Phase Identification of Smart Meters Using a Fourier Series Compression and a Statistical Clustering Algorithm

Jeremy Chiu
Mathematics and Statistics
Langara College
Vancouver, Canada
0000-0002-0737-9055

Albert Wong
Mathematics and Statistics
Langara College
Vancouver, Canada
0000-0002-0669-4352

James Park
Mathematics and Statistics
Langara College
Vancouver, Canada
0000-0002-3714-9138

Joe Mahony
Research and Development
Harris SmartWorks
Ottawa, Canada
JMahony@harriscomputer.com

Michael Ferri
Research and Development
Harris SmartWorks
Ottawa, Canada
mferri@harriscomputer.com

Tim Berson
Research and Development
Harris SmartWorks
Ottawa, Canada
TBerson@harrisutilities.com

***Abstract*—Accurate labeling of phase connectivity in distribution systems is important for maintenance and operations but is often erroneous or missing. In this paper, we present an algorithm to identify which smart meters must be in the same phase using a hierarchical clustering method on voltage time series data. Instead of working with the time series directly, we apply the Fourier transform to represent time series in their frequency domain, remove 98% of the Fourier coefficients, then cluster the remaining coefficients to estimate which meters belong in the same phase. We validate results by verifying they do not change phase in time and by comparing our results to available network-distribution data.**

***Index Terms*—Phase identification, clustering, Fourier series, Fourier series compression**

I. Introduction

Managing an electricity distribution network efficiently requires accurate phase connectivity models [18]. However, electricity companies usually do not have accurate information of phase connectivity and often require the use of measurement-based phase identification methods. [8].

To deliver high-voltage power from the generation station to customers, voltage in the primary distribution circuit is stepped down at a distribution substation. Then through feeders electricity is distributed to transformers. In North America, power is stepped down again from transformers and distributed to the customers using a three-phase system [18]. Which phase is used for the customers is often not recorded, and therefore creating a phase identification problem if phase connection information is required for network management tasks.

We would like to acknowledge and thank the Post Degree Diploma program, the Work on Campus program, and the Applied Research Centre at Langara College for supporting our research.

There are many ways in research to tackle this identification problem:

Micro-synchrophasors - One can use a micro-synchrophasor to measure voltage magnitude and phase angle of a meter [19]. The higher the correlation between the voltage magnitude of the substation and that of smart meters, the more accurate the phase labelling. To complete the identification, signal generators are set up at the substations and signal discriminators at the smart meters to accurately identify the phase. This method is quite accurate but expensive as it requires deployment and maintenance of additional equipment and human resources.

Integer Programming algorithms ([2], [3], [7], [22]) - Phase connection of smart meters are represented as binary variables, then integer linear programming methods are used to determine the most-likely phase network. However, this approach requires a new variable for every new meter, making the problem computationally intensive, especially for feeders with thousands of meters.

Correlation-based method ([14]–[16]) - Data is first collected over time from the smart meters to be identified. The correlation coefficient is then calculated using voltage time series between two smart meters – the closer a coefficient is to one, the more likely the pair of smart meters have the same voltage pattern and therefore the same phase. The correlation coefficients are then transformed to a distance measure as input to a clustering algorithm. The method is logical and seems promising. However, based on results from unpublished research by a project team at Langara College (personal communication), when applied to the data set in this research, this method suffers from issues with a number of performance criteria that we have

identified and discussed below.

Constrained k-means clustering - Voltage time series data is first normalized using standard deviation, then principal component analysis is applied to reduce the data's dimension. A k-means clustering algorithm is then used to cluster the smart meters. The phase of each cluster is then identified by solving a minimization problem [9], [14], [18].

Other phase identification methods proposed include the use of supervised learning models or different types of clustering algorithm, such as spectral clustering [4]–[6], [10], [11], [17], [20], [21].

In this research, we will take a new approach in the phase identification problem. The central idea is to extract as much information as possible from the voltage time series using a Fourier series compression process. A hierarchical clustering routine is then applied on the compressed data to produce accurate identification.

II. Research Data Set

For this research, we use a voltage data set that was provided by a utility company in the United states, which contains hourly voltage data for a number of smart meters in the month of June and July 2021. The data set also include the linkage between the smart meters and their associated transformers and feeders. This information is critical for the assessment of appropriateness and accuracy in the clustering results.

We removed smart meters with any missing entries from June and July 2021. We then normalize each smart meter by dividing each voltage value by its mean. We chose two of the smaller feeders (Feeder F with 26 smart meters and Feeder D with 55 smart meters) to conduct our research so that we can easily visualize and evaluate the results.

