 100%. Field verification accuracy could only be tested using GMSV on sections of the feeder with overhead wiring. This process was time consuming, and so it was only completed thoroughly for the meter-to-transformer and phase identification examples where there was a discrepancy between the voltage-correlation result and the field verification result. GMSV analysis suggested that the utility field verification accuracy for transformer to phase connectivity was 95% on Feeder A. Feeder B did not have enough overhead wiring to merit a thorough analysis.

## 4.1 Findings Discussion

For phase identification, the findings of this demonstration are straightforward. A voltage correlation solution using data for two meters per transformer achieved accuracies on par with those of field verifications. Also, assuming the presence of all six possible phases, was an important configuration change that drastically improved the results on these feeders.

For meter-to-transformer, the limitations of the dataset led to the problem of incorrect suggestions generated for each correct suggestion. In particular, the incorrect suggestions are the necessary result of attempting to correct meter-to-transformer errors when there is only voltage data for two meters per transformer. This situation has a direct parallel in the computer science field of error correction, namely attempting to correct one-bit errors with a single bit message. The simplest code capable of correcting a single bit error in a single bit of data is the triple repetition code, which uses two parity bits for each bit of data. Similarly, when error correcting meter-to-transformer connectivity, three meters per transformer are required to detect and correct a single error. Four meters on the transformer are required to both correct a single error and detect the presence of two errors. With only two meters per transformer, the best that can be hoped for is the detection of an error. Even then, the problem is more complicated than in the binary data example. While two bits can either match exactly or mismatch completely, two meters' voltage readings can correlate anywhere on the range [-1, 1], where +1 indicates a perfect positive linear relationship – as one variable increases in its values, the other variable also increases in its values through an exact linear rule; and −1 indicates a perfect negative linear relationship – as one variable increases in

its values, the other variable decreases in its values through an exact linear rule. In practice, correlations near one are indicative of a "match", while low correlations near zero are indicative of a mismatch, but where to draw the line can be difficult. The findings from this experiment highlight some of the mathematical limitations of this approach to correcting meter-to-transformer. A voltage correlation-based meter-to-transformer solution that only has access to meter voltages will never be able to detect or correct connectivity errors on transformers with one or fewer connected meters. Such a solution will also be unable to reliably *correct* errors on transformers with two connected meters. Even in the case of three meters per transformer, this methodology would make incorrect suggestions in cases with two errors (Error Correction Code, 2021). Generally, the accuracy of this type of solution will increase with increasing numbers of meters per transformer due to the increased capability for error detection and correction. This was demonstrated in the results where meter-to-transformer accuracy was 80% for the transformers with two connected meters and 95% for the transformers with three connected meters.

Beyond the problems of error correcting, there is an inherent issue with only collecting data for anything less than *every* meter per transformer. This goes back to the value proposition for meter-to-transformer. In one example use case, utilities need to notify every customer affected by a planned outage. Failure to do so can result in the cancellation and subsequent rescheduling of the planned outage. In a use case like this, any solution that does not account for every single meter on a given transformer is inadequate. The answer to the question, "Can voltage data for two meters per transformer be used to accurately predict and correct meter-to-transformer connectivity on a given feeder?" is quite plainly, "no," without the need for any tests. Even if the voltage correlation solution could achieve 100% accuracy on that dataset, the results would still be insufficient for most of the use cases outlined in the value proposition. In almost all of them the benefit comes from knowing every meter that is attached to a given transformer. As such, the only solutions worth pursuing for meter-to-transformer are those which account for every meter. A major consideration that led to SDG&E opting to collect data for only two meters per transformer in this demonstration, was that of network traffic. To collect data for every meter on a feeder, further research should be conducted into the maximum network capacity. If it is the case that longer voltage intervals could reduce network traffic, then it is possible that the optimal data collection scenario on the given network requires longer voltage intervals. Preliminary explorations were conducted at the tail end of this demonstration which indicate that 10-minute intervals offer similar accuracies for phase identification. More research is needed in this area.

## 5.0 Conclusions

In the case of phase identification, the demonstrated technology successfully performed the desired functions by achieving accuracies comparable to those of field verification accuracies. With regards to the value proposition, adopting a similar voltage correlation based back-office solution to phase identification would save time and money, without sacrificing accuracy in every non-safety related use case. In the safety-related use cases, such a solution would also bolster utility confidence in field verified results.

In the case of meter-to-transformer, the demonstrated technology performed the desired functions with 95% accuracy when provided with voltage data for three meters per transformer and with 80% accuracy when provided data for only two meters per transformer. With data for two meters per transformer, while the technology was able to correct several of the field verified meter-to-transformer connections,

the output included a much greater number of erroneous predictions. Many of these erroneous predictions were related to "good" predictions. This was the direct result of limiting the dataset to two meters per transformer. The mathematics of error correction suggest that accuracies even greater than 95% should be expected when data is provided for more than three meters per transformer. Separately, when analyzing the use cases outlined in the value proposition for meter-to-transformer, it is evident that a back-office solution for meter-to-transformer would only be worth implementing in a production setting if it collected data for every meter per transformer. Thus, it seems natural that follow-up tests should be conducted using a dataset with voltage data for every meter per transformer. A potential hurdle is that of network bandwidth. One proposed method for mitigating network traffic is longer voltage intervals, as similar accuracies were observed when performing phase identification using every other voltage datapoint. Even without further testing, the results from this demonstration indicate that with voltage data for every meter per transformer, the accuracy achieved by the meter-to-transformer solution would likely be greater than 95%.

## 6.0 References


| References | Document Title |
|---|---|
| 1 | Wikipedia contributors. (2021, September 15). Error correction code. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:59, October 11, 2021, from https://en.wikipedia.org/w/index.php?title=Error_correction_code&oldid=1044553578 |
| 2 | San Diego Gas & Electric Company. (2018, May 1). *Application of San Diego Gas and Electric Company (U902-E) for Approval of Electric Program Investment Charge Terminal Plan for Years 2018 – 2020.* https://www.sdge.com/sites/default/files/FINAL_313383_SDG%2526E_Third_Triennial_EPIC_Application_2018-2020.pdf |
| 3 | Neelam Tyagi. (2020, December 21). What is Confusion Matrix? https://www.analyticssteps.com/blogs/what-confusion-matrix |

# PART III

Part III captures the results of Methodology B, the second of two methodologies where SDG&E worked with an external vendor.

## Part III List of Illustrations



## Part III List of Tables

## Part III List of Acronyms

| Acronym | Acronym Description |
|---|---|
| AMI | Advanced Metering Infrastructure |
| ADMS | Advanced Distribution Management System |
| AssetID | Asset Identification number |
| CIS | Customer Information System |
| CPUC | California Public Utilities Commission |
| COTS | Commercial off-the-shelf product |
| DERMS | Distributed Energy Resource Management System |
| DER | Distributed Energy Resource |
| EAM | Enterprise Asset Management |
| EV | Electric Vehicle |
| GIS | Geographic Information System |
| Phase ID | Phase Identification |
| SCADA | Supervisory Control and Data Acquisition |
| SaaS | Software as a Service |
| UI | User Interface |

# 1.0 Overview

Methodology B demonstrated use of an established, data analytics platform to ingest, analyze, evaluate, and display results for end point phase identification and meter to transformer mapping. The project was organized into three tasks.

Task 1: Phase Identification

The phase identification algorithm was validated to ensure its applicability to the selected SDG&E feeders. This step was important to ensure assumptions made (e.g., secondary network topology, number of available measurements, switch status, etc.) when developing the algorithms would apply for the selected feeders, and modifications to the algorithms could be made as necessary. This task included network data information collection from SDG&E's planning models, development of test systems, AMI data cleansing, and development of visualizations for the selected feeder.

The validation was performed using the information from the planning network model and through field verification. The former approach was used to perform the first stage of validation. The field verification was performed as part of Task 3, discussed below. The algorithm was recursively tuned using machine learning to improve the overall accuracy of the prediction.

The output of this task was phase identification of the selected feeders. This was provided through a secure access portal to the vendor platform. Supporting documentation was provided and knowledge sessions were conducted with SDG&E stakeholders.

Task 2: Meter-to-Transformer Mapping

An algorithm was configured to determine the association of meters to service transformers using AMI data. The AMI data and information collected for Task 1 was used in this algorithm. Additional data such as geographical location, impedance parameters, and existing meter mapping information was used in the optimization of the algorithm. The solution used existing static meter-to-transformer mapping information for initial validation of results during the algorithm development phase. The final validation included verification of meter-to-transformer connectivity at selected locations in the field and fine-tuning the algorithm as needed. The final validation was addressed as part of Task 3 discussed below.

The output of this task was twofold:

1. Implement analytics based on geospatial data of meters and transformers, such as methods using latitude/longitude location data to evaluate meter-to-transformer association and suggest new transformers for service points located implausibly far from their assigned transformer.
2. Implement analytics based on five-minute AMI voltage data and hourly (in some cases 15-minute) AMI consumption data, such as methods which identify meters on the same transformer based on short term voltage and current patterns on the individual meters.

The output of this task was provided through a secure access portal to the vendor application. Supporting documentation was provided and knowledge transfers sessions were conducted with SDG&E stakeholders.

Task 3: Field Validation

To determine the validity of the demonstrated phase identification and meter-to-transformer mapping algorithms, existing information in the GIS served as a reference to check the accuracy of the results. However, neither the results nor the GIS database can be perfect. Mismatches between the results from the algorithms and the GIS database are expected, which requires checking the ground truth in the form of field visits. In this task, SDG&E performed field checks to verify the phase connectivity and meter-to-transformer associations at selected locations.

The output of this task was field validation and application configuration. Secure access to the vendor application was provided in order to review the validated Supporting documentation was provided and knowledge transfers sessions were conducted with SDG&E stakeholders.

