SAND2019-6614C

AMI Data Quality and Collection Method Considerations for Improving the Accuracy of Distribution Models

Logan Blakely[1], Matthew J. Reno[1], Kavya Ashok[2]

[1] Sandia National Laboratories, Albuquerque, New Mexico, 87185, USA
[2] Georgia Institute of Technology, Atlanta, Georgia, 30332, USA

***Abstract* — Spectral clustering is applied to the problem of phase identification of electric customers to investigate the data needs (resolution and accuracy) of advanced metering infrastructure (AMI). More accurate models are required to accurately interconnect high penetrations of PV/DER and for optimal electric grid operations. This paper demonstrates the effects of different data collection implementations and common errors in AMI datasets on the phase identification task. This includes measurement intervals, data resolution, collection periods, time synchronization issues, noisy measurements, biased meters, and mislabeled phases. High quality AMI data is a critical consideration to model correction and accurate hosting capacity analyses.**

***Index Terms* – AMI, AMI recommendations, distribution system errors, distribution system models**

I. INTRODUCTION

The availability of advanced metering infrastructure (AMI) data provides an opportunity to analyze the distribution system at a level that was previously impossible and to accelerate the installation of high penetrations of residential PV systems. One of the many challenges facing the installation of high-penetration of residential PV systems and other distributed energy resources (DER) is the necessity for rapid and accurate hosting capacity analysis simulations [1]. AMI data has the potential to aid in determining the placement of solar installations, as well as correcting errors present in existing utility models, [2]–[6]. Accurate models are necessary for accurate hosting capacity analyses which are critical to interconnect high penetrations of DER, and research using AMI data is showing promising results in correcting common types of errors. However, there are many open questions regarding what type of AMI data should be collected to best facilitate the new data science techniques. Each utility implements their own version of AMI data collection with different collection intervals, meter precision, etc. See [7] for an overview of AMI and smart meter deployment.

This research attempts to answer some of these questions about what type of AMI data is required for model correction tasks using a synthetic dataset generated to test common AMI configurations as well as common errors present in AMI data. The task of phase identification is used to evaluate the effect of possible AMI configurations and how those configurations may interact with common dataset errors. This paper provides an overview of these issues and concludes with recommendations for AMI data collection based on this application.

II. RELATED WORK

AMI meters are quickly becoming the standard in the United States as well as worldwide. There were more than 85 million smart meters deployed in the U.S. at the end of 2018, which is ~60% of households, and 95 million are projected to be deployed by the end of 2020 [8]. Worldwide adoption is rising as well; [9] provides an overview of smart metering adoption trends around the globe. Although smart meters are rapidly becoming ubiquitous, there are few guidelines or best practices in place regarding the specific data collection techniques or quality considerations.

The AMI reporting interval is one of the few areas of AMI data collection that does have some analysis in the literature. In 2013, the National Renewable Energy Laboratory (NREL) recommended collecting AMI data at 15-minute intervals but noted that it often gets down-sampled to 1-hr intervals. In 2015, the U.S. Department of Energy (DOE) released a report entitled Metering Best Practices: A Guide to Achieving Utility Resource Efficiency in which they recommend different collections intervals based on the use case, from monthly to 15-minutes or shorter [10].

The Smart Grid Investment Grant Program (SGIG) summary from 2016 [11] provides some insight into the diversity of measurement intervals in place in the U.S. The project collected data on the AMI collection intervals from 70 of its projects with 16.3 million smart meters in use, ~14.5 million residential meters, ~1.7 million commercial meters, and ~50 thousand industrial meters. Figure 1 breaks down the measurement intervals for those customers. This subset of meters shows that most residential smart meters are collecting at 1-hr intervals, followed by 15-minute intervals, and most commercial and industrial customers are collecting at 15-minute intervals.

The recommendations for measurement interval have wide ranges, and [10]–[12] and the current literature suggestions primarily approach these recommendations from a cost-savings analysis perspective rather than the perspective of using the AMI data to validate and correct distribution system models. There is little research into what types of data analysis can be done at varying levels of AMI quality and different data collection techniques.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Looking at meter accuracy or precision, the American National Standards Institute (ANSI) defines Accuracy Classes 0.1, 0.2, and 0.5 for meters, defining the maximum allowable percentage of measurement error for each meter type [13]. The Electric Power Research Institution (EPRI) did a study on meter accuracy [14], and customer concerns drove a third-party assessment of meter accuracy in Texas [15]. These standards and studies define limits on the measurement error, but they do not quantify how that affects data analysis on the AMI data generated by those meters.