III. Fourier Compression

Clustering the smart meters using its time series (voltage vs time) is challenging because of its size – measurements are hourly, so in a month of 30 days, each time series would be in $\mathbb{R}^{720}$. We reduce the dimension by using a compressed Fourier series, then cluster the smart meters using the compressed Fourier series. Figure 1 shows a high level overview of how we use Fourier series to reduce the dimension.

The compression is done as follows. We represent each smart meter in its frequency domain by applying the Fourier transform to the normalized time series. Recall the Fourier series (sine-cosine form) representation of a periodic function $f(t)$ is

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\tfrac{2\pi}{P} nt\right) + b_n \sin\left(\tfrac{2\pi}{P} nt\right) \right), \quad (1)$$

where $a_n,\ b_n$ are real coefficients and P is the function's period. We then delete coefficients that are 'small' (either by deleting frequencies that are smaller in magnitude than a predetermined magnitude, or by only keeping a predetermined number of the largest terms), thus giving us a compressed Fourier representation. We also delete the 0th harmonic a_0 because it is constant across all smart meters due to normalization. In practice, we used 12 Fourier coefficients to represent a month of data, thus reducing the dimension from $\mathbb{R}^{720}$ to $\mathbb{R}^{12}$ (a 98% reduction in size).

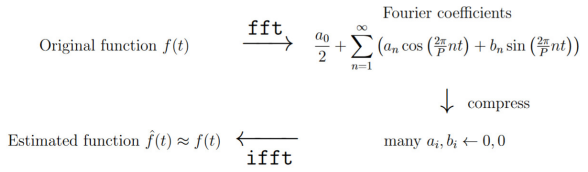

Fig. 1. A high level overview of how we use Fourier series to reduce the dimension of a smart meter. We performed clustering on the compressed Fourier series. The functions $f(t)$ and $\hat{f}(t)$ are time series, where $\hat{f}(t) \approx f(t)$.

As demonstrated in Figure 2, most of the Fourier coefficients are very small, which suggests the compressed Fourier series could provide a high-accuracy, low-dimension approximation of the time series. To verify the accuracy of the compression, we obtain an approximate time series by applying the inverse Fourier transform to a compressed Fourier series, and then comparing the approximate time series to the original time series. Figure 3 shows approximate time series alongside the original time series – the general trend of the time series is captured, but the 12-coefficient approximation does poorly at the spikes. As Figure 4 demonstrates, keeping more coefficients yields better accuracy. Notice that with about 10% of the coefficients, we maintain about 90% accuracy of the time series. Ultimately, the accuracy of the time series is not too important, so long as the clustering results are sensible.

The compression was done in Matlab. Given a smart meter's time series, we use Matlab's `fft` function, which returns complex coefficients corresponding to the Fourier series in exponential form. We convert the complex coefficients into a_n and b_n, the real coefficients of the Fourier series in sinusoidal form (we used `get_harmonics` [1]). In practice, a time series in June would be in $\mathbb{R}^{720}$, corresponding to $0 \leq t \leq 720$ hours, and so $P = 720$. Matlab's `fft` would return the complex coefficients $c_{-360}, \ldots, c_{359}$, which we convert to real coefficients, then only keep $a_1, \ldots, a_{360}$ and $b_1, \ldots, b_{360}$ (note a_{360} and b_{360} were computed from a_{-360} and b_{-360}). We then compress by using a mask to set most coefficients to zero. In practice, we kept a_n and b_n where $n = 30, 60, \ldots, 180$ (these coefficients correspond to the large frequencies in Figure 2), a total of 12 coefficients.

IV. Clustering of Smart Meters

After the dimension of the data is reduced through a Fourier compression, distance between smart meters'

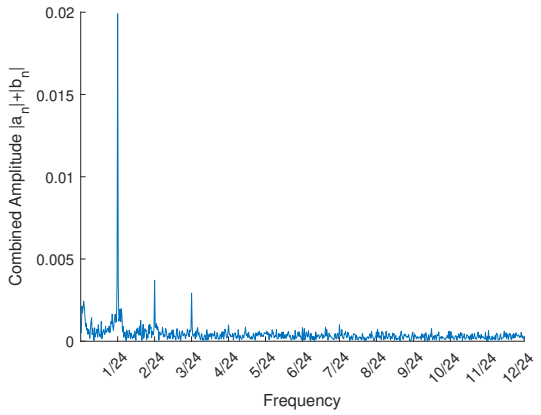

Fig. 2. Combined magnitude of the Fourier coefficients ($|a_n|+|b_n|$) vs frequency. Notice most of the coefficients are small. The largest amplitude occur at the frequency $1/24$; this is unsurprising because energy usage follow daily patterns.

