Assessment of Measurement-Based Phase Identification Methods

FRANCIS THERRIEN1, LOGAN BLAKELY2 (Member, IEEE), AND MATTHEW J. RENO2 (Senior Member, IEEE)

1 Eaton Corporation, CYME International T&D, Saint-Bruno-de-Montarville J3V 3P8, Canada
2 Sandia National Laboratories, Albuquerque, NM 87123 USA

CORRESPONDING AUTHOR: F. THERRIEN (FrancisTherrien@eaton.com)

This material is based upon work supported by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 34226. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

ABSTRACT The task of determining the phase connection of customers, known as phase identification, is crucial to obtain accurate distribution system models. This paper starts with a thorough literature review of the existing phase identification methods, which are broadly divided into three categories: hardware-based, real power-based, and voltage-based methods. This is followed by multiple case studies assessing the accuracy of six real power- and voltage-based phase identification algorithms on four realistic distribution test systems. Synthetic load profiles along with network models are used to quantify accuracy of each method for different scenarios: varying advanced metering infrastructure (AMI) coverage, number of initially mislabeled customer phases, number of available samples, and measurement noise. A case study using a real AMI data set, including field verification, is also provided. Finally, several aspects key to accurate phase identification are discussed in detail.

INDEX TERMS AMI, distribution systems, literature review, phase detection, phase identification, smart meters.

I. INTRODUCTION

UTILITY companies are under increasing pressure to have accurate models of the low-voltage distribution system. This pressure stems in part from the ever-increasing penetration of distributed energy resources (DER) such as photovoltaic (PV) installations and electric vehicle adoption inside the low-voltage networks. Inaccuracies in the distribution system models used by utilities are well documented [1]–[3]. The phase of individual customers (see Fig. 1) is one potential example of distribution system model error, and accuracy in phase labels has several critical benefits to a utility company. One example is the benefits accrued by balancing the load of the phases using a more precise network model, e.g., reducing losses and overloads [4], [5]. Another benefit is the increased accuracy of hosting capacity analyses for increasing the penetration of DER, [2], [6], [7]. The availability of advanced metering infrastructure (AMI) data streams has dramatically increased the number of viable approaches to algorithmically identify customer phases, which is known as phase identification.

Although there are a large number of different phase identification algorithms in the literature; there has not been a direct comparison of promising methods on the same data.
sets, in-depth discussions of when different methods may be appropriate, or their advantages and disadvantages under different conditions. This type of analysis is necessary for these methods to achieve widespread use, incorporation into mainstream software packages, and for utilities to have confidence in the algorithm results for their particular situation.

The primary contributions of this paper are as follows:

i) A comparison of six state-of-the-art phase identification algorithms on four different circuits with over 3,000 customers, including four voltage-based methods and two real power-based methods. The algorithms are compared under differing circuit conditions, AMI penetration, number of available samples, different quantities of measurement noise, and differing quantities of initially mislabeled phases. To our knowledge this is the first comparison of this scale and breadth, directly comparing so many algorithms with different approaches and data requirements;

ii) An additional set of results obtained from a real AMI data set, providing more insight on phase identification accuracy under realistic conditions;

iii) A comprehensive literature review of the current state-of-the-art in phase identification methods.

II. LITERATURE REVIEW

The literature on the phase identification task can be broadly divided into three categories depending on the approach. Those categories are i) hardware-based approaches, covering both signal-injection methods and micro-synchrophasor (PMU) methods; ii) real power-based methods covering load-summering methods, salient event analysis, and regression methods; and iii) voltage-based methods which are often based on correlation analysis of voltage time series either between customers or between customers and the substation or feeder head. At the time of writing, the hardware-based methods and the real power-based methods account for approximately one quarter each of the available literature, and the voltage-based methods account for approximately half of the available literature.

A. HARDWARE-BASED APPROACHES

Historically, the hardware-based approaches have been the most widely used for phase identification. In the signal injection approach, a signal is injected onto a particular phase at the substation and a device is used at each meter in the system to read that signal, [8]–[10]. This is done individually for each phase and each customer. The approach based on PMU technology uses a PMU device to measure the voltage phase angle at individual customers and this is referenced to the voltage phase angles measured at the substation, [11]–[15]. There is also an approach that leverages video analysis of light sources within buildings, [16]. The advantage of the hardware-based methods is that they are extremely accurate; the signal injection and PMU approaches are well-established and accurate. The disadvantages are the cost associated with the hardware devices themselves and the cost in both time and manpower for utility field personnel to take readings at each individual metering device. Despite the significant investment required, they may be the only option in the absence of available AMI data.