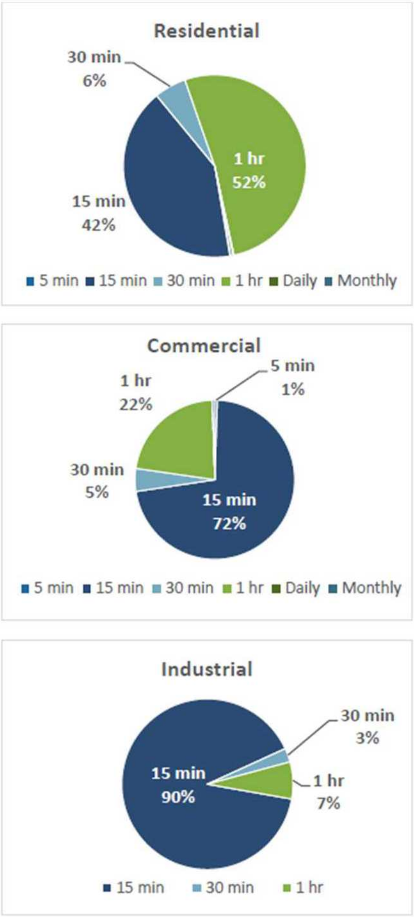


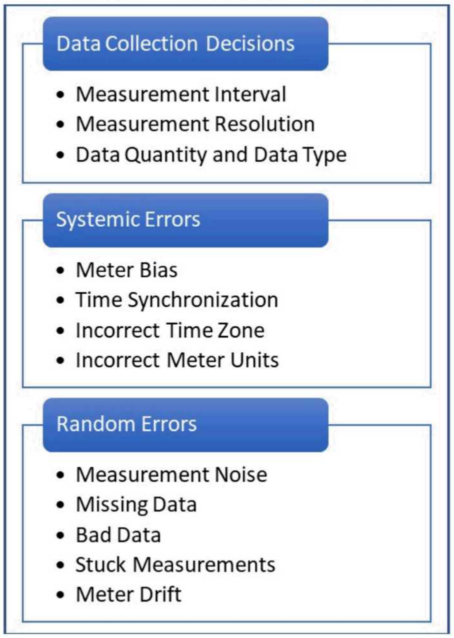
Figure 1 - Measurement intervals for the SGIG program, [11]

[16] used both 15-minute interval AMI data and 1-hr interval AMI data for a phase identification task and concluded that the 15-minute interval data produced better results on the phase identification task. [5], [17], [18] use 5-minute, 15-minute, and 1-hr AMI data respectively for phase identification. Both [16] and [18] speculate that shorter intervals may be needed to deal with seasonal variation in AMI data. However, it is difficult to directly compare these methods and results due to high variability in location, data quantities, seasons, availability of substation voltages and a variety of other factors. The discussion of AMI data collection techniques and quality has been a side discussion of the available data, not the purpose of the work.

There has been little rigorous analysis on the effects of AMI data quality issues on the ability to use the AMI data to validate and correct existing distribution system models, and that is the main contribution of this research.

Table 1 shows a, non-exhaustive, list of AMI data concerns divided into three categories. This list is compiled from literature referenced here, general measurement error types, and issues observed in AMI data. The first category consists of data collection concerns that are not errors, but simply collection decisions to be made at the time of installation of the AMI metering system. The second category is systemic errors, which are errors that are consistent over repeated measurements. The third category is random errors which are not necessarily consistent over repeated measurements. For further treatment of types of measurement error see [19], [20]. A subset of these data concerns is explored in further detail in the Results section below.

TABLE 1 - LIST OF AMI CONCERNS



Data Collection Decisions

- Measurement Interval
- Measurement Resolution
- Data Quantity and Data Type

Systemic Errors

- Meter Bias
- Time Synchronization
- Incorrect Time Zone
- Incorrect Meter Units

Random Errors

- Measurement Noise
- Missing Data
- Bad Data
- Stuck Measurements
- Meter Drift

III. TEST SYSTEM AND DATA

A year-long synthetic AMI dataset was created by running an OpenDSS [21] simulation of EPRI's Test circuit [22]. It is a 12.47 kV network, single feeder with 1379 Residential loads and 584 transformers. Average real power (kW) data at 1-minute intervals was extracted from Pecan Street [23] to create 12-month period load profiles for 1379 customers. Reactive power (kVAr) was created by assuming a uniform distribution of power factors between 0.79-0.99, that varied every 30 minutes. This range was chosen by analyzing real 15-minute P & Q load data provided by a utility. The experiments and results shown in this research require only the voltage time series recorded from the simulation.

IV. METHODOLOGY

A set of AMI data quality manipulations was chosen that includes both data alterations (e.g. differing measurement intervals) and error injections (e.g. measurement noise). The data quality manipulations that were tested were measurement interval, meter precision, biased meters, measurement noise, time synchronization, missing data, available data, and mislabeled phases. These are described in more detail in the Results section and in Table 3 Each experiment follows the three steps shown in Table 2. First, a data manipulation or a set of manipulations is chosen for an experiment and those manipulations are performed on the dataset. For example, the dataset is first averaged to 15-minute intervals, then a maximum of 0.15% noise is added to 50% of the meters. Second, the phase identification algorithm from [24] is run using the altered version of the dataset. Third, the results are analyzed to determine, in this example, what the effects of using 15-minute interval data with that level of noise has on the phase identification task.