# 2.0 Methodology Approach

## 2.1 Supporting SDG&E Infrastructure and Data Requirements

The solution was configured in a secure vendor hosted environment, hence there were no specific infrastructure requirements for the project scope. The following data was used in this demonstration:

- One year of SCADA (voltage and consumption) data
- One year of five-minute AMI (voltage) read data available for four circuits/feeders (both for three-wire and four-wire systems*)
- One year of coincidental interval load data for the set of AMI meters under study
- Applicable AMI events and exceptions data for the set of AMI meters under study
- Baseline system topology and asset relationship model (e.g., GIS, CYME, distribution network planning model, etc.)
- Geospatial information and locational data were provided for applicable assets, including but not limited to meters, transformers, substations, and medium voltage assets under study

** - The four circuits used in this methodology are identified as Feeder/Circuit A, B, C, and D. Feeder/Circuits A and B are the same circuits used across all three methodologies. While Circuits C and D are only used in this methodology.*

In addition to the input data outlined above, SDG&E provided the field collected data for Circuits A and B to evaluate the prediction accuracy. These data included the phase information and meter-to-transformer connectivity.

## 2.2 Execution of Demonstrations

As per the project tasks detailed in Section 1.0, demonstrations were carried out at four key milestones:

- Demo #1: Initial run of the algorithm
- Demo #2: Results based on field data comparisons
- Demo #3: Results from second optimized run of the algorithm against additional two circuit data
- Demo #4: Final demonstration of the complete results

In addition, SDG&E was provided access to the hosted solution for their independent review and evaluation. Due to COVID-19 restrictions, there was no opportunity to conduct onsite in-person reviews.

## 2.3 Use Case Execution

The use cases were executed within the vendor's commercial off-the-shelf platform in their secure hosted environment. The use cases were initially evaluated using assorted options of statistical algorithms to identify and reconcile transformer phase assignment and meter-to-transformer relationships. Based on the quality and availability of the circuit data, appropriate statistical algorithms such as K-means, Gaussian mixture models, and Bayesian model were chosen for the use case execution.

The use case outcomes were regressively improved through machine learning, geo clustering, and appropriate data filtering techniques to compare the signal with meter voltage (interval Vh) and association with one another and to a given transformer.

Once the computational analysis was complete, the system generated a representative connectivity model from meter to substation of the circuits. This representative model highlights differences between the computed model and the "as-found" model, where the "as-found" model represents the current GIS and distribution network planning model.

Figure 1 provides a snapshot of various visualizations available to the users of the system to review, analyze, and validate the use case results.

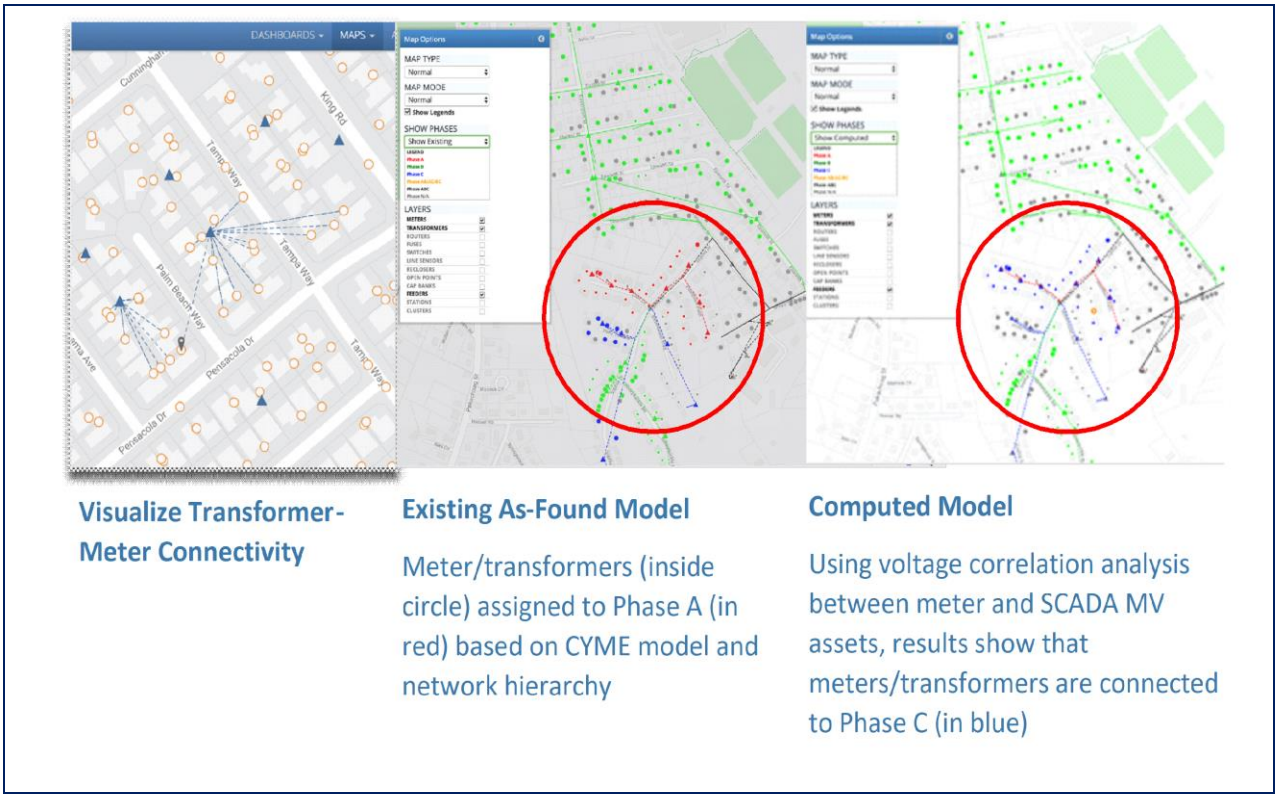



*Figure 1. Indicative Interface to Analyze Results of Automated Mapping*

One of the key elements of the use case demonstration is the ability to support an intuitive interface to various stakeholders within SDG&E. The use case outcomes were communicated via dashboards and mapping views of the circuit data. PART III, Appendix A provides snapshots of relevant interfaces that supported the overall data analysis for the stakeholders.

In addition to providing demonstrations of the results for phase and connectivity use cases, the vendor also utilized SDG&E provided data to implement and demonstrate the realization of advanced analytics such as Transformer Utilization and Voltage Management. PART III, Appendix B provides an overview of additional potential to use the AMI and SCADA data aggregated through this project.

## 3.0 Results

The summary below provides the key results from use case execution:

- Automated algorithm results for the sample SDG&E Circuits A and B closely correlated with actual field data for phase mismatches at approximately 92% predictability and connectivity issues at 89% predictability. The solution flagged approximately 9% false positives. A false positive occurs when identified meter-to-transformer mismatches are not correct.
- The solution was able to generate a high level of correlation despite the availability of less than 45% of valid/available voltage data for the sampled meters within the SDG&E territory.
- Circuits C and D were successfully processed; however, accuracy was not defined due to lack of field data.
- A major source of data gaps was associated with the availability of voltage data for the critical assets (meters and transformers).
- The solution enabled an interactive analytical interface in a secure hosted environment for easy access to the results and to conduct further investigation. The solution provided a view of data quality gaps that can be addressed to improve the overall quality of prediction.
- Availability of SCADA voltage as a reference improved accuracy of overall results.

Summary results of the use cases executed for the four circuits are presented below and grouped by two runs (iterations) of the algorithm.

*Run 1: Circuits Validated with Field Data*

Run 1 of the algorithm included analysis for phase identification and meter-to-transformer mapping for circuits A and B. The vendor first ran the algorithm using the circuits without the benefit of field validated results, and then compared the results to the field validated data. The tables below summarize the results and associated metrics:

*Table1: Summary Base Meter Statistics for Circuits A and B*

| Circuit Reference | Total Meters |
|---|---|
| Circuit A | 5,172 |
| Circuit B | 2,393 |

Importantly, there was limited AMI meter voltage data for both circuits, especially Circuit A. The table below summarizes the valid source meters that were used for the prediction based on valid data:

*Table 2: Valid Meter Data Statistics Referenced for Final Configuration*

| Reference | Meters with Voltage Data | Percentage Circuit |
|---|---|---|
| Circuit A | 674 | 13.0% |
| Circuit B | 947 | 39.6% |

The raw data from AMI and SCADA was normalized and the waveforms correlated using the vendor's algorithm to predict the phase ID mismatch and meter-to-transformer corrections. Tables 3 and 4 below summarize the results of the algorithm:

*Table 3: Phase ID Prediction Accuracy Statistics for Circuits A and B*

| Circuit Reference | Total Meters | Phase ID prediction accuracy compared with field validation | |
|---|---|---|---|
| | | # of Meters | % Accuracy |
| Circuit A | 5,173 | 284 | 83% |
| Circuit B | 2,393 | 145 | 92% |

*Table 4: Connectivity Mismatch Accuracy Statistics for Circuits A and B*

| Circuit Reference | Total Meters | Meter-to-transformer connectivity mismatch prediction accuracy compared with field validation | |
|---|---|---|---|
| | | # of Meters | % Accuracy |
| Circuit A | 5,173 | 186 | 65% |
| Circuit B | 2,393 | 129 | 89% |

### *Run 2: Circuits C & D Summary Results*

Run 2 results are based on the execution of the algorithm against the two additional circuits, C and D. Run 2 was primarily focused on prediction and enunciating the gaps in the source data that needed attention. Field validation was not carried out for Run 2 circuits; hence the accuracy of prediction is not applicable.