B. REAL POWER-BASED APPROACHES

A number of methods are based on using the real power AMI time series, or equivalent transformations of AMI energy readings (kWh). There are several approaches that leverage the fact that the sum of all customer loads on a phase should be equal to the power measurements at the substation, less losses. Arya et al. [17], [18] take this load summing approach and use optimization to solve for the customer phases. The approach in [19] extends that methodology to AMI data collected in the United Kingdom where the AMI data is required to be aggregated due to privacy laws. Pappu et. al [20], [21] incorporate principal component analysis (PCA) and graph theory to leverage the load summing concept to build a tree from individual customers back to the substation. Dilek et al. [22] solve the power flow equations and then use Tabu search among all possible phase assignments to find the arrangement that best matches the power flow. This approach requires all of the information necessary to solve the power flow equations. A disadvantage of the load summing approach is the degradation in results if the AMI penetration is not 100%.

Work conducted in [23] first filters out low frequencies from the customer load profiles using a Fourier transform, and then extracts ‘salient’ or distinguishing events from each customer and uses correlation coefficients with the substation load measurements to assign a phase to each customer. This is one of the methods that was implemented for analysis in this paper, and results are shown below in Section IV. Jimenez et al. [24] extend that approach, incorporating a customer processing order and including a statistical test on the correlation coefficients as a measure of statistical significance for the results. This approach does not depend on having high penetrations of AMI. Hosseini et al. [25] also propose preprocessing the AMI real power data using a high-pass filter to remove the seasonal and other low-frequency trends in the data, and then apply a modified version of k-means. One advantage of this method is that it does not depend on 100% AMI penetration.

Multivariate linear regression between customer load profiles and the measured substation load is applied in [26]. The author also includes a detailed analysis of a variety of data issues including measurement error and clock synchronization.

Tang and Milanovic [27] propose an approach using the least absolute shrinkage and selection operator (LASSO) method. The LASSO method is a regression methodology that utilizes an l1-norm penalty. The regression methods calculate the regression between individual customers and the substation. The LASSO methodology in [27] is one of the methods implemented and tested below. The primary
advantage of real power-based phase identification methods is that utilities are collecting this data—often as equivalent energy measurements—whereas they may not be collecting voltage data. Even as the penetration of AMI continues to increase, utilities rarely collect all of the available measurements from the AMI devices so the power-based methods are appealing for that reason. One disadvantage is that, in many but not all cases, less than 100% penetration of AMI meters degrades performance of the power-based algorithms.

C. VOLTAGE-BASED APPROACHES
Approximately half of the current literature on phase identification uses a voltage-based approach, although some methods also require additional data. Many of these methods leverage the fact that customers on the same phase will have voltage time series that are more correlated than that of customers on different phases. Seal and McGranaghan [28] propose an approach that finds ‘events’ that are unique to each phase within the substation supervisory control and data acquisition (SCADA) measurements and then calculate correlation coefficients between customer voltage time series and the identified events. This is similar in concept to the approaches from [23], [24] which used power data. This approach is further tested and refined in [29].

In [30], [31] reference profiles for each phase are created by averaging the time series per phase from the three-phase customers on a feeder. Then an exhaustive comparison is performed between individual customer profiles and the reference profiles.

Linear regression was also proposed between the customer voltage and substation time series in [32]. Luan et al. [1], [33] calculate a point-of-coupling (POC) voltage for each customer to remove the effects of individual customer load and then calculate correlation coefficients to do phase matching. Note that calculating POC voltage requires real power, reactive power, and information about line resistance and reactance in addition to the voltage measurements.

The work in [34]–[36] uses k-means applied to a voltage difference representation of the data. Predicted phases are assigned based on a majority vote from the original utility phase labels. PCA and must-link constraints are included in the k-means formulation in [37]. The dimensionality of the voltage profiles is first reduced using PCA and then k-means is applied, leveraging must-link constraints in the form of the labeling of the single-phase laterals. This works well, but it assumes that the labeling for the laterals is correct. If it is not completely accurate, those inaccuracies are carried through into the phase identification results. This method was implemented below with results, but the must-link constraints were not included in implementation. In [38], t-SNE dimensionality reduction and DBSCAN clustering with must-link constraints are proposed instead of PCA and k-means clustering.

Supervised machine learning approaches are proposed in [39], [40]. These approaches discuss intelligently selecting a subset of customers to credibly verify the phase connections using field inspections, and that subset is used as a training data set for the machine learning algorithms. The algorithms then predict the remaining customers on the feeder. These approaches still require some manual verification of phases in each feeder and the model requires retraining for each feeder.

Olivier et al. [41], [42] propose a constrained multi-tree (CMT) algorithm for the phase identification task. This approach creates a tree structure for each phase with the substation as the root, leveraging the radial structure of the distribution system. Correlation coefficients are calculated between the remaining customers and the existing nodes in the tree, the customer with the highest correlation coefficient is added to the tree until all customers have been added. This method was implemented, and results are shown below. The work in [43] follows a similar approach but adds a POC calculation.

In [44] the topology of the whole system and AMI measurements are used together to formulate the phase identification problem as a maximum marginal likelihood (MMLE) problem. The results in this work are promising, but the data requirements for the method are quite high.

Various formulations that include spectral clustering of the voltage time series are shown in [45]–[48]. An ensemble using spectral clustering of data windows of the voltage time series is used to create a co-association matrix in [47] which is then used as input to another spectral clustering algorithm to assign the phase predictions. In [48] an ensemble using spectral clustering uses the original utility labels within each cluster to vote on predicted phases; this method was implemented here and the results are shown below.