TABLE 2 - EXPERIMENTAL PROCESS

1. Alter dataset and/or inject errors
2. Run phase identification algorithm
3. Analyze the effects of the data alterations and/or errors on the phase identification task

A. Standardized Data Processing

There are two steps of data processing on the AMI voltage time series data prior to using it as input to the phase identification algorithm. First, the data is normalized to a mean of one. Then the time series is transformed into a voltage fluctuations representation, by simply taking the difference of adjacent measurements. This was proposed in [5] and its efficacy was also demonstrated in our prior research on the phase identification algorithm [24].

B. Phase Identification Algorithm

The phase identification algorithm that is used here to test the alterations and error injections on the synthetic dataset was proposed in our prior work [24]. Figure 2 shows a summary of the algorithm. A 'window' of 4 days is selected from the available data, customers with missing data points are removed from the window, the remaining customers are clustered using the spectral clustering methodology, and each customer is assigned a predicted phase based on the majority vote of its resulting cluster. This process is repeated in subsequent windows until all available data has been used. Using the window approach leverages the power of ensemble machine learning, gives a way to deal with missing data, and allows the algorithm to be more scalable, as the entire dataset is not in use at once. On the dataset referenced in Section III and in the presence of unaltered data and no mislabeled phases, this algorithm was 100% accurate on the phase identification task.

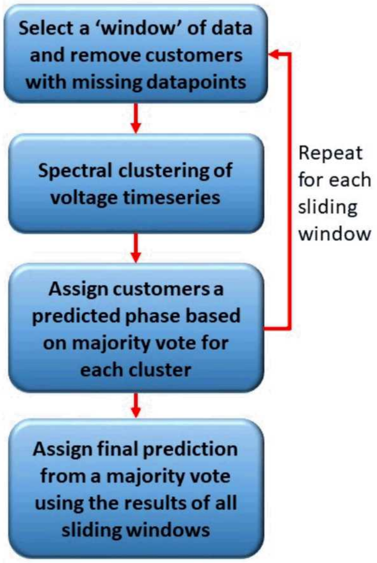

Select a 'window' of data and remove customers with missing datapoints
Spectral clustering of voltage timeseries
Assign customers a predicted phase based on majority vote for each cluster
Assign final prediction from a majority vote using the results of all sliding windows
Repeat for each sliding window

Figure 2 - Phase identification algorithm flowchart [24]

C. Confidence Score Analysis

A simple accuracy metric, the number of customers where the phases is correctly predicted over the total customers, does not give a realistic picture of the algorithm performance. We have also chosen to assign a confidence score to each prediction of the algorithm to give further insight into the performance of the phase identification algorithm in the presence of differing data formats or errors. The confidence score metric leverages the information present due to the ensemble nature of the methodology. The confidence score is defined as the number of 'winning' votes divided by the total number of votes. For example, say there are 20 windows of data, two were not used because of missing data, in 15 windows the phase prediction was phase A, and in 3 windows the prediction was phase B. In this case, the confidence score would be 15/18 or ~0.83. This can be interpreted as 83% of the total number of predictions were in agreement that this is a phase A customer. Note that the confidence score does not necessarily indicate a correct or incorrect prediction, simply the confidence of the algorithm in the final prediction. Figure 3 shows the confidence scores for the dataset that has no errors injected, all customer phase labels are correct, and the dataset is at the highest resolution. There remain a few customers with relatively lower confidence scores, and research is ongoing to determine why some customer have lower confidence even in the presence of perfect data. See Table 5 and Figure 4 for an example of an instance of the confidence score metric providing significantly more information than a percentage of correctly predicted customers.