*Table 5: Meter Data Statistics for Circuits C and D*

| Circuit Reference | Total Meters |
|---|---|
| Circuit C | 1,422 |
| Circuit D | 357 |

*Table 6: Percentage of Successfully Processed Assets for Prediction – Circuit C*

| Circuit C | Assets processed |
|---|---|
| Successfully processed meter analysis | 30% |
| Successfully processed transformer analysis | 8% |

*Table 7: Percentage of Successfully Processed Assets for Prediction – Circuit D*

| Circuit D | Assets processed |
|---|---|
| Successfully processed meter analysis | 84% |
| Successfully processed transformer analysis | 80% |

The tables below highlight the nature of source data issues that were flagged during the analysis that degraded the overall accuracy. In some cases, there was no run due to lack of valid data.

*Table 8: Source Data Issues Categorized by Assets – Circuit A*

| Device | Issue Type | Count | Total Devices | Issue % |
|---|---|---|---|---|
| Meter | Missing Existing Phase | 2,888 | 5,173 | 55.83% |
| Meter | Missing GIS Data | 5 | 5,173 | 0.10% |
| Meter | Missing Nominal Voltage | 4,496 | 5,173 | 86.69% |
| Transformer | Missing Existing Phase | 2 | 325 | 0.62% |
| Transformer | Missing GIS Data | 1 | 325 | 0.30% |
| Transformer | Missing KVA rating | 2 | 325 | 0.60% |
| Transformer | Missing Primary Voltage | 325 | 325 | 100.00% |
| Transformer | Missing Secondary Voltage | 325 | 325 | 100.00% |

*Table 9: Source Data Issues Categorized by Assets – Circuit B*

| Device | Issue Type | Count | Total Devices | Issue % |
|---|---|---|---|---|
| Meter | Missing Existing Phase | 283 | 2,393 | 11.83% |
| Meter | Missing GIS Data | 54 | 2,393 | 2.26% |
| Meter | Missing Nominal Voltage | 1437 | 2,393 | 60.05% |
| Transformer | Missing Existing Phase | 2 | 649 | 0.30% |
| Transformer | Missing GIS Data | 2 | 649 | 0.30% |
| Transformer | Missing KVA rating | 2 | 649 | 0.30% |
| Transformer | Missing Primary Voltage | 649 | 649 | 100.00% |
| Transformer | Missing Secondary Voltage | 649 | 649 | 100.00% |

*Table 10: Source Data Issues Categorized by Assets – Circuit C*

| Device | Issue Type | Count | Total Devices | Issue % |
|---|---|---|---|---|
| Meter | Missing Existing Phase | 18 | 1,422 | 1.30% |
| Meter | Missing GIS Data | 22 | 1,422 | 1.50% |
| Meter | Missing Nominal Voltage | 412 | 1,422 | 29.00% |
| Transformer | Missing Existing Phase | 14 | 733 | 1.90% |
| Transformer | Missing GIS Data | 14 | 733 | 1.90% |
| Transformer | Missing KVA rating | 14 | 733 | 1.90% |
| Transformer | Missing Primary Voltage | 733 | 733 | 100.00% |
| Transformer | Missing Secondary Voltage | 733 | 733 | 100.00% |

*Table 11: Source Data Issues Categorized by Assets – Circuit D*

| Device | Issue Type | Count | Total Devices | Issue % |
|---|---|---|---|---|
| Meter | Missing Existing Phase | 2 | 357 | 0.60% |
| Meter | Missing GIS Data | 6 | 357 | 1.70% |
| Meter | Missing Nominal Voltage | 106 | 357 | 29.70% |
| Transformer | Missing Existing Phase | 3 | 197 | 1.50% |
| Transformer | Missing GIS Data | 3 | 197 | 1.50% |
| Transformer | Missing KVA rating | 3 | 197 | 1.50% |
| Transformer | Missing Primary Voltage | 197 | 197 | 100.00% |
| Transformer | Missing Secondary Voltage | 197 | 197 | 100.00% |

## 3.1 Results Discussion

In this methodology, the vendor executed the two use cases in a hosted environment with the data provided by SDG&E. The vendor platform provided predictions with high accuracy for the sample circuits that matched with field verified data for Circuits A & B. This was especially prominent for circuits that provided adequate AMI voltage and SCADA reference data, meeting the data input requirements of the use cases.

*Data Processing*

The project successfully processed all the data from the four circuits for both use cases based on the availability of AMI and SCADA data on the assets.

*Prediction Accuracy*

The demonstration proved that using data analytics to automatically identify the phase of meters is possible with accuracy ranging from 83% - 92%**.** Prediction accuracy for identifying incorrect transformer to meter connectivity was assessed at 65% - 89% accuracy.

Accuracy of the prediction correlated with the availability of AMI data as demonstrated in Circuit A which had only 13% coverage of voltage data resulting in lower accuracy compared to Circuit B. Prediction accuracy for Circuits C and D was not done due to nonavailability of field data.

*Data Quality*

Source data quality was a key metric that defined the overall percentage of processing and accuracy as summarized in Table 11. The main data issue was the absence of nominal voltage in approximately 60% of the assets. Results of the analysis persisted in the hosted solution with an interactive interface that supported the overall results. PART III, Appendix A provides snapshots of visualizations that supported SDG&E stakeholder analysis of the algorithm results.

## 3.2 Updated Benefits Analysis

This demonstration provided considerable insights into SDG&E circuit data, format, and highlighted the quality of data that helped articulate the following additional benefits.

- Provide advanced analytics using AMI and SCADA data to establish transformer utilization metrics. This will benefit operational teams to prioritize and more importantly, proactively resolve issues that may cause outages.
- Voltage metrics from AMI can be assessed for power quality and benefit the operations team to identify and remediate voltage quality issues.
- Accurate phase ID prediction helps with improved phase balancing. This will help the operations team to reduce losses and associated outages and improve customer satisfaction.
- Increased penetration of DERs and related impacts to circuits is a major challenge to SDG&E's service territory. Accurate phase ID prediction can assist the overall interconnection process and thereby contribute to the overall carbon offset/de-neutralization goals.

# 4.0 Findings

This methodology demonstrated analytical approaches to phase identification and meter-to-transformer mapping using two meters per transformer. The vendor first ran the circuits with no reference to field data and subsequently compared the results with the field validated results. Results were then presented in the vendor's commercial off-the-shelf application. As presented in the summary section above, the match between the algorithm and field results were closely correlated as captured in Table 3 and Table 4. For phase identification, results of 83% and 92%, and for meter-to-transformer connectivity, results of 65% and 89% were achieved for circuits A and B respectively. As surmised, the input data constraint of two meters per transformer affected the overall accuracy results in both phase identification and meter-to-transformer mapping. Publicly available studies, Pacific Gas & Electric (2018) and Wenyu Wang (2016), indicate phase identification accuracy levels ranging from 90% to 97% without the two meter per transformer constraint employed in this demonstration.

## 5.0 Conclusions

While not superior to field validated results, estimated to be above 95%, analytical approaches for phase identification using this methodology appear adequate for operational use cases as discussed in PART I, Section 6.1.2. Further, the accuracy achieved using the analytical approach in this methodology are comparable with the results from known studies, Pacific Gas & Electric (2018) and Wenyu Wang (2016), where minimal constraints were applied. However, results for meter-to-transformer mapping may not be sufficient. While there are no industry standards for meter-to-transformer mapping, 65% and 89% may be too low to warrant full-scale deployment. Minimum standards must be established by SDG&E to determine whether this solution should be pursued.

Data cleansing plays a significant role in the overall accuracy. This methodology highlighted data quality gaps that warranted further investigation and prior actions to support full-scale deployment. This is evident in the phase identification results for circuit A at 83%. The source data gap issues are provided in Table 9.

Of note, this methodology also highlighted the effectiveness of a user-friendly interface to study the results of the algorithm in an engineering view for user interactions and various optimization assessments to improve the overall performance of the algorithm. While beyond the scope of this demonstration, several other use cases were identified that went beyond phase identification and meter-to-transformer mapping that could be of benefit to SDG&E. These additional use cases are identified and discussed in PART III, Appendix B.

## 6.0 References


| References | Document Title |
|---|---|
| 1 | Pacific Gas & Electric Company. (2018, November 28). *EPIC 2.14 Automatically Map Phasing Information*. https://www.pge.com/pge_global/common/pdfs/about-pge/environment/what-we-are-doing/electric-program-investment-charge/PGE-EPIC-Project-2.14.pdf |
| 2 | Wenyu Wang, Nanpeng Yu, Brandon Foggo, and Joshua Davis. "*Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data*", 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). |

## Part III Appendix A – Automated Mapping Interactive Interface

The vendor platform provides an indicative interface for business stakeholders to visualize and investigate the results of the algorithm for computed phase vs. assigned phase. Mapping based reviews of the circuits is an effective way to validate the results and compare the results in a spatial context.

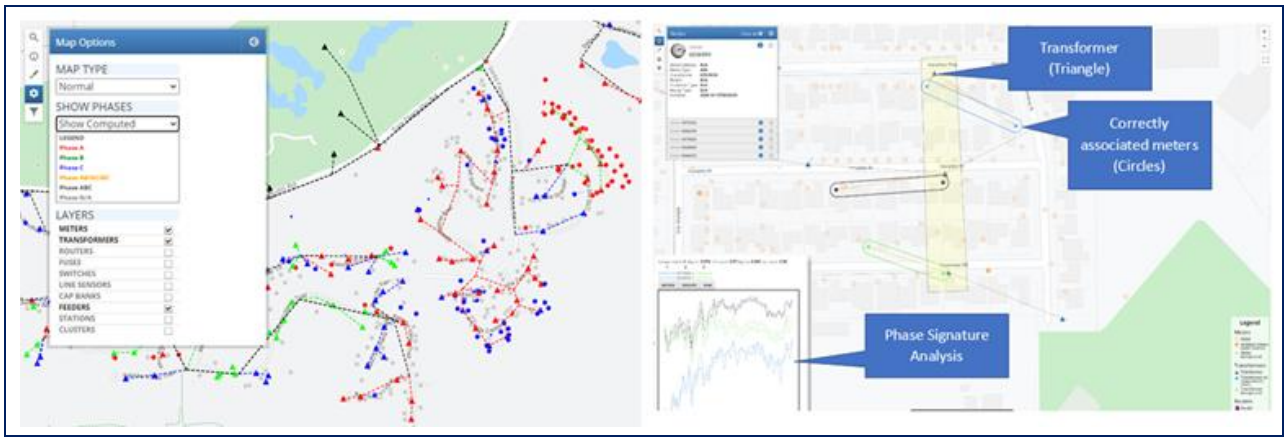



*Figure 2: Visualization of Results of the Algorithm in Interactive Mapping Views*

The key to understanding the algorithm is to visualize the voltage analysis in a convenient graphical format. Figure 3 below provides a sample view of the options available for the users to investigate the results of AMI and SCADA analysis.