Looking at the literature above, it is difficult to directly analyze the efficacy of different phase identification algorithms due to differences in tested data sets, differing data requirements, etc. Out of the 26 data-based methods, one compared their proposed method to two other methods and six compared their proposed method to one other method, and the comparison methods were always within the same family, either voltage-based or power-based. The other literature did no comparison of other methods. This makes it very difficult to accurately compare algorithms, even among the same type of method, but particularly power methods versus voltage methods; to our knowledge there is no literature that directly compares voltage-based methods to power-based methods. This work addresses these issues by directly comparing six methods on the same data sets and quantifying the advantages and disadvantages of each method. Four of the methods are voltage-based methods and the other two are power-based methods. This is the first comparison of this type and scope in the literature for phase identification algorithms.

III. METHODOLOGY
A. ALGORITHMS
The algorithms listed below have all been implemented in a Python prototype that interfaces with the CYME Power Engineering Software. The first four use voltage measurements
TABLE 1. Input data requirements of the six tested methods.

| Method                        | AMI       | SCADA   | Initial customer phases |
|-------------------------------|-----------|---------|-------------------------|
| ESC-GIS [48]                  | ✓         | ✓       | ✓                       |
| ESC-SCADA                     | ✓         | ✓       | ✓                       |
| PCA [37]                      | ✓         | ✓       | ✓                       |
| MT [41]                       | ✓         | ✓       | ✓                       |
| LASSO [27]                    |           | ✓       | ✓                       |
| SF [23]                       |           | ✓       | ✓                       |

while the last two are based on power measurements. The overview of each algorithm was described above, and the implementation details are shown here. The data requirements of all six methods are also summarized in TABLE 1, where a checkmark means that data is required as an input to the algorithm.

1) ENSEMBLE SPECTRAL CLUSTERING WITH GIS PHASING (ESC-GIS) [48]
Scikit-learn’s spectral clustering function is used with a Gaussian kernel ($\gamma = 0.01$) [49]. Unless stated otherwise, 12 clusters are used in the k-means stage. Each window covers four days.

2) ENSEMBLE SPECTRAL CLUSTERING WITH SCADA MEASUREMENTS (ESC-SCADA)
This method is similar to ESC-GIS, except that the initial customer phase – used for voting – is found by computing the Pearson correlation coefficients (PCC) between the customer and the three single-phase SCADA feeder head voltage time series. The largest PCC defines the initial phase.

3) PRINCIPAL COMPONENTS ANALYSIS (PCA) [37]
The problem is solved using scikit-learn’s PCA function with two components [49]. Must-link constraints [37] are excluded to avoid dependency on a potentially erroneous network model.

4) MULTI-TREE (MT) [41]
The single-phase SCADA feeder head measurements serve as the tree’s root. PCC are used to form the distance metric.

5) LASSO [27]
Scikit-learn’s LASSO function is used with a penalty coefficient $\alpha = 0.05$ [49].

6) SALIENT FREQUENCY (SF) [23]
The first 10 components computed by the Discrete Fourier Transform are filtered out [23]. The threshold $TH$ [23] is equal to the number of AMI meters divided by 5. Only the load variations of the closest 20 samples are considered.

TABLE 2. Properties of the four distribution test systems.

| Network            | Nodes | AMI Meters | Substation Regulators | Inline Regulators |
|--------------------|-------|------------|-----------------------|-------------------|
| EPRI’s CKT5        | 3003  | 1373       | 0                     | 0                 |
| North #1           | 2369  | 615        | 1                     | 3                 |
| North #2           | 4065  | 963        | 1                     | 6                 |
| South              | 1778  | 447        | 0                     | 1                 |

B. TEST SYSTEMS
Four existing, large-scale North American radial distribution systems are considered in this paper; some of their features are presented in TABLE 2. All substation and inline voltage regulators have individual phase control. Only three-phase four-wire systems with customers connected between line and neutral are considered, since many power-based methods are not designed to handle customers connected line to line (some voltage-based methods have been shown to support both line-to-line and line-to-neutral customers, e.g., [37], [38]). Moreover, only the phases of single-phase customers are identified.

To compare the accuracy of the phase identification methods, the network model of each system is considered as the ground truth. Base sets of equivalent AMI and SCADA measurements are generated by solving time-series power flows on each network model for 10,000-timesteps at a 15-minute interval (around 104 days) with synthetic AMI active and reactive power profiles [50] as inputs. In order to see the impact of the voltage regulation equipment on the algorithms, other sets of equivalent measurements are generated for the North #1, North #2, and South systems where the regulator taps are fixed during the entire simulation. They are denoted by the suffix (Fixed).