TABLE 3 - AMI DATA QUALITY MANIPULATIONS FORMULAS

Variable Definitions	$T_{total} = the\ total\ number\ of\ measurements\ available\ at\ the\ 1 - minute\ resolution$
	$t\ \in \{1, 2, 3, \ldots, T_{total}\} - individual\ time\ step\ at\ 1 - \min resolution$
	$i\ \in I\ where\ I = \left\{1, 2, 3, \ldots, \frac{T_{total}}{k}\right\}$ $-individual\ measurement$
	$v\ \in\ V, where\ V = time\ series\ of\ voltage\ measurements$
	$C_{Total} = The\ total\ number\ of\ customers$
	$c\ \in C, where\ C = \{\ 1,2,3, \ldots, C_{Total}\}$ $-set\ of\ all\ customers$
	$\mu = ideal\ mean\ of\ the\ time\ series$ $(240\ in\ this\ case)$
	$U - uniform\ distribution$
Measurement Interval	$For\ each\ c$: $v_{c,k}(i) = \frac{1}{k} \sum_{t_0 = (i-1)*k+1}^{t = k+t_0-1} v_c(t), \qquad \forall i$
	$k = \{\ 1, 5, 15, 30, 60\}$ – measurement interval in minutes
Meter Resolution	$For\ each\ c$: $v_{c,k}(i) = round\left(v_{c,k}(i), d\right), \qquad \forall i$
	$d\ \in D\ where\ D = \{0, 1, 2\} - decimal\ places$
Meter Bias	$For\ Each\ c$: $V_{k,bias} = V_k + (b *\ \sigma_c)$
	$p\ \in \{0, 0.5, 1.0, 1.5, 2\}$ $-\max\ allowable\ percent\ bias$ $\sigma_c \sim U(-1,1) - bias\ scaling\ factor$ $b = (p/100 *\ \mu) - max\ allowable\ bias$
Measurement Noise	$For\ each\ c$: $v_{k,noise}(i) =\ v_k(i) + (n *\ \sigma_i), \qquad \forall i$
	$p\ \in \{0, 0.05, 0.25, 0.5, 0.75, 1.0, 1.25, 1.50, 1.75, 2.0\}$ $-maximum\ allowable\ percent\ noise$ $\sigma_i \sim U(-1,1) - noise\ scaling\ factor$ $n = (p/100 *\ \mu) - max\ allowable\ noise$
Time Synch	$For\ each\ c$: $\forall i, \qquad v_{c,1,timeSynch}(i) =\ v_1(i +\ s_c)$ $V_{c,timeSynch} = Truncate(V_{c,timeSynch}, (2 * f))$
	$f\ \in F, where\ F = \{1,2,3,4,5\}$ $-\max\ offset\ in\ minutes$ $s_c\ = U(-f, f)$ $-random\ scaling\ factor$ *Truncate(timeseries, truncationAmount) – truncates the beginning and end of the specified timeseries by the amount specified*
Missing Data	For Each Customer c: For ctr from 0 to h: $startPosition = U_{int}(0, \lvert I \rvert)$ $V_{c,k,Missing}\ [startPosition : (startPosition + g)] = NaN$
	$p = \{\ 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ $-percentages\ of\ missing\ data$ $h = floor\left(\frac{(p * \lvert V \rvert)}{g}\right)$ $-\ the\ number\ of\ missing\ data\ instances$ $g = number\ of\ samples\ missing$
Quantity of Data	$For\ Each\ c$: $V_{c,k,m} =\ V_{c,k,12}(1),\ V_{c,k,12}(2),\ V_{c,k,12}(3), \ldots, V_{c,k,12}(30 * m)$
	$m\ \in M\ where\ M = \{12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$ $-\ months\ of\ data$

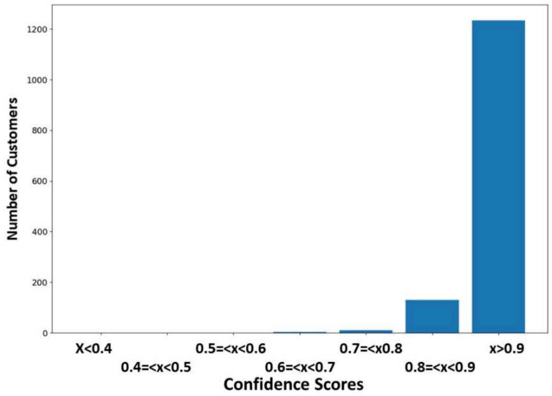


Figure 3 - Confidence scores for unmanipulated data at 15-minute measurement intervals

V. RESULTS

The following sections detail each of the data quality manipulations that were simulated on the synthetic dataset. Table 3 gives detailed formulas and definitions for each of the manipulations performed on the dataset.

The customer phase identification task was used to compare the efficacy of the data given each of these manipulations. A base case was obtained for each of these in isolation, meaning that the first experiments tested only one of these manipulations at a time. Subsequently a more realistic combination of data quality manipulations was constructed for a more practical test case of the effects. For example, in the base case measurement interval sweep, no other data manipulations are done, the otherwise unmanipulated data is used to test various measurement interval values. In the test case, the data is given each of the data manipulations shown in Table 4 for the test case and then the different values for the measurement interval are tested. The results shown in the sections below show both the base case accuracy and well as the test case accuracy. The values used for the test case are listed in Table 4.

The spectral clustering algorithm used for the phase identification task from [24] is a correlation-based method. It should be kept in mind that we expect these results and conclusions to be true in the context of correlation-based analysis of AMI data, and that other methodologies may show other sensitivities. Research is ongoing to test other types of methodologies under this framework.