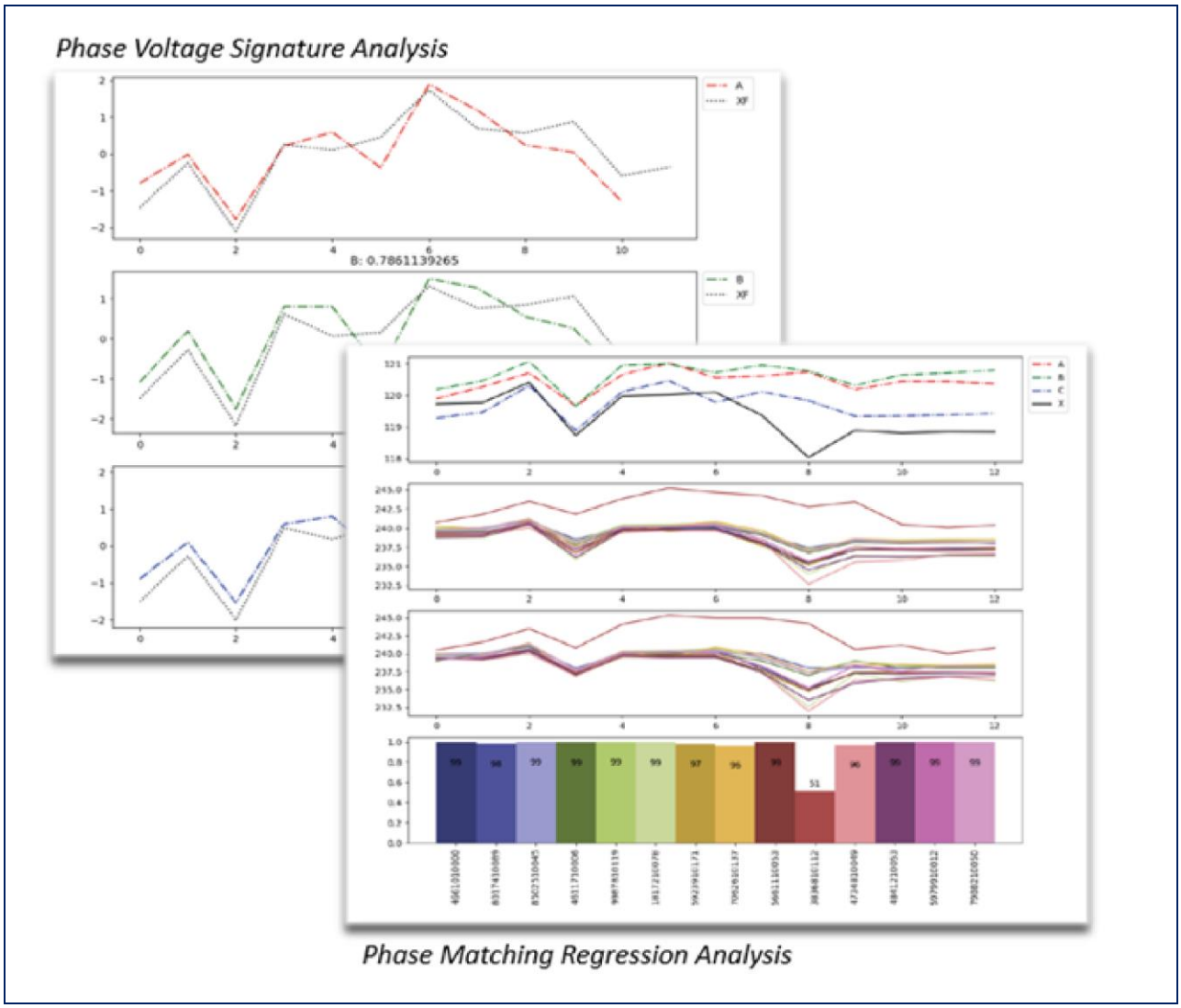



*Figure 3. Visualization of Results of the Algorithm in Interactive Mapping Views*

Configurability is a critical feature to review, optimize and iterate the results of algorithm until the desired accuracy is reached. Figures 4 and 5 below demonstrate the options to tune the algorithm interactively and visualize the results.