C. METRICS
The accuracy of the algorithms is defined as

$$\text{Acc}_k = \frac{1}{3} \sum_{\sigma \in \{a,b,c\}} \sum_{n \in M^\text{ref}_\sigma} \frac{F_\text{eq}(P_n^\text{pred}, \sigma)}{|M^\text{ref}_\sigma|} \times 100\%$$

where $M^\text{ref}_\sigma$ is the set of customers with valid measurements originally connected to phase $\sigma$ (in the network model), $P_n^\text{pred}$ is the predicted (identified) phase of the $n$th customer, $F_\text{eq}(x, y)$ returns 1 if $x = y$ and 0 otherwise, and $|S|$ represents the cardinality (number of elements) of set $S$.

For ensemble methods (ESC-GIS and ESC-SCADA), a confidence score $CS_n$ can be computed to get a feel of the prediction accuracy. For a given customer, $CS_n$ represents the percentage of windows whose predicted phase equals the final predicted phase (see Section IV-C of [51] for more detail).

IV. CASE STUDIES–SYNTHETIC DATA SET
This section presents several case studies using the phase identification methods and networks described in Section III.
A. IDEAL CONDITIONS
The accuracy metric $\text{Acc}_\%$ is first computed for all methods under ideal conditions (e.g., noiseless measurements, full AMI coverage, ...); the corresponding results are presented in Fig. 2.

ESC-GIS provides almost perfect results (>99%) for all systems. The accuracy of ESC-SCADA and PCA is similar: both are very accurate with no active tap changers, but their accuracy plunges for networks with multiple regulators (<60% for North #1 and #2). MT behaves similarly to ESC-SCADA and PCA but is more accurate for North #1 (84.6%). LASSO’s accuracy is greater than 90% for all networks except CKT5 (67.9%). The accuracy of SF ranges from 60 to 80%.

The main reason why ESC-SCADA and PCA are inaccurate for North #1 and #2 is because inline voltage regulators cause occasional but significant voltage changes that are not seen by the SCADA voltage measurements. Consequently, the step in these methods that tries to correlate the AMI and SCADA voltage measurements often fails. As for ESC-GIS, as long as an adequate number of clusters is used (see Section VI-A), the presence of voltage regulators has a negligible impact since it does not use SCADA measurements.

MT almost always correctly identifies large groups of nearby same-phase customers independently of the presence of tap-changing devices. However, it often has difficulty connecting these groups to the proper tree when tap changers are present. An extreme case is North #2, where 957 of the 963 customers were assigned to phase A while only 397 are truly on this phase.

For CKT5, the LASSO algorithm fails to identify any phase for several customers (the corresponding element in the three solution vectors [27] all equal 0), which explains its very low accuracy. Nevertheless, if these unidentified customers are removed from $M_\sigma^x$, $\text{Acc}_\%$ only jumps from 67.9% to 84.4%. Since CKT5 is the network with the most customers, followed by North #2 where several customers also remain unidentified, it appears that LASSO has difficulties with larger systems.

B. AMI COVERAGE
A well-known practical concern with power-based methods is low AMI coverage, e.g., due to partial AMI deployment or opt-out clauses. To assess the robustness of LASSO to different AMI coverage levels, the case study is repeated by removing different numbers of AMI meters from the base set. The location of these meters is chosen randomly. To compare with voltage-based methods, the same study is executed with ESC-GIS. The corresponding results are presented in Fig. 3.

LASSO’s accuracy decreases as the AMI coverage is reduced. The accuracy remains above 80% for the South network even with a coverage of only 50%, which shows some robustness. However, for North #2, the accuracy falls below 80% and 90% for coverage levels of 90% and 95%, respectively. This limits the applicability of LASSO in practical situations. Note that as in Fig. 2, the presence of tap changers has a limited impact on LASSO’s accuracy. As for ESC-GIS, the coverage level has no effect on accuracy.

C. INITIAL PHASE MISLABELING
ESC-GIS assigns a unique phase to all customers of a given cluster using a voting scheme based on the initial guess of each customer phase [48]. This initial customer phase typically comes from a GIS system. In the ideal conditions test case (Section IV-A), all initial phases were exact, which is an unrealistic assumption. Note that among the methods tested in this paper, only ESC-GIS uses this initial phasing.

The study of Section IV-A is repeated by intentionally mislabeling different numbers of initial customer phases. The method remains very accurate with up to half of the phases incorrectly labeled (>95% for all networks), showing high robustness. The accuracy decreases drastically afterwards. This perfectly aligns with the results presented in [51].
D. NUMBER OF SAMPLES

Fig. 4 presents the impact of the number of measurement samples on Acc on the CKT5 and South networks. For the four voltage-based methods, 1,000 samples (around 10 days of data at a 15-minute interval) is sufficient to achieve maximum accuracy under ideal conditions on all test networks. The same is true with LASSO on South (the smallest network of the set) but not CKT5 (the largest). With CKT5, the accuracy keeps increasing as a function of the sample count, although it starts plateauing between 5,000 and 10,000 samples. Finally, as mentioned in Section IV-A, Acc monotonically increases as a function of the sample count with SF. At least 500 samples were needed for the ensemble methods to have at least one full window of four days.