TABLE 4 - TEST CASE PARAMETERS

Data Quality Manipulation	**Default Values Used in the Base Case**	**Default Value Used in the Test Case**
Measurement Interval	15-minutes	15-minutes
Meter Resolution	2 decimals	1 decimal
Meter Bias	0% maximum bias	0.2% maximum bias
Measurement Noise (Meter Precision)	0% maximum noise	0.2% maximum noise
Time Synchronization	No time synch issues	No time synch issues
Missing Data	0% missing data	0.2% missing data.
Available Data	12 months	6 months
Mislabeled Phases	0% mislabeled	10% mislabeled

A. Measurement Interval

This data manipulation tests the different intervals that utility companies may use for data collection. The original synthetic data was created at a 1-minute granularity, other common choices include 5, 15, 30, or 60-minute collection intervals. To obtain these intervals, the 1-minute granularity measurements were averaged using the appropriate number of measurements to obtain the new intervals. Table 5 shows the results of the phase identification task sweeping through plausible values for the AMI data collection interval. The number in parentheses in this table, as well as subsequent tables, represents the number of customers, out of 1379 customers total, that were predicted with an incorrect phase label. The phase identification algorithm [24] uses windows of 4 days, so if the measurement interval is larger, then each window includes fewer data points, shown in column 2. Columns 3 and 4 demonstrate that given otherwise perfect data, the accuracy on the phase identification task does not begin to degrade until the interval reaches 60-minutes when customers are identified on the wrong phase. However, this is a case where the accuracy metric does not show the whole picture of algorithm performance. Looking at the confidence scores shown in Figure 4, we can see the algorithm confidence is nearly identical for measurement intervals of 1-minute and 5-minutes, slight degradation at 15-minute intervals, and then significant degradation at 30-minute and 60-minute intervals. This suggests that the measurement granularity should be 15-minutes or less.

TABLE 5 - MEASUREMENT INTERVAL RESULTS

Measurement Interval	Window Size – 4 days	Base Case Accuracy	Test Case Accuracy
1-min	5760	100% (0)	100% (0)
5-min	1152	100% (0)	100% (0)
15-min	384	100% (0)	100% (0)
30-min	192	100% (0)	100% (0)
60-min	96	100% (0)	99.93% (1)

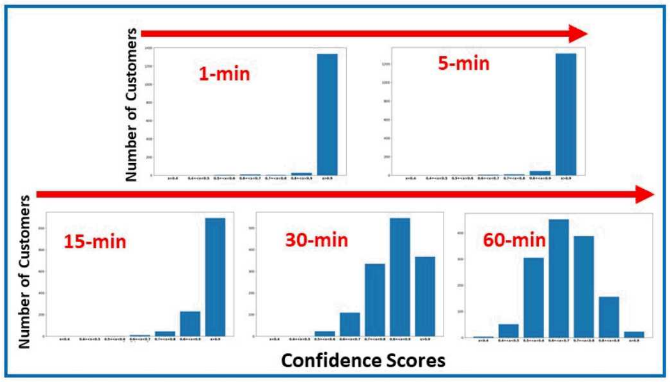


Figure 4 - Confidence scores for the measurement interval experiment for the test case

B. Meter Resolution

Meter resolution is the resolution with which the measurement is collected at the meter. Different resolutions are obtained by rounding each measurement point to the desired resolution. All simulations, both for the base case and the test case, returned with perfect accuracy on the phase identification task. However, looking at the confidence scores in Figure 5, we can see that at least one decimal point is required for the algorithm to have high confidence in the predictions.

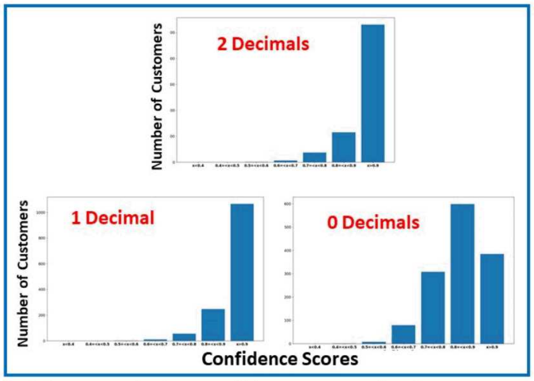


Figure 5 - Meter resolution confidence scores for the test case

C. Biased Meters

Meter bias is a common issue for meters in the field. We define meter bias here as a constant factor that is added to each measurement for a given meter. A percentage of the total meters and a range of potential biases are chosen and for each of those meters a bias value within the specified range is selected and added to each measurement in that meter. Meter bias was simulated using 100% of the meters, using maximum bias percentages from 0-2%, shifting all voltage measurements for that meter up or down. The meter bias had no effect on the accuracy of the phase identification task; all simulations had 100% accuracy. The primary reason for this is that the phase identification algorithm that is used here converts the voltage time series into a voltage difference representation, only considering the difference between adjacent measurements in each timeseries. This representation removes the effect from the injected meter bias.

D. Measurement Noise

Noise is defined as a measurement error that affects individual measurements in a random way. A percentage of the total meters is injected with noise, and individual meters up to that percentage are selected randomly. For each individual measurement in that set of selected customers, noise is randomly selected within an acceptable range and added to the measurement.