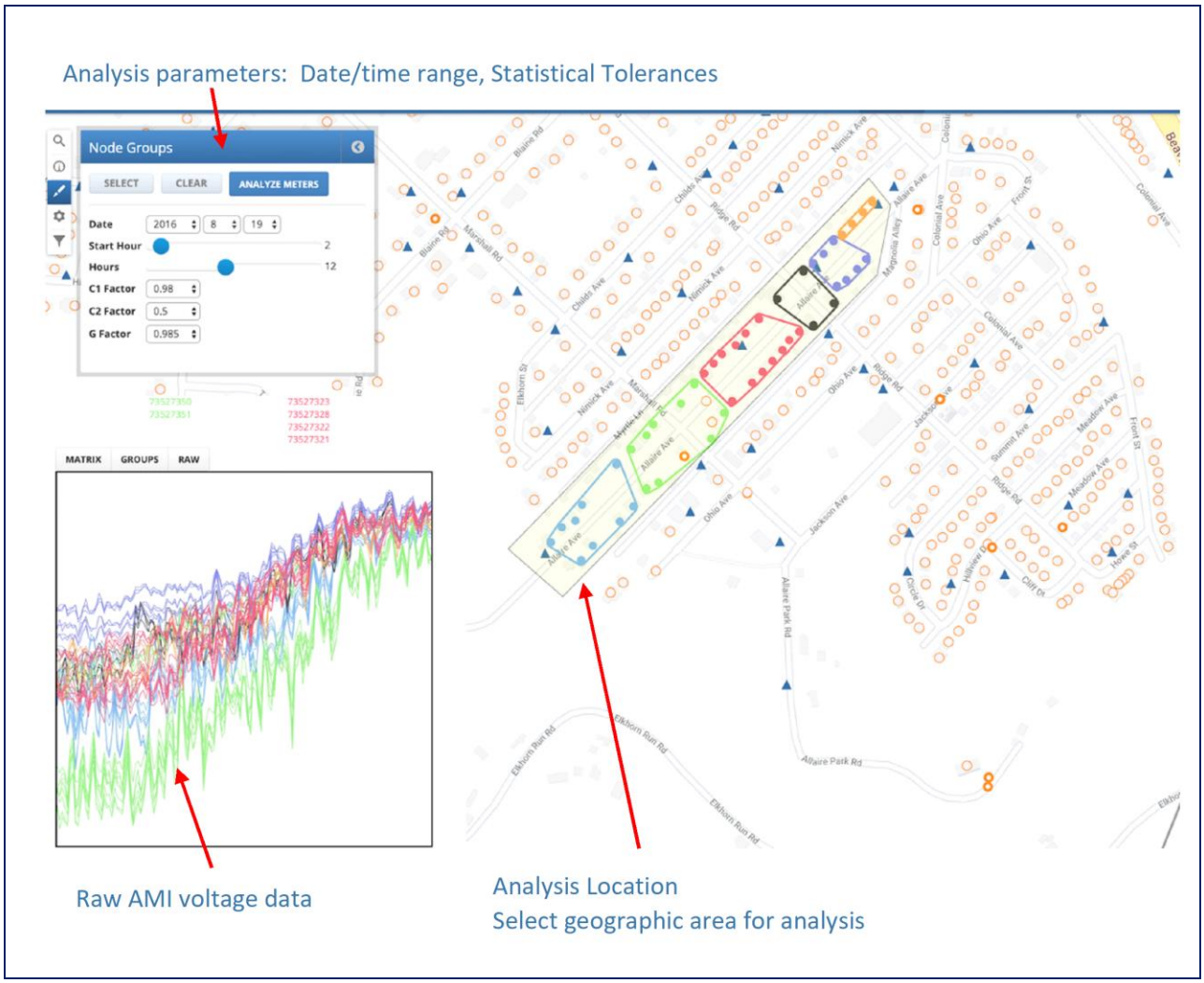



*Figure 4: Options to Tune and Optimize the Algorithm - Engineering Tool 1*

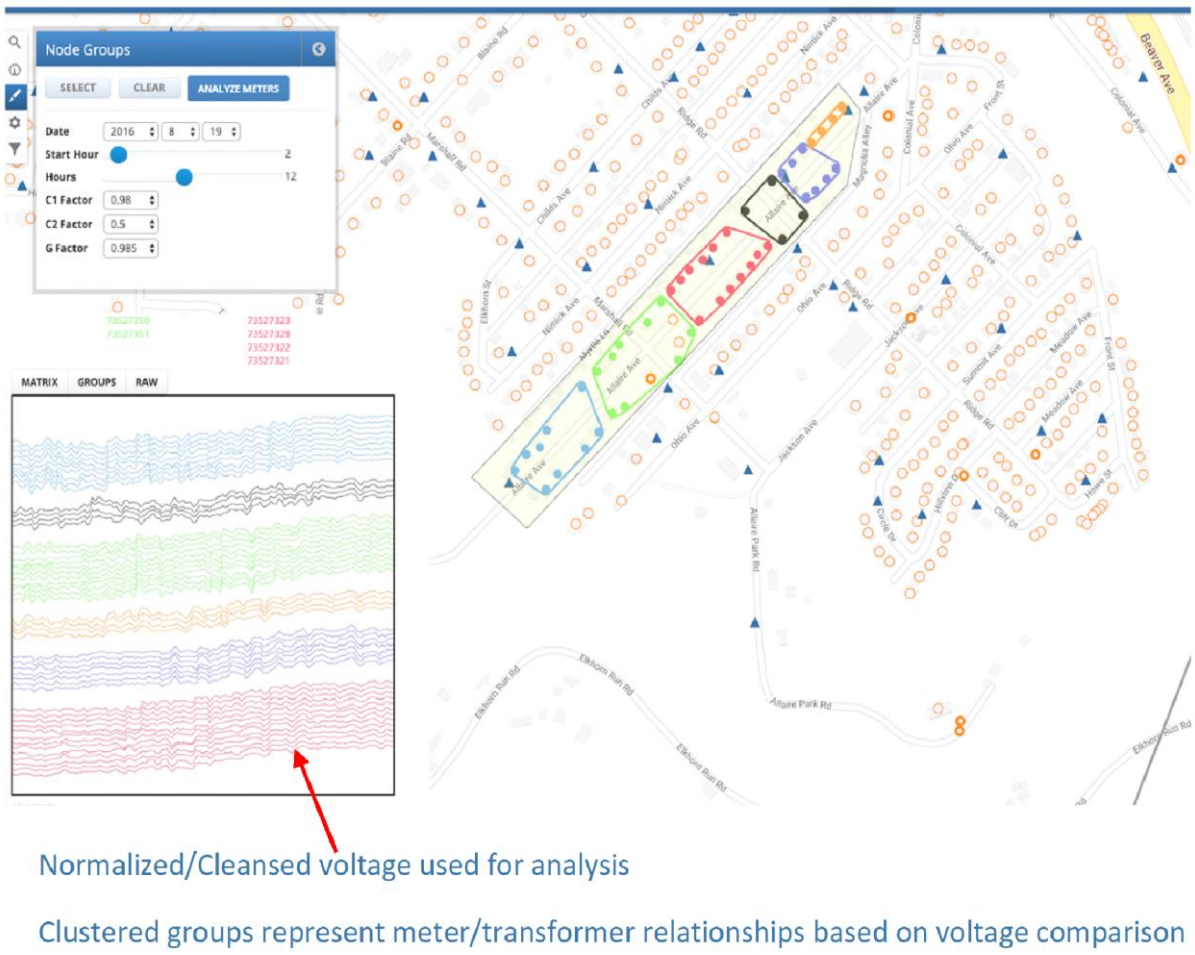
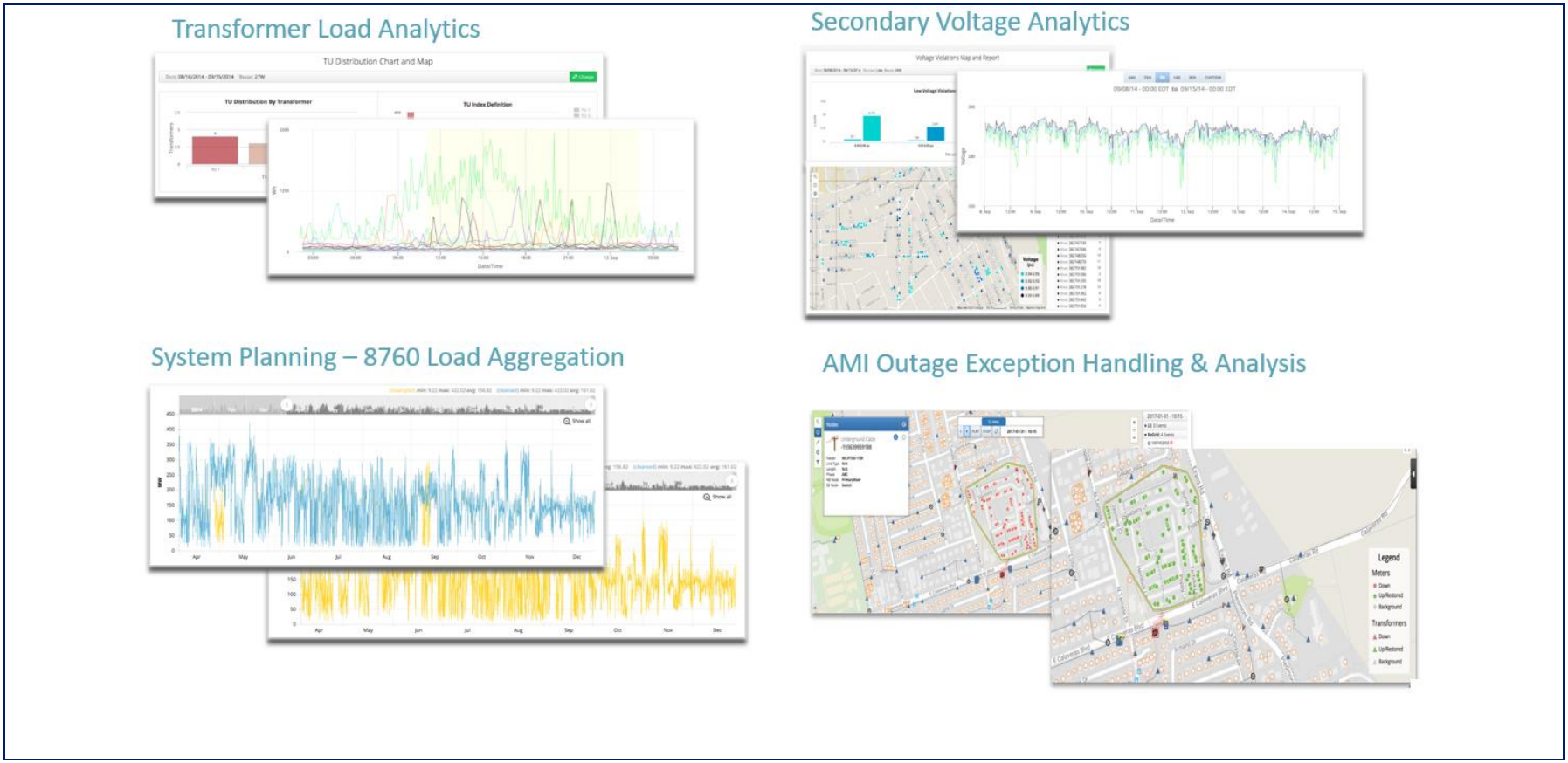



*Figure 5: Options to Tune and Optimize the Algorithm - Engineering Tool 2*

## Part III Appendix B – Automated Mapping Extended Use Cases

This section highlights additional use cases that can be supported by extension of the AMI and SCADA data for operational purposes and thereby improve the overall return on investment in the project. These are recommended use cases for SDG&E's validation as part of the next steps.



Transformer Load Analytics (Transformer Utilization)

Use cases: Asset management, planning

Combining the increased adoption of distributed generation and electric vehicle (EV) systems on the grid with an aging utility infrastructure, the importance of monitoring transformer utilization is a growing necessity to mitigate against accelerated loss of life and prevent downstream impacts to customer reliability. The Transformer Utilization (TU) analytics focuses on evaluating overload conditions on individual transformers. The system collects meter load data and assesses the aggregate load at the transformer level. The meter-to-transformer connectivity hierarchy uses a known relationship model or leverages the output from the identified use cases. The TU application module interprets the connected aggregate load data and assesses the load condition based on load percentage above nameplate rating and relative time duration at each overload state. When ambient, top-oil, and/or winding hotspot temperature is available, we can overlay the temperature data to characterize the overload condition.

The analytics uses a ranking system based on a calculated "severity index" to prioritize overloaded transformers. A severity index is assigned to transformers that have encountered an extensive overload state over the given analysis period. As overload conditions are not necessarily always an issue, the system's severity index calculation uses a weighting algorithm to only rank and classify noteworthy overload conditions. The weighting system uses a combination of analyzing load percentage above nameplate rating, duration at each percentage level, and system peak load data.

The primary factors used to analyze and characterize transformer overload are aggregate meter load percentage above nameplate rating and time duration at each overload state.

System Planning Analysis

Use cases: Annual and monthly planning and capacity analysis support

System Planning focuses on profiling and normalizing interval load data at different aggregation points at the substation and/or along the feeder. With the increased adoption of distributed energy resources, the ability to understand its impact to the load curve has become more critical for system planning purposes. The data available within SDG&E from the current project can be used to aggregate and correlate against the network topology to a common point. The common point can be at the substation and/or a strategic node along the feeder, such as a recloser. After the aggregation is complete, an appropriate algorithm can be configured to identify load anomalies (i.e., switching events, load transfers, etc.) and normalizes the data to generate a representative load profile. The normalization process incorporates historical load pattern, seasonal load trends, and weather data to fill atypical load behavior, as well as using DER load shape library to disaggregate between gross and net loads.

Voltage Analytics

Use cases: Voltage quality, compliance and planning. With the emergence of distributed energy resources, increased customer demand, and continued initiatives in energy efficiency programs, secondary voltage management has become a critical task for electric utilities. Voltage Management Analytics is designed to monitor, analyze, and identify voltage issues for individual meters, transformers,

and/or the full feeder. The purpose of the approach is to characterize voltage profiles along the feeder and identify meter and transformer voltage violations as defined by ANSI C84.1 specifications. Voltage issues are summarized and presented using various tabular reports, charts, bar graphs, and geospatial views.