E. MEASUREMENT ERROR

All previous studies assumed ideal, noiseless, and synchronized measurements. To analyze some of the impact of imperfect measurements, the study of Section IV-A is repeated by adding different levels of uniformly distributed noise (as a percentage of each measured value) to all AMI and SCADA measurements; Fig. 5 presents the corresponding results for North #1 and North #1 (Fixed). All methods see virtually no change with an error level of 0.1%, while Acc starts decreasing with a noise level of 0.2% for some of the voltage-based methods. Due to the inherent randomness, there is no clear trend. For instance, ESC-GIS has better accuracy with noises of 0.4% and 0.5% than 0.2% and 0.3%. LASSO and SF appear immune to these levels of noise, since power measurements see natural variations of much larger amplitude than the injected noise.

V. CASE STUDY WITH UTILITY DATA SET

Section IV used synthetic data sets with network models as ground truths, enabling to quantify phase identification accuracy under controlled conditions. It is nevertheless difficult to create a completely realistic scenario solely using synthetic data and network models.

In this section, an AMI voltage data set is provided by a utility, and ESC-GIS is applied to predict the actual customer phases compared to the utility’s distribution system model phases. ESC-GIS is used since SCADA measurements are unavailable and because it provided the most accurate results in Section IV. The set spans over five months with a 15-minute time interval and covers a single feeder with 290 metered single-phase customers. The feeder has one substation and two inline voltage regulators. There are several instances of missing data through parts the five months. Out of 11,310 potential results (39 windows times 290 customers), 2,615 could not be computed [48]. Six clusters are used in the k-means step.

The results are summarized in TABLE 3. All customers initially labeled on phase A are predicted to be on this phase; while no additional customers are assigned on phase A. Out of 107 customers on phase B according to the original network model, ESC-GIS predicted that 29 of them should be on phase C. The confidence score of each of these 29 customers is very high (CS > 89%), with most being in the 96-97% range. According to the network model, the 29 customers belong to the same lateral. After viewing these results, the utility realized that they had planned to move this lateral to phase B and changed the model accordingly but had forgotten to do it in the field. ESC-GIS therefore correctly identified an entire mislabeled lateral. The phase identification algorithm also predicted that six customers initially connected to phase C should be on phase B, with relatively low CS ranging from 72.2% to 75.7%. According to the network model, all six customers are directly tapped to three-phase lines in the same part of the feeder. The utility did not confirm nor infirm this prediction.

In the spectral clustering algorithm, the final k-means stage – which finds clusters of customers on the same phase – uses the eigenvectors of a Laplacian as the input [48]. Each eigenvector is a representation of a single customer. Visualizing them is another means of assessing the prediction quality.
The first three non-trivial dimensions of the eigenvectors of the utility data set are plotted in Fig. 6 for a single window with few missing measurements. Most customers are neatly grouped in three clusters, implying a strong correlation between the voltage time series of customers connected on the same phase, thus giving more confidence in the predicted results. The six customers initially labeled on phase C discussed in the previous paragraph belong to a compact cluster mostly comprised of phase B meters (bottom left of Fig. 6), strengthening the prediction. The 29 customers belonging to the mislabeled lateral are also grouped compactly with the majority of customers initially assigned on phase C (bottom right).

VI. DISCUSSION

There are several topics that merit further discussion arising from the results in the case studies shown in Sections IV and V. These topics are the impact of voltage regulators, the impact of the number of clusters used in the ESC-GIS and ESC-SCADA algorithms, the impact of missing measurements, the impact of measurement distortion, and the impact of the number of samples available.

A. VOLTAGE REGULATORS

As outlined in Section V, voltage-based phase identification methods are very sensitive to the presence of multiple voltage regulators, especially when both AMI and SCADA measurements are used. The reason is intuitive. Many voltage-based methods are derived on the assumption that the voltage variations are more similar between customers on the same phase than on different phases. Since voltage regulators cause large variations that are only seen by downstream customers, the voltage correlation between customers on the same phase but on other sides of the tap changer is weakened.

To visualize this explanation, the first three non-trivial dimensions of the eigenvectors obtained by ESC-GIS on North #2 and North #2 (Fixed) under ideal conditions are shown in Fig. 7 and Fig. 8, respectively.

Due to the voltage regulators, many more clusters can be observed in Fig. 7 than in Fig. 8; nevertheless, same-phase customers are mostly grouped together, enabling proper delimitation of clusters. While clusters of the same phase are mostly closer to each other, one group of customers on phase A is closer to clusters of phases B and C. This makes it challenging to correctly map each cluster with the corresponding SCADA phase voltage measurement. Finally, as observed in Section IV, voltage regulators do not impact power-based methods since they have a negligible impact on load power consumption. It is therefore suggested to either use power-based methods (e.g., LASSO and SF) or voltage-based methods that do not rely on SCADA measurements (e.g., ESC-GIS) for networks with multiple voltage regulators.