A baseline simulation for noise injection was run using a uniform noise injection, where the added noise was pulled from a uniform distribution. The maximum allowable noise percentages tested ranged from 0-2% maximum allowable noise, and noise was added to all meters under consideration. Looking at Table 6, for maximum allowable noise percentages up to 0.45%, the results on the phase identification were quite good on the test case. Starting with 0.45% maximum noise, accuracy began to decrease as the noise increased. Figure 6 plots the confidence scores for the test case simulation. We can see that starting at 0.25% maximum allowable noise, the confidence scores of the phase prediction begins to degrade rapidly (although accuracy remains high). This suggests a guideline of at least <0.25% maximum allowable noise in the meter. That guideline corresponds well to the Accuracy Case 0.2 from the ANSI standards [13].

It is worth noting the effect of randomness in the test case. At 0.05% noise there was one customer incorrectly identified. In the test case, given the configuration of other data manipulations present in the dataset as a whole, and for that customer, even that level of noise was enough to cause a misclassification. Thus, although the algorithm performance overall is quite good at low levels of noise it is important to keep in mind the random factors at work.

TABLE 6 - MEASUREMENT NOISE RESULTS

Max Noise Percentage	Base Case Accuracy	Test Case Accuracy
Original	100% (0)	100% (0)
0.05%	100% (0)	99.93% (1)
0.15%	100% (0)	100% (0)
0.25%	100% (0)	100% (0)
0.35%	100% (0)	100% (0)
0.45%	100% (0)	99.93% (1)
0.55%	100% (0)	99.71% (4)
0.65%	100% (0)	98.40% (22)
0.75%	99.93% (1)	96.37% (50)
0.85%	99.42% (8)	94.92% (70)
0.95%	99.13% (12)	91.88% (112)
1.0%	99.10% (13)	91.81% (113)

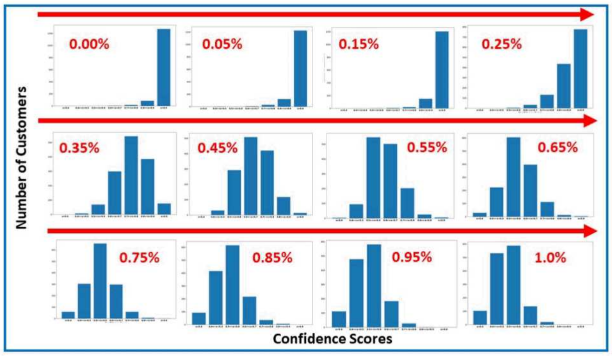


Figure 6 - Measurement noise confidence scores for the test case

E. *Time Synchronization*

This alteration simulates a meter's clock being out of synch with the other meters by a specified number of minutes. This has the effect of shifting the values of that meter relative to the other meters. Time synchronization issues were simulated from 1-5 minutes of allowable offset. For this simulation, all customers were given a random time synchronization error within the range [- max offset, + max offset]. Using a measurement interval of 15-minutes the simulations remained completely accurate. We believe this is because the effect of using 15-minute averages alleviates the consequences of time synchronization issues. Investigation of the effects of time synchronization at 1-minute measurement intervals shows that the accuracy quickly degrades in the presence of time synchronization issues. This suggests that there is value in the averaging effect of using larger measurement intervals than 1-minute.

F. *Missing Data*

A percentage of the total data is specified to be missing and the number of measurements totaling that percentage is removed from the dataset and replaced with 'NaN' as a marker. Data can be missing for a variety of reasons, resulting in varying lengths of periods where data is missing. For this experiment we have chosen to remove data in 4-hour blocks. 4-hours was chosen because that is representative of the median missing data block size for the dataset used in the phase identification research in [24]. Values tested for missing data percentages were in the range [0%,1.0%] incrementing by 0.1%. All simulations for both the base case and test case were 100% accurate in the phase identification task.

It is important to recognize that this particular phase identification algorithm is sensitive to the *distribution* of missing data. Recall from Figure 2 that in each window, any customer whose time series contains missing values is removed from consideration during that window. Thus, missing data simply decreases the total number of available windows for a customer, and provided that enough windows remain, the clustering and phase prediction is unaffected. However, consider the case where there happens to be missing data approximately uniformly distributed such that there is a missing data point in each available window; in such a case this phase identification algorithm would fail because there would be no windows without missing data for that customer. This 12-month dataset, using 4-day windows and 4-hour missing data instances requires < ~4% total missing data.

G. *Available Data*

This alteration changes the amount of data that is used in the simulations. There are 12-months total in the synthetic dataset, and this alteration simply removes portions of that data from consideration. Values were tested in one month increments from 12-months to 1-month. Table 7 shows the results from this sweep. The results indicate that more than 4 months of available data is required for accurate phase identification under the test case conditions. Larger percentages of missing data may result in a larger data availability requirement.