The approach can establish voltage profiles for any grid-connected device with available time series-based voltage data. Grid-connected devices include but are not limited to smart meters, transformers, line sensors, capacitor banks, and line regulators. The algorithm uses a ranking system based on a calculated "severity index" to prioritize voltage performance relative to each transformer based on the voltage.  measurements from connected meters. A severity index is assigned to transformers that have meters that have experienced a voltage violation (+/- 5% from nominal) over the given analysis period. To eliminate noise and excessive flagging of voltage violation anomalies, the severity index calculation uses a weighting algorithm to only rank and classify noteworthy voltage violations. The weighting system uses a combination of analyzing voltage magnitude, duration of voltage violation, frequency of violation, and coincidence violation between neighboring meters.

Voltage Management

Use cases: Asset management, capacity planning and additional analytics for AMI outage events

The AMI Outage Exception Analytics focuses on analyzing real-time outage exceptions received from meters and SCADA assets. Given the value of real-time outage notifications, system operators can be better equipped to utilize this data to improve operational efficiency and response time to outages. The algorithm processes the outage exceptions and uses a series of steps to confirm the outage, time bound the outage, quantify impacted customers, and identify a fault perimeter location.

# PART IV

Part IV captures the results of the internal methodology executed by SDG&E personnel and describes the demonstration based on publicly available studies.

## Part IV List of Illustrations



## Part IV List of Tables



## Part IV List of Acronyms

| Acronym | Acronym Description |
|---|---|
| AMI | Advanced Metering Infrastructure |
| DER | Distributed Energy Resources |
| EPIC | Electric Program Investment Charge |
| EV | Electric Vehicle |
| GIS | Geographical Information System |
| L-L | Line to Line (phasing) |
| L-N | Line to Neutral (phasing) |
| MDMS | Meter Data Management System |
| O&M | Operations and Maintenance |
| OMS | Outage Management System |
| Phase ID | Phase Identification (meter to phase connectivity) |
| RFI | Request for Information |
| RFP | Request for Proposal |

| Acronym | Acronym Description |
|---|---|
| SCADA | Supervisor Control and Data Acquisition |
| SDG&E | San Diego Gas and Electric Company |

# 1.0 Overview

The purpose of this internal methodology was to assess and demonstrate pre-commercial analytical approaches to phase identification to enhance utility system operations. Unlike the methodologies in Part II and Part III, this methodology focused on the internal effort by SDG&E personnel to identifying endpoint phasing based on publicly available studies. No work was done on meter-to-transformer mapping in this pre-commercial demonstration.

# 2.0 Methodology Approach

A single use case was demonstrated – analytic phase identification. This involved making predictions for the meter-to-phase connectivity within a feeder by using an internally developed algorithm based on the research and results in references, (Wenyu Wang, 2016) and (Roelofsen, 2018). On a feeder, electricity is typically distributed using three powered lines. Each line has a different phase of alternating current. Often these three phases are labeled A, B, and C. In between the powered distribution lines and residential electric meters, transformers are used to reduce voltages to operating levels. There are many ways to wire transformers between the power distribution lines. The result is the low voltage wires coming from a single-phase transformer can transmit electricity in one of six possible phases (A, B, C, AB, BC, AC), depending on the wiring configuration of the transformer. These phases are split into two groups. The L-N phases occur when the transformer is wired between a powered distribution line and a neutral line (phases A, B, and C). They conduct electricity with a phase corresponding to the phase of the powered line. The L-L phases occur when the transformer is wired between two powered distribution lines (phases AB, BC, and AC). They conduct electricity with a phase corresponding to the difference between the two powered distribution lines. Utilities typically keep track of the transformer to phase connectivity because all the meters connected to a single-phase transformer share the same phase. For this use case, however, meter-to-phase connectivity is predicted. The primary reason for this is the meter-to-transformer connectivity is also in question. Accurate meter-to-phase connectivity is sufficient for use in phase balancing. Meter-to-phase connectivity can also be used to cross-validate meter-to-transformer connectivity. Another reason that meter-to-phase connectivity is predicted, and not transformer-to-phase connectivity, is that voltages are not metered on the transformers.

## 2.1 Software Requirements

The internally developed phase clustering algorithm uses Python 3 for circuit analysis and the Julia Programming Language for voltage data analysis. Appendix A is provided as pseudocode to trace the logic of the algorithm. Voltage data is stored and preprocessed in a local SQLExpress instance. Results from the clustering algorithm and voltage data are displayed in a Power BI report.

## 2.2 Supporting SDG&E Infrastructure and Data Requirements

The internally developed clustering algorithm requires data found in the SDG&E OSI PI time series database, the SDG&E ESRI GIS system, and the SDG&E Engineering Data Warehouse.

The algorithm requires a voltage data extract from the OSI PI system containing five-minute interval Volt Hour readings over a time range. These data are stored in a relational database and used by the main

algorithm written in Julia. The data are normalized to a per-unit voltage value based on nominal voltage. Meters with long periods of stale or missing data are removed from the analysis.

From the GIS system, an extract is pulled that contains data for service transformer and conductors in the GeoJSON format (GeoJSON, 2021). The attribute data from these sources are analyzed in the Python circuit tracing script to identify single phase branches in the circuit.

From the engineering data warehouse, a metadata extract is pulled which is used to map meter IDs to service transformer IDs. These data are stored in a relational database and joined with the voltage data.

## 2.3 Execution of Demonstrations

The algorithm used is a k-means constrained clustering algorithm. A k-means clustering algorithm is defined by (Pedamkar, 2020) as an unsupervised learning method that uses an iterative process in which the datasets are grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space, it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum. At this position the center of the cluster is the arithmetic mean of the data points in the clusters.

The k-means algorithm is enhanced by generating constraints for the circuit programmatically using a GIS extract. The program starts at a chosen structure ID and recursively follows GIS conductors based on structure IDs. The program marks each conductor segment with a group ID. The algorithm reuses the last group ID only if both conductors are single phase and there are no multi-phase conductors attached to the structure. The program then outputs a mapping of transformer IDs to group IDs. The constraint data is then fed into the k-means clustering algorithm. Python and Julia pseudocode is contained in PART IV, Appendix A.

# 3.0 Results Discussion

## 3.1 Methodology Limitations

The internally developed phase clustering algorithm was designed to give the team a baseline metric of results accuracy using simple time-series clustering. Missing features of the internally developed solution compared to the vendor products are important to consider before looking at the accuracy of the results. The two main limitations of this algorithm are 1) the results output is provided as “phase groups” rather than “phase IDs” and 2) the analysis is currently restricted to single-phase, line to neutral meters.

The first limitation could likely be overcome by bringing in time series bus voltages and SCADA device voltages. After calculating a time series correlation coefficient between each phase group voltage and known phase voltages, a map between the predicted phase group and the actual phase ID can be created. With reasonable confidence in the existing GIS data, identifying phase IDs could also be accomplished by mapping the groups to the IDs that result in the least mismatched data. For demonstration purposes, this limitation was overcome using a manual analysis step.

The second limitation requires a more detailed understanding of the GIS data to ensure properly created constraints for meters electrically connected to more than one phase. The same clustering algorithm could then be run against the line-to-line meters to group them.

## 3.2 Results

The accuracy of the clustering algorithm for single phase meters on Feeder A and Feeder B was 72.5% (190/262) and 95.5% (741/776) respectively. Table 1 and 2 below provide the breakdown between predicted phase and true phase for each feeder.

*Table 1. Feeder A Predicted vs. True Phase*

| Prediction Phase | True Phase | Count of Prediction |
|---|---|---|
| A | A | 42 |
| A | B | 2 |
| A | C | 31 |
| B | A | 4 |
| B | B | 94 |
| B | C | 11 |
| C | A | 18 |
| C | B | 6 |
| C | C | 54 |
| *Total* | | *262* |

*Table 2. Feeder B Predicted vs. True Phase*

| Prediction Phase | True Phase | Count of Prediction |
|---|---|---|
| A | A | 235 |
| A | B | 11 |
| A | C | 3 |
| B | A | 1 |
| B | B | 269 |
| B | C | 0 |
| C | A | 13 |
| C | B | 7 |
| C | C | 237 |
| *Total* | | *776* |

When viewing the field confirmed phasing from Feeder A, it is clear that the algorithm struggled grouping A and C phase meters because of the similar voltage signature of the two phases. The reason for this can be seen visually when comparing Figure 1 and 2 below. Feeder A phases are much more tightly coupled than Feeder B phases.

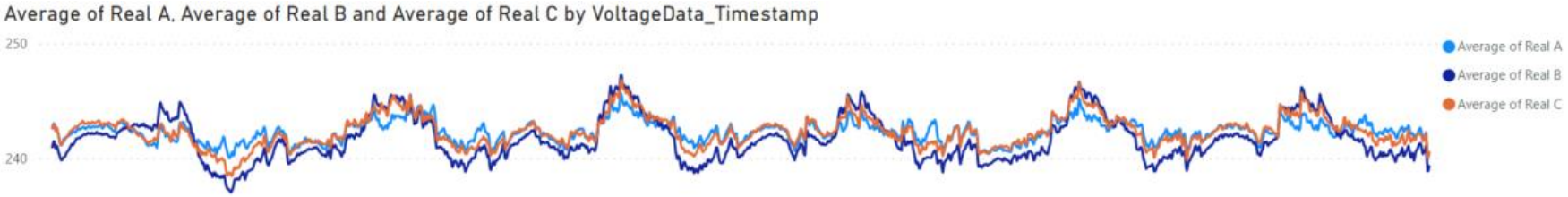



*Figure 129. Feeder A Confirmed Phase Groups from 10/21/2018 to 10/26/2018*

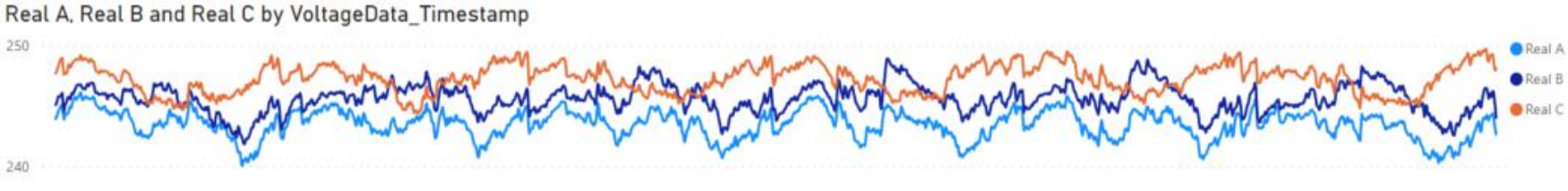



*Figure 2. Feeder B Confirmed Phase Groups from 10/21/2018 to 10/26/2018*

When running the algorithm, a metric is calculated indicating how consistently the meters were placed into the same bucket through multiple time ranges. For Feeder B, the consistency was 96.0%, while for Feeder A the consistency was only 84.8%. A confidence metric calculated based on the average Pearson correlation coefficient between each meter and each group it is not a member of, would also be a good indication of confidence in prediction accuracy. Meters that have close voltage profiles to other groups on average indicate the clustering algorithm is probably not a good choice for phase identification on a circuit.