B. NUMBER OF CLUSTERS

It could appear that using one cluster per phase in the k-means stage of ESC-GIS and ESC-SCADA is sufficient for networks with all customers connected between line and neutral. How-
ever, as seen in Fig. 7, all customers connected to the same phase are not necessarily grouped in one cluster even under ideal conditions. This is particularly manifest for feeders with multiple voltage regulators. Consequently, more clusters are used with ESC-GIS and ESC-SCADA than the number of phases (i.e., 12 in Section IV and six in Section V).

To further substantiate this claim, the study of Section IV-A is repeated with ESC-GIS using only three clusters, i.e., one per phase. The accuracy of North #1 and North #2 is only 68.9% and 66.4%, respectively; whereas Acc% > 99% for all five other networks. It is recalled that accuracies greater than 99% were obtained for North #1 and North #2 using 12 clusters. On the other hand, using too many clusters may also deteriorate the accuracy, especially for networks with fewer customers. As an example, the study of Section V is repeated with 12 clusters instead of six. Only six of the 29 customers belonging to the mislabeled lateral discussed in Section V are correctly identified when using 12 clusters; whereas all are identified with only six clusters. The creation of smaller clusters within the ESC-GIS algorithm may therefore also prevent the detection of multiple highly correlated mislabeled customers, e.g., belonging to the same lateral.

C. MISSING MEASUREMENTS

Section IV-B showed and discussed the impact of AMI coverage, where loads were either fully metered or unmetered. In practice, in addition to customers with no AMI metering, AMI voltage and/or power time series will often comprise single or multiple missing measurements, e.g., due to communication failures and system maintenance.

Many phase identification methods require all customers being identified to have a full measurement set for the entire period of study. For methods requiring large data sets to achieve accurate predictions, it can be difficult to define a long enough study period where all meters have no missing measurements.

Ensemble methods such as ESC-GIS and ESC-SCADA therefore have a practical advantage. In these methods, multiple smaller windows (e.g., 4 days) are studied individually [48]. If one or multiple measurements are missing, the customer is only ignored for the corresponding window(s). The final voting stage is still applied, except with fewer windows for customers with missing measurements [48]. However, caution must be taken when including windows with several excluded customers, as they are more prone to yield inaccurate predictions. This may unintentionally decrease the confidence scores.

D. MEASUREMENT DISTORTION

The study of Section IV-E considered the impact of measurement noise on accuracy by adding different levels of uniformly distributed noise to otherwise ideal and synchronized measurement sets. While informative, this does not represent the full spectrum of measurement distortions that can affect practical phase identification studies. For instance, the set of SCADA and/or AMI measurements used at a given time step may not be synchronized. These measurements may represent instantaneous values at the end of the interval, peak or average values for the entire interval, etc. Moreover, many SCADA meters provide power measurements while AMI systems often send energy measurements, requiring approximations to convert to equivalent power measurements.

E. SAMPLE COUNTS

Section IV-D showed that unlike with SF, a relatively small number of samples (e.g., 1,000) is needed for voltage-based methods to reach their peak accuracy. This is to be expected since SF identifies all customers individually (through their salient variations) as opposed to grouping them. It therefore requires a considerable number of samples to find sufficient high-frequency components (salient events) for each AMI meter to be associated with the corresponding SCADA meter.

While several methods can handle years of measurements with little computational cost, using too much data comes with practical concerns. For one, customer phases are not always static. An example thereof is provided in Section V. Another issue is algorithms that can only use complete data sets (see Section VI-C). Longer periods increase the possibility of having to remove customers due to missing measurements.

VII. CONCLUSION

This work tested six state-of-the-art phase identification algorithms using AMI data. Four methods were based on voltage time series: ensemble spectral clustering with GIS phasing (ESC-GIS) [48], ensemble spectral clustering with SCADA measurements (ESC-SCADA), principal component analysis (PCA) [37], and multi-tree algorithm (MT) [41]. Two were power-based methods, LASSO [27] and salient frequency (SF) [23]. The ESC-GIS method performed the best in all of testing conducted in this work, showing robustness to differing feeder configurations and data collection concerns. ESC-GIS requires AMI voltage time series and existing utility phase labels where more than 50% are believed to be accurate. ESC-SCADA removes the requirement for existing phase labels but adds a requirement for SCADA data at the substation, and in that case, there is a decrease in performance. If utility labels and SCADA measurements are unavailable, ensemble spectral clustering can be conducted and the final phase left to manual verification [48]; this would require only AMI voltage time series but adds a small manual verification step at the end of the method. If voltage AMI data is not available, the LASSO method performed the best out of the two power methods; it requires real power time series and real power SCADA measurements at the substation. If AMI voltage and either real power AMI or real power SCADA data is not available, then traditional phase identification methods must be used, such as manual verification or hardware-based methods. The ensemble spectral clustering methods (if voltage AMI data is available) and the LASSO method (if only real power AMI data is available) are both shown to be good
choices for the distribution system phase identification task under a variety of conditions.