TABLE 7 - DATA AVAILABILITY RESULTS

Available Data (months)	Base Case Accuracy	Test Case Accuracy
12	100% (0)	100% (0)
11	100% (0)	100% (0)
10	100% (0)	100% (0)
9	100% (0)	100% (0)
8	100% (0)	100% (0)
7	100% (0)	100% (0)
6	100% (0)	100% (0)
5	100% (0)	100% (0)
4	100% (0)	99.93% (1)
3	100% (0)	100% (0)
2	99.78% (3)	99.50% (7)
1	99.42% (8)	98.48% (16)

H. Data Quality Impact with Mislabeled Customer Phases

The next set of experiments evaluates the effect of injecting customers with mislabeled phases into the dataset and varying the percentage of mislabeled customers. Since some utility distribution system models may have more problems than others, the accuracy of the algorithm is tested for a range of percentage of customers that have their phase mislabeled in the original utility model before doing the phase identification spectral clustering. This experiment more clearly indicates the effects of the test case combination. A percentage of the total customers is selected for their phase label to be replaced by an incorrect label. Values tested for the percentage of mislabeled customers were from 0%-50% and the results are shown in Figure 7.

The mislabeled phases experiment was run as a Monte Carlo simulation because the configuration of customers chosen to have incorrectly labeled phases is significant in the ability of the phase identification algorithm to correctly label those customers. There are configurations that are more difficult to correct the than other configurations. The 500 run Monte Carlo simulation is designed to quantify those affects. Each value for the percentage of customers mislabeled was run 500 times with different customers randomly mislabeled each time. Thus, the accuracies shown represent the average of all 500 runs. We can see that even at low percentages of the customers mislabeled, it is possible to occasionally misclassify a customer's phase. This illustrates that which customers are mislabeled is a crucial issue for algorithm performance. The (average) accuracy slowly degrades as the percentage of customers mislabeled increases up to 45% where we still see a reasonable overall average accuracy of ~95%. The base case and test case accuracies are nearly identical, further illustrating that the choice of mislabeled customers dominates the accuracy in this case.

In this version of the algorithm, the utility phase labels are used to assign the predicted phases after the clustering of each window. Therefore >50% of the phase labels must be accurate. However, the algorithm could be adjusted to assign each customer one of three 'placeholder' phases, resulting in an accurate final clustering of the customers. In that case, it would remain up to the utility to determine which 'placeholder' phase label corresponds to which actual phase label.

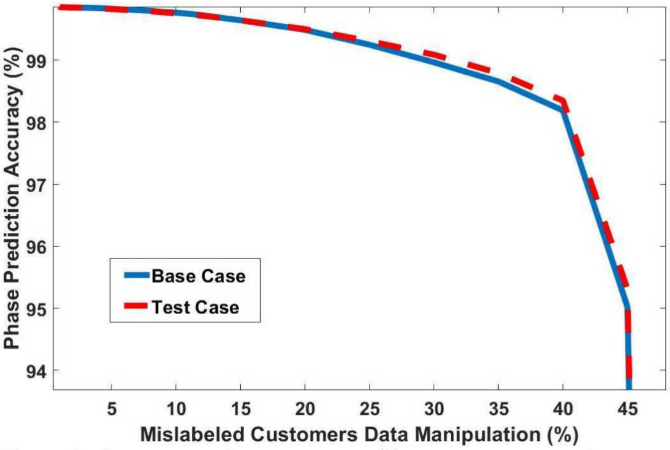


Figure 7 - Sweep over the percentage of incorrect customer phase labels

VI. DISCUSSION AND RECOMMENDATIONS

Table 8 shows a summary of conclusions that can be drawn from the experiments described previously. These conclusions are in context of the phase identification task using a correlation-based method. Although that focus is relatively narrow, the following conclusions do begin to shape the considerations for data quality in AMI data collection. For the measurement interval parameter, the time synchronization experiments demonstrated that the averaging to intervals larger than 1-minute smoothed some of the effects of time synchronization errors, and the confidence scores from the measurement intervals experiments (Figure 4) show that the measurement interval should be less than 30-minutes. That leaves acceptable measurement interval choices of 5-minutes and 15-minutes, and the confidence scores were slightly improved at the 5-minute interval. The meter resolution experiments (Figure 5) clearly show that at least one decimal point is required. The confidence scores begin to degrade at ~0.25% maximum noise in meter measurements. The results also show that more than 4 months of AMI data is preferred for this task Table 7. The algorithm is sensitive to the distribution of missing data points, and so the more data that is missing, the longer period of available data will be required to make up for the missing data. Meter bias is not a consideration for this methodology but may need to be considered in other contexts.