# 4.0 Findings

*Lessons Learned*

The internally developed phase identification algorithm provided significant insight into the implementation of a simple, scalable solution for phase ID. One of the main lessons learned from this proof of concept is that a simple k-means clustering approach can be effective at phase ID in certain circumstances. Analysis of the results also indicates that it is possible to use confidence metrics from the results to decide whether voltage groups are unique enough to effectively cluster meters based only on voltage readings. Lastly, the results have sparked ideas on improving accuracy and confidence with different data preprocessing.

The difference in accuracy between the two circuits indicates that constrained k-means clustering is very effective in grouping meters within circuits that have a large sample size of meters and have distinct voltage signatures for each phase. For well suited circuits, accuracy in the mid-90% range can be achieved with this methodology. If changes to the GIS phasing data are only made in cases where the meter voltage signature is very similar to its group voltage signature while being relatively far from the other group voltages, confidence would be very high that the GIS model is being improved.

As discussed in the results section, one of the important findings from this proof of concept is there are circuits for which clustering is not as effective. In the case of Circuit A, there was difficulty in grouping meters between A and B phase because their voltage signatures were very similar. More detailed confidence metrics must be calculated with the results and confidence thresholds established so that phasing information isn't changed to the wrong value.

Developing the phase clustering algorithm and analyzing the results have also brought forth different ideas on improving its accuracy. There are additional ways to identify constraints based on the GIS model that can be applied to the algorithm. By increasing the number of meter groupings, single meter spikes average out and the results are improved. There are also likely meters that can be guaranteed to not belong to the same phase based on the GIS model. These additional restrictions could improve performance of the analysis and improve accuracy. In addition to working on more GIS model analysis, improvements can be made to the algorithm by better selecting a time window to run the algorithm against. Depending on the time of year, weather, and many other factors the true phasing can vary in similarity. With a long time range of voltage data available, the algorithm can be run multiple times and the results with the most distinct phase characteristics can be used. The lessons learned from this process

are valuable findings that can be used when analyzing an in-house developed phase identification algorithm or a vendor product using a similar method.

## 5.0 Conclusion

The internally developed phase identification algorithm has proven to be a very good baseline for which other vendor products can be compared. Keeping all preprocessing and GIS analysis done programmatically means the algorithm is easily scalable to all circuits without extensive cost. By understanding the shortcomings of the developed method, the team better understands where similar unsupervised methods may have accuracy issues. Analysis of the two circuit results has shown the importance of developing metrics and thresholds related to confidence of the groupings. It is also clear that more effort needs to be invested into static analysis of the GIS model to improve accuracy on circuits that are not as well suited to time series voltage clustering.

## 6.0 References


| Reference | Document Title |
|---|---|
| 1 | Wenyu Wang, Nanpeng Yu, Brandon Foggo, and Joshua Davis, "*Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data"*, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). |
| 2 | Pjotr Roelofsen, 2018, "Time series clustering", Vrije Universiteit Amsterdam Faculty of Science, De Boelelaan 1081a 1081 HV Amsterdam, https://www.math.vu.nl/~sbhulai/papers/thesis-roelofsen.pdf |
| 3 | Priya Pedamkar, 2020 "K-Means Algorithm", EDUCBA, https://www.educba.com/k-means-clustering-algorithm/ |
| 4 | Wikipedia contributors. (2021, November 3). GeoJSON In *Wikipedia, The Free Encyclopedia*. Retrieved 11:59, December 5, 2021, from https://en.wikipedia.org/wiki/GeoJSON |

# Part IV Appendix A – Python and Julia Algorithm Scripts

*Note – This pseudocode cannot be run as-is, rather, it provides the logic that can be used in other programming languages and was derived from the publicly available studies* - (Wenyu Wang, 2016) and (Roelofsen, 2018).

## *Python Circuit Tracing Algorithm*

```
PROCEDURE MAIN
SET firstStructureID to the structure ID where the tracing should start on the circuit
SET conductors to the contents of the parsed conductor JSON file
CALL generateConstraints function with firstStructureID, conductors
SET transformers to the contents of the parsed transformer JSON file
FOR transformer in transformers
   SET allNodes to a list of conductors upstream or downstream from the current structure
   IF allNodes have matching groupIDs THEN
      APPEND transformer.ID to an array stored in groupData[groupID]
   ENDIF
ENDFOR
OUTPUT JSON file with the transformerID to groupID mapping


PROCEDURE generateConstraints
INPUT currentStructureID, conductorMap, branchID DEFAULT 0
SET conductor to conductorMap[currentStructureID]
SET conductor.branchID to branchID
SET connectedNodes to a list of conductors with IDs matching conductor.UPSTREAMSTRUCTUREID and
conductor.DOWNSTREAMSTRUCTUREID
SET maxPhases to the maximum number of phases designated to conductor and connectedNodes
FOR node IN connectedNodes
   CONTINUE IF node.branchid is set
   IF maxPhases is 1 THEN
      CALL generateConstraints with node.ID, conductorMap, branchID
   ELSE
```

CALL generateConstraints with node.ID, conductorMap, getNextBranchID()

ENDIF

ENDFOR

*Julia Time Series Voltage Clustering Algorithm*

PROCEDURE MAIN

SET data to a DataFrame containing columns MeterID, Timestamp, Vh, and TransformerID for a 3 month time period

SET constraints to the output data from the circuit tracing algorithm

JOIN data with constraints adding a new groupID column to data

SET dataGroup to data grouped by the id column

SET windowVotes to a zero matrix with dimensions (length(groupID), 3)

SET windowBreaks to a list of evenly spaced ranges for which the clustering algorithm will operate on

FOR windowRange IN windowBreaks

SET analysisDF to an empty dataframe

Iterate through each group in dataGroup and append the subset of data to analysisDF

SET groupedAnalysisDF to analysisDF grouped by the groupID column

CALL cluster WITH groupedAnalysisDF, 3, length(windowRange)

Calculate which permutation of the returned groups aligns most closely with windowVotes

Apply the permutation to the returned groups and add the votes to windowVotes

ENDFOR

OUTPUT a CSV file with each meter ID and the prediction containing the most votes from its row in windowVotes

PROCEDURE CLUSTER

INPUT groupedDataFrame, clusterCount, seriesLength

SET estimatedClass to a random selection of 1:clusterCount of size length(groupedDataFrame)

SET iterationNumber to 0

WHILE true

```
    INCREMENT iterationNumber
    Calculate the average voltages for each estimatedClass group at each time in the analysis window
    FOR gdf in groupedDataFrame
        Calculate the sum of the Pearson distance between each group member and each cluster group
        SET estimatedClass[groupIdx] to the group with the smallest Pearson distance
    ENDFOR
    IF no group changes were made then break out of the loop
END WHILE
RETURN estimatedClass
```

# PART V

PART V summarizes Module 2 project outcomes.

## Part V List of Tables

# 1.0 Module 2 Findings

All three methodologies agree that automatic phase identification is achievable at acceptable levels of accuracy using data from only two meters per transformer, as the project module sought to confirm. Meter-to-transformer connectivity, however, proved less precise with demonstrations revealing added complexity when the use case included correction to meter-to-transformer mismatches. A summary of findings by methodology is provided in Table 1.

*Table 1: Summary Findings by Methodology*

| | Methodology A | | Methodology B | | Internal Methodology | |
|---|---|---|---|---|---|---|
| | Circuit A | Circuit B | Circuit A | Circuit B | Circuit A | Circuit B |
| **Accuracy Phase ID** | 98% | 97% | 83% | 92% | 72.5% | 95.5% |
| **Accuracy Meter-to-Transformer (two connected meters)** | 82% | 79% | 65% | 89% | NA | NA |
| **Accuracy Meter-to-Transformer (three connected meters)** | 95% | | NA | | NA | |
| **Key Challenges** | For meter-to-transformer mapping, a sufficient number of connected meters is necessary. At two meters per transformer, it is possible to detect the presence of a single error, but it is not possible to correct that error without introducing more errors into the system. | | Quality of source data and data availability impacts accuracy results | | Phase ID limited to line to neutral phasing. Line to line phase identification will require future research. | |
| **Lessons Learned** | For phase ID, the voltage correlation solution using data for two meters per transformer achieved accuracies on par with those of field verifications<br>More tests are required to determine if voltage data for | | The demonstration proved that using data analytics to automatically identify the phase of meters is possible.<br>For meter-to-transformer, accuracy of the prediction correlated with the availability of AMI data as | | The clustering algorithm can be effective at phase ID.<br>Clustering is not as effective where meters have similar voltage signatures (A and B phase). | |

| | Methodology A | | Methodology B | | Internal Methodology | |
|---|---|---|---|---|---|---|
| | Circuit A | Circuit B | Circuit A | Circuit B | Circuit A | Circuit B |
| | two meters per transformer can be used to accurately predict and correct meter-to-transformer connectivity on a given feeder | | demonstrated in circuit A which had only 13% coverage of voltage data resulting in lower accuracy compared to Circuit B. | | | |
| **Future Considerations** | To achieve data collection on every meter on a feeder, further research into the maximum network capacity is recommended.<br>If it is the case that longer voltage intervals could reduce network traffic, then it is possible that the optimal data collection scenario on the given network requires longer voltage intervals. | | Extension of the utilization of AMI and SDADA data for operational purposes beyond the Module 2 project scope (provided in PART III, Appendix B). | | Use of different data pre-processing techniques for improving accuracy and confidence levels. | |

## 2.0 Updated Value Proposition

If commercially adopted, each of these methodologies could improve workforce safety by reducing the frequency at which SDG&E employees and contractors must field verify phase ID, hence lowering the potential of hazard. In cases where manual verification is still needed, such as in situations where correct connectivity information affects safety, better understanding of the circuit distribution will help to streamline the process. The data analytics approach could also increase the safety for SDG&E customers by enhancing grid reliability.