VIII. ACKNOWLEDGMENT

This article describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapunuk, “Smart meter data analytics for distribution network connectivity verification,” IEEE Trans. Smart Grid, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.
[2] B. Palmintier et al., “On the path to SunShot: Emerging issues and challenges in integrating solar with the distribution system,” Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/TP-5D00-65331, 2016.
[3] L. Blakely, M. J. Reno, and J. Peppanen, “Identifying common errors in distribution system models,” in Proc. IEEE 46th Photovoltaic Specialists Conf. (PVSC), Chicago, IL, USA, Jun. 2019, pp. 3132–3139.
[4] J. Zhu, M.-Y. Chow, and F. Zhang, “Phase balancing using mixed-integer programming,” IEEE Trans. Power Syst., vol. 13, no. 4, pp. 1487–1492, Nov. 1998.
[5] R. Yan and T. K. Saha, “Voltage variation sensitivity analysis for unbalanced distribution networks due to photovoltaic power fluctuations,” IEEE Trans. Power Syst., vol. 27, no. 2, pp. 1078–1089, May 2012.
[6] M. J. Reno, K. Coogan, S. Grijalva, R. J. Broderick, and J. E. Quiroz, “PV interconnection risk analysis through distribution system impact signatures and feeder zones,” in Proc. IEEE PES Gen. Meeting Conf. Expo., Harbor, MD, USA, Jul. 2014, pp. 1–5.
[7] V. Schwarzer and R. Ghorbani, “Transient over-voltage mitigation: Explanation and mitigation options for inverter-based distributed generation projects,” Elect. Vehicle Transp. Center, Univ. Hawaii Manoa, Honolulu, HI, USA, Tech. Rep. HNEI-02-15, Feb. 2014.
[8] P. Kulkarni and A. R. Kolwalkar, “Phase identification system and method,” EP Patent 2 562 554 A3, Jun. 25, 2014.
[9] Z. Shen, M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, “Three-phase AC system impedance measurement unit (IMU) using chirp signal injection,” in Proc. 28th Ann. IEEE Appl. Power Electron. Conf. Expo. (APEC), Long Beach, CA, USA, MAR. 2013, pp. 2666–2673.
[10] K. J. Caird, “Meter phase identification,” US Patent 8 143 879, Mar. 27, 2012.
[11] S. V. Achanta, “Automatic phase identification for electric power delivery lines,” U.S. Patent 2019 0 383 869 A1, Sep. 15, 2020.
[12] M. H. F. Wen, R. Arghandeh, A. von Meier, K. Poolla, and M. Belkhayat, “Three-phase AC system impedance measurement unit (IMU) using chirp signal injection,” in Proc. 28th Ann. IEEE Appl. Power Electron. Conf. Expo. (APEC), Long Beach, CA, USA, MAR. 2013, pp. 2666–2673.
[13] Z. Shen, M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, “Three-phase AC system impedance measurement unit (IMU) using chirp signal injection,” in Proc. 28th Ann. IEEE Appl. Power Electron. Conf. Expo. (APEC), Long Beach, CA, USA, MAR. 2013, pp. 2666–2673.
[14] J. L. T. Santos, “Phase identification in low voltage smart grids using smart meter data,” Ph.D. dissertation, Dept. Elect. Comput. Eng., Técnico Lisboa, Lisbon, Portugal, 2019.
[15] X. Tang and J. V. Milanovic, “Phase identification of LV distribution network with smart meter data,” in Proc. IEEE Energy Power Eng. Soc. Gen. Meeting (PESGM), Portland, OR, USA, Aug. 2018, pp. 1–5.
[16] B. K. Seal and M. F. McGranaghan, “Automatic identification of service phase for electric utility customers,” in Proc. IEEE Power Energy Soc. Gen. Meeting, Detroit, MI, USA, Jul. 2011, pp. 1–7.
[17] V. Vycital, M. Pateck, P. Toman, D. Topolanek, J. Drapela, and J. Zamhiriopoulos, “Phase identification in smart metering pilot project Komorany,” in Proc. Conf. Electr. Distrib. (CRED), Madrid, Spain, Jun. 2019, pp. 1–5.
[18] H. Pezeshki and P. J. Wolfs, “Consumer phase identification in a three phase unbalanced LV distribution network,” in Proc. 3rd IEEE PES Innov. Smart Grid Technol. Eur. (ISGT Europe), Berlin, Germany, Oct. 2012, pp. 1–7.
[19] H. Pezeshki and P. J. Wolfs, “Correlation based method for phase identification in a three phase LV distribution network,” in Proc. Australas. Univ. Power Eng. Conf. (AUPEC), Bali, Indonesia, 2012, pp. 1–7.
[20] T. A. Short, “Advanced metering for phase identification, transformer identification, and secondary modeling,” IEEE Trans. Smart Grid, vol. 4, no. 2, pp. 651–658, Jun. 2013.
[21] W. Luan, J. Peng, M. Maras, and J. Lo, “Distribution network topology error correction using smart meter data analytics,” in Proc. IEEE Power Energy Soc. Gen. Meeting, Vancouver, BC, Canada, Jul. 2013, pp. 1–5.
[22] R. Mirra et al., “Voltage correlations in smart meter data,” in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Sydney, NSW, Australia, Aug. 2015, pp. 1999–2008.
[23] V. Arya et al., “Voltage analytics to infer customer phase,” in Proc. IEEE PES Innov. Smart Grid Technol. Conf. (ISGT), Istanbul, Turkey, Oct. 2014, pp. 1–6.
[24] V. Arya and R. Mitra, “Voltage-based clustering to identify to identify connectivity relationships in distribution networks,” in Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm), Vancouver, BC, USA, Oct. 2013, pp. 7–12.
[25] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, “Phase identification in electric power distribution systems by clustering of smart meter data,” in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Anaheim, CA, USA, Dec. 2016, pp. 259–265.
[26] W. Wang, N. Yu, and Z. Lu, “Advanced metering infrastructure data driven phase identification in smart grid,” in Proc. 2nd Int. Conf. Green Commun. Comput. Technol. (GREEN), Rome, Italy, 2017, pp. 1–32.
[27] B. Foggo and N. Yu, “A comprehensive evaluation of supervised machine learning for the phase identification problem,” WASET Int. J. Comp. Syst. Eng., vol. 12, no. 6, pp. 419–427, 2018.
[28] B. Foggo and N. Yu, “Improving supervised phase identification through the theory of information losses,” IEEE Trans. Smart Grid, vol. 11, no. 3, pp. 2337–2346, May 2020.
[29] F. Olivier, A. Suteru, P. Geurts, R. Fonteneau, and D. Ernst, “Phase identification of smart meter data by clustering voltage measurement data,” in Proc. Power Syst. Comput. Conf. (PSCC), Dublin, Ireland, Jun. 2018, pp. 1–8.
[30] F. Olivier, D. Ernst, and R. Fonteneau, “Automatic phase identification of smart meter measurement data,” in Proc. Conf. Electr. Distrib. (CRED), Dublin, Ireland, Jun. 2017, pp. 1579–1583.
[31] Y. Liao, Y. Weng, G. Liu, Z. Zhao, C. Tan, and R. Rajagopalan, “Unbalanced multi-phase distribution network topology measurement and bus phase identification,” IET Smart Grid, vol. 2, no. 4, pp. 557–570, Dec. 2019.
[32] W. Wang and N. Yu, “Maximum marginal likelihood estimation of phase connections in power distribution systems,” IEEE Trans. Power Syst., vol. 35, no. 5, pp. 3906–3917, Sep. 2020.
[33] Y. Ma et al., “Phase identification of smart meters by spectral clustering,” in Proc. 2nd IEEE Conf. Energy Internet Energy Syst. Integ. (EI), Beijing, China, Oct. 2018, pp. 1–5.
[46] S. Liu et al., “Practical method for mitigating three-phase unbalance based on data-driven user phase identification,” IEEE Trans. Power Syst., vol. 35, no. 2, pp. 1653–1656, Mar. 2020.