TABLE 8 - AMI DATA QUALITY CONSIDERATIONS

Data Quality Manipulation	AMI Considerations Based on the Phase Identification Task
Measurement Interval	5 - 15-minute intervals are recommended
Meter Precision	At least 1 decimal on voltage measurements (240V) is required
Meter Bias	Bias does not impact phase identification results with this algorithm
Measurement Noise	< 0.25% maximum uniform random noise is recommended
Time Synchronization	> 1-min measurement intervals are required to account for the time synchronization errors
Missing Data	Sensitive to the *distribution* of missing data. Given uniformly distributed 4-hr missing data instances, with this algorithm, the percentage of missing data is required to be < ~4%
Data Availability	> 4 months of AMI voltage data are required

VII. CONCLUSIONS

These preliminary results show the importance of what type of AMI data collection techniques are employed. These results can also begin to inform AMI data collection techniques and illustrate the types of considerations that utilities need to consider when implementing data collection policies. Table 8 shows a set of recommendations based on this research. AMI data collection considerations have a significant impact on the ability of utilities to validate and correct errors in their models of the distribution system, and the accuracy of the model is directly important for DER planning, interconnection studies, and operations. Knowing what type of data to collect, what its limitations are, and what can be done with that data will allow rapid progress towards high-penetrations of DER safely and in optimal locations.

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 multi-mission 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-NA-0003525

VIII. REFERENCES

[1] B. Palmintier *et al.*, "On the Path to SunShot: Emerging Issues and Challenges in Integrating Solar with the Distribution System," *Natl. Renew. Energy Lab.*, vol. NREL/TP-5D00-65331, 2016.

[2] W. Luan, J. Peng, M. Maras, B. Harapnuk, and J. Lo, "Smart Meter Data Analytics for Distribution Network Connectivity Verification," *IEEE Trans. Smart Grid*, vol. 6, p. 1, Jul. 2015.

[3] J. Peppanen, S. Grijalva, M. J. Reno, and R. J. Broderick, "Distribution System Low-Voltage Circuit Topology Estimation using Smart Metering Data," *IEEE PES Transm. Distrib. Conf. Expo. Dallas TX*, 2016.

[4] X. Zhang and S. Grijalva, "A Data-Driven Approach for Detection and Estimation of Residential PV Installations," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sep. 2016.

[5] R. Mitra *et al.*, "Voltage Correlations in Smart Meter Data," *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1999–2008, 2015.

[6] Blakely, M. J. Reno, and J. Peppanen, "Identifying Common Errors in Distribution System Models," *Photovolt. Spec. Conf. PVSC*, Jun. 2019.

[7] R. R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A Survey on Advanced Metering Infrastructure," *Int. J. Electr. Power Energy Syst.*, vol. 63, pp. 473–484, Dec. 2014.

[8] "Smart Meters At A Glance." The Edison Foundation Institute for Electric Innovation (IEI), Mar-2019.

[9] N. Uribe-Pérez, L. Hernández, D. de la Vega, and I. Angulo, "State of the Art and Trends Review of Smart Metering in Electricity Grids," *Appl. Sci.*, vol. 6, no. 68, 2016.

[10] S. Parker *et al.*, "Metering Best Practices: A Guide to Achieving Utility Resource Efficiency, Release 3.0," *Pac. Northwest Natl. Lab. PNNL - Prep. US Dep. Energy DOE*, Mar. 2015.

[11] "Advanced Metering Infrastructure and Customer Systems: Results From the Smart Grid Investment Grant Program," *US Dep. Energy Off. Electr. Deliv. Energy Reliab.*, Sep. 2016.

[12] M. Sheppy, A. Beach, and S. Pless, "Metering Best Practice Applied in the National Reneweable Energy Laboratory's Research Support Facility," *Natl. Renew. Energy Lab. NREL Tech. Rep. - TP-5500-57785*, Apr. 2013.

[13] "American National Standards Institute," 2015. [Online]. Available: http://www.nema.org/stds/c12-20.cfm.

[14] "Accuracy of Digital Electricity Meters," *Electr. Power Res. Inst. EPRI*, May 2010.

[15] "Evaluation of Advanced Meter System Deployment in Texas - Meter Accuracy Assesment," *Navig. Consult. PI LLC*, Jul. 2010.

[16] 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.

[17] H. Pezeshki and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2012, pp. 1–7.

[18] 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 *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 259–265.

[19] J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd Edition. University Science Books, 1997.

[20] P. R. Bevington, *Data Reduction and Error Analysis*, 3rd Edition. McGraw-Hill, 2003.

[21] D. Montenegro, R. C. Dugan, and M. J. Reno, "Open Source Tools for High Performance Quasi-Static-Time-Series Simulation Using Parallel Processing," *IEEE Photovolt. Spec. Conf.*, 2017.

[22] J. Fuller, W. Kersting, R. Dugan, and S. C. Jr., "Distribution Test Feeders," *IEEE PES AMPS DSAS Test Feeder Working Group*, 2013. [Online]. Available: http://sites.ieee.org/pes-testfeeders/.

[23] "Pecan Street Database," *Pecan Street*. [Online]. Available: http://www.pecanstreet.org.

[24] L. Blakely, M. J. Reno, and W. Feng, "Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries," *Power Energy Conf. Ill. PECI*, Feb. 2019.