Added value of the approach is improved reliability and power quality and improved performance of the distribution system by enabling better phase balancing and ensuring transformers are not over or underloaded by using an analytical approach. Accurate connectivity models also support a growing body of advanced data analytics for solving problems from load management issues with electric vehicle (EV) to outage management.

By reducing system electrical losses and enhancing grid efficiency, accurate connectivity models will help reduce the need for electric generation, thereby also reducing greenhouse gas emissions.

If operationalized, this project will lead to more efficient, reliable, and safe electric power, with lower cost and higher quality. All of these are consistent with the objectives of the EPIC program and provide value to SDG&E's customers.

# 3.0 Commercialization

The following discussions offer guidance and cost estimates for commercialization of the vendor and internally developed methodologies.

## 3.1 Methodology A

Commercial adoption of the system used in this methodology should include ongoing analysis of phase identification and meter-to-transformer on a regular basis. This analysis is important to ensure utility enterprise systems that increasingly rely on these data are correct and up to date as new customers come online, crews perform maintenance, and proactive activities like phase balancing and feeder reconfigurations occur.

Commercial cost components:

1) Data loading package, and initial endpoint configuration and subscription to AMI headend system. This is a one-time cost to set up a service to load initial and ongoing AMI measurements.
2) One-time cost: approximately $100,000
3) Support services for systems integration and data validation
4) Annual services contract: approximately $75,000 - $120,000
5) Software as a Service (SaaS) subscription to cloud-based analysis software, including web-based user interface and reporting: Annual SaaS subscription approximately $700,000 - $1,200,000
6) Ongoing analysis of phase identification and meter-to-transformer connectivity
7) Internal resources

## 3.2 Methodology B

Strong emergence of data analytics, information technology (IT) and operations technology (OT) convergence is helping utilities capitalize inherent value of data aggregated and maintained in AMI, SCADA, GIS, enterprise asset management (EAM), and customer information systems (CIS). The demonstrated methodology highlighted the commercial elements of various components for consideration. Table 2 below highlights the commercial implications and consideration for full-scale implementation of this methodology.

*Table 2. Methodology B Commercial Considerations*

| Project Component | Commercial Implications | Recommendations and Opportunities |
|---|---|---|
| Technology platform to support data ingestion, processing, and aggregation | SDG&E to consider investment in base technology platforms that support enterprise grade ETL, hosting the solution in a big | Commercial options exist for either hosted services as demonstrated in the project or investment in the full license of |

| Project Component | Commercial Implications | Recommendations and Opportunities |
|---|---|---|
| | data platform and visualizing in an intuitive interface | the platform within the SDG&E on-premises environment. |
| Proven algorithm that can be configured, improved, and include visualization of results | SDG&E to consider proven algorithm that can be easily configured and scaled for their service territory | Market offers configurable algorithms that can be deployed on-premises or as demonstrated through this project in a SaaS model. |
| Prepare, validate, and define data transport to SDG&E's source data (AMI, SCADA, GIS etc.) | SDG&E to plan to build the data bridges to continuously move data to the platform solution | A recommended approach is to plan for professional services to build scalable bridges to ingest data from the source system. Ideally, the solution chosen for the automated mapping will offer the capabilities for SDG&E's consideration. |
| Address gaps in data quality that may limit the accuracy of automated approach | SDG&E to carefully evaluate the source data quality that might prevent the required level of accuracy for automated mapping. This is a key component that should be budgeted and addressed effectively. | Level of data quality and source of the issue drives the cost of resolution. SDG&E should also carefully consider the critical data gaps vs. non-critical in consultation with the chosen partner solution for optimal accuracy in prediction. SDG&E should also evaluate the needs for deploying sensors at bellwether asset locations to address the gaps in voltage data. |

## 3.3 SDG&E Internally Developed Methodology

The unsupervised nature of the internally developed algorithm allows scale up of phase identification to all circuits for a relatively low cost. All the steps taken during the data pre-processing phase would be trivial to automate for any number of circuits. An estimated 370 hours of work with an internal developer, GIS analyst, and engineering resource would be sufficient to automate the current algorithm to run for all circuits.

In the results discussion in Part IV, two limitations were discussed that would need to be addressed for commercialization. The first limitation of the groupings to phase translation would require more collaboration between engineering groups and IT. If SCADA data or a static circuit analysis tool proves sufficient, this solution would take an estimated 90 hours.

The second limitation to overcome with this solution is to include line to line and polyphase meters in the analysis. This could be overcome easily with some additional analysis of the circuit and meter metadata. Incorporating this into the existing algorithm would take an estimated 160 hours.

Prior to commercialization, some additional postprocessing metrics would be required to gauge confidence in the results of the clustering algorithm. The time series clustering algorithm works much better in circuits with a greater distance, or differentiation between the voltages of each group. Because of this, it is important to provide additional confidence metrics to get an idea on when the algorithm might have done a poor job at grouping meters. These simple calculations would take an estimated 10 hours of work.

The last requirement prior to commercialization is to enhance the results display to allow for more detailed analysis of the algorithm output. A Power BI and ArcGIS map layer would provide out of the box functionality to display the results geospatially with added context from existing circuit and service transformer layers. Enhancements to the result dashboard would take an estimated 120 hours of work with out-of-the box solutions although this cost could grow if a custom application with alternate functionality is required.

In total, the enhancements and changes needed for an adequate internally developed phase identification solution would start at an estimated 370 hours of work with an internal developer, GIS analyst and engineering resource. A project champion and funding sources would need to be determined if this methodology is pursued.

## 4.0 Tech Transfer Plan

The results of this project will be disseminated throughout the industry in several ways.

### *SDG&E Website*

This comprehensive final project report is the main tech transfer documentation for the project. All EPIC final project reports are posted to the SDG&E website at: https://www.sdge.com/epic. The website also includes annual updates that were made over the life of the projects. These documents are also filed with the CPUC.

### *EPIC Symposium*

The project results will be shared with California Investor-Owned Utilities through the annual EPIC symposiums. During these meetings, information on various EPIC projects is shared with the personnel from the other IOUs in the state.

*Industry Conferences and Publications*

SDG&E personnel worked with the product vendors to develop presentation material outlining the results of this report. These

presentations will be offered, as may be appropriate, for inclusion at industry conferences such as DISTRIBUTECH, IEEE conferences, Utility Week, Grid Modernization Forum, and others. Papers may also be submitted to industry publications, such as IEEE Transactions.

# 5.0 Recommendations

## 5.1 Transition for Commercial Use

Based on the findings and results in this demonstration, phase identification and meter-to-transformer mapping are not ready for commercial use with the given constraint of two meters per transformer. While phase identification has shown promising results with this constraint, meter-to-transformer mapping has not. To achieve higher levels of accuracy for meter-to-transformer mapping, this constraint must be removed and data from as many meters as possible used in calculations.

## 5.2 Implementation Recommendation

Advancements in machine learning, advanced data mining, and artificial intelligence coupled with reduced data storage costs and improved network throughput have created numerous opportunities to use AMI data beyond the use case of meter reading and billing. The successful use case in this demonstration, analytical based phase identification, is just one example of this, but there are many more. Pursuing only this singular use case would be an inefficient use of resources when additional valued could be derived from the data collected. The key recommendation for this study is to identify additional use cases that use AMI data, and then to pursue an application or suite of applications that can fulfill them. This will require further investigation and coordination with operations personnel. At a minimum, the following high-level activities should be pursued:

*Use Case Development/Business Case*

Identify a core team of human resources (internal staff supported by consultants) with expertise in various areas of business operations. These areas may include distribution grid operations, distribution planning, AMI operations, Electric Regional Operations, and others. Conduct a brainstorming session to determine various use cases that could potentially use AMI data as its source. Sample use cases are listed in Part I, Section 6.1.2. Part and parcel to identifying the use cases ensuring that operational requirements are identified for each use case. This is where the minimum acceptable accuracy for each use case must be identified in addition to other functional and non-functional requirements. From these use cases, a business case must be developed that clearly identifies the benefits and the associated cost.

*Request for Information/Request for Proposal*

Once the use cases are identified, further investigation will be required to determine the availability of products or services that can satisfy each use case. Steps to accomplish this task include requirements gathering (beyond those that are identified for each use case); preparation and submission of the

Request for Information (RFI) or Request for Proposal (RFP); evaluation of the vendor/service provider responses; and finally, vendor selection.

Future potential use cases have a wide variety of organizations across the enterprise that will benefit from implementation. However, solving the needs of all potential users may be too large of an endeavor. The project team therefore recommends limiting use cases to organizations that perform planning functions and grid modernization functions. It is recommended that the stakeholder business units in SDG&E that served on the team for this EPIC project, define and implement an action plan to pursue the steps outlined above.