[47] L. Blakely and M. J. Reno, “Phase identification using co-association matrix ensemble clustering,” IET Smart Grid, vol. 3, no. 4, pp. 490–499, 2020.

[48] L. Blakely, M. J. Reno, and W.-C. Feng, “Spectral clustering for customer phase identification using AMI voltage timeseries,” in Proc. IEEE Power Energy Conf. Illinois (PECI), Champaign, IL, USA, Feb. 2019, pp. 1–7.

[49] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.

[50] K. Ashok, M. J. Reno, L. Blakely, and D. Divan, “Systematic study of data requirements and AMI capabilities for smart meter analytics,” in Proc. IEEE 7th Int. Conf. Smart Energy Grid Eng. (SEGE), Oshawa, ON, Canada, Aug. 2019, pp. 53–58.

[51] L. Blakely, M. J. Reno, and K. Ashok, “AMI data quality and collection method considerations for improving the accuracy of distribution models,” in Proc. IEEE 46th Photovoltaic Specialists Conf. (PVSC), Chicago, IL, USA, Jun. 2019, pp. 1–8.

FRANCIS THERRIEN received the Ph.D. degree in electrical and computer engineering from The University of British Columbia in 2015. Since 2015, he has been working as a Lead Power Systems Engineer with Eaton. He contributes to the development of several computational engines and analyses that are part of the CYME Power Engineering Software. His research interests include modeling and simulation of electrical distribution systems, electrical machines, electromagnetic transients, and numerical analysis.

LOGAN BLAKELY (Member, IEEE) is a member of Technical Staff with the Electric Power Research Department, Sandia National Laboratories. His research focuses on machine learning applied to power systems challenges, particularly in the intersection of merging physics domain knowledge with machine learning techniques. This research area combines advances in machine learning and data analytics, the recent available of large quantities of new types of data, along with the reliability and interpretability of physics-based approaches.

MATTHEW J. RENO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Georgia Institute of Technology. He is a Principal Member of Technical Staff with the Electric Power Systems Research Department, Sandia National Laboratories. He leads several projects that are applying cutting edge machine learning algorithms to power system problems. His research focuses on distribution systems modeling and analysis with Big Data and high penetrations of PV.