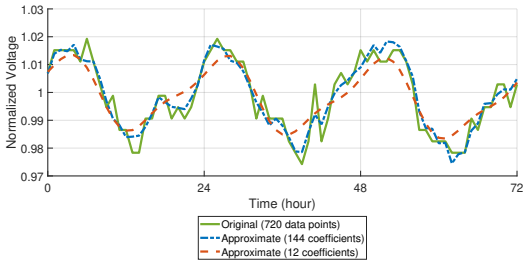

Fig. 3. Original time series alongside approximate time series. The domain was reduced to 3 days for a better viewing rectangle. The 12-coefficient approximation does poorly at the spikes, but captures the general trend. The 144-coefficient approximation captures most spikes.

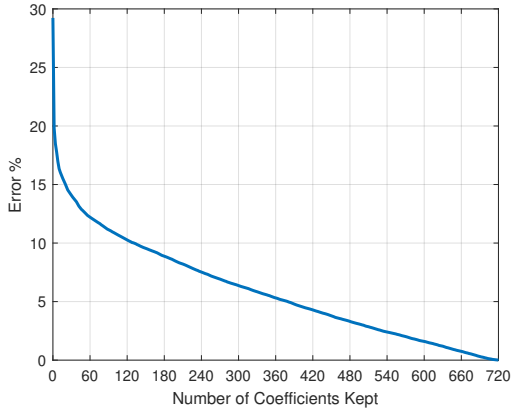

Fig. 4. Error percentage is computed as $\frac{\|y-\hat{y}\|_2}{\bar{y}}$, where y is the original time series, $\hat{y}$ is the approximate time series, and $\bar{y}$ is the average of y (note $\bar{y}=1$ due to normalization). The 12-coefficient approximation has 16% error.

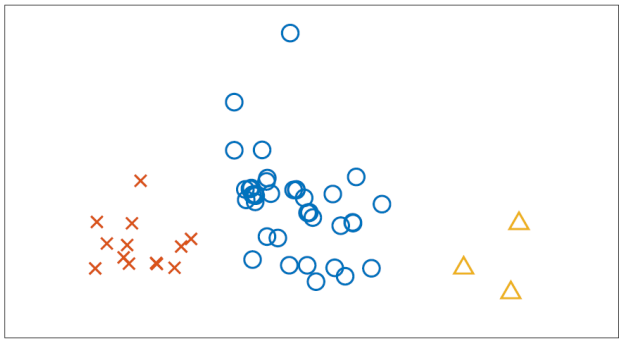

Fig. 5. Visualizing the clustering of Feeder D using June 2021 Data via Matlab's `mdscale` function.

Fourier coefficients $D(X,Y)$ is calculated using the traditional Euclidean distance metric:

$$D(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)^2. \tag{2}$$

Using this distance, we cluster the set of smart meters in Feeder F (then repeat for Feeder D) using the Ward hierarchical clustering algorithm in Matlab [12]. Since all smart meters should be in one of the three phases, the number of resulting clusters is set to be three. Hence, meters clustered together would mean they belong to the same phase.

V. Validation of Clustering Results

A. Visualizing Clustering Results

A useful way to visualize the result of clustering a multi-dimensional data set is to somehow "project" the data set into a two dimensional space. We could then visualize clusters with a scatter diagram in the xy-plane. Since we are using the Euclidean distance as the basis for clustering, a natural way to achieve this is to use Matlab's multidimensional scaling technique [13]. Given the distance between points, `mdscale` reconstructs where the points could be in 2D so that the distance is still roughly preserved. In Figure 5, we see a visualization of the clustered meters from Feeder D. Notice that there are clear boundaries between different clusters.

Moreover, a hierarchical clustering algorithm such as Ward would allow us to visualize the formation of the clusters hierarchically via a dendogram (Figure 6). However, it is less useful here because the number of clusters is required to be three.

B. Same Transformer, Same Phase

Meters within the same transformer must be in the same phase, and thus should be clustered together. We can use this fact to see how well our method performs – after we cluster the smart meters, each transformer should only have meters of a single phase. As seen in Tables I and II, the clustering of Feeder F is almost perfect while that for

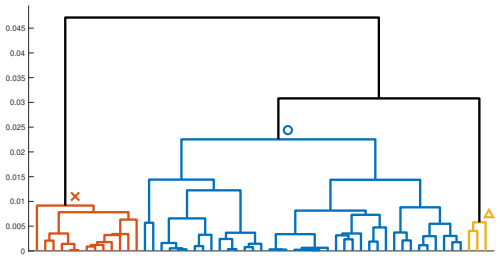

Fig. 6. Dendogram of clustering Feeder D using data from June 2021. Dendograms are useful to see how clusters are being formed.

Feeder D is perfect, giving us hope that this approach has promise.

	Cluster		
Transformer	A	B	C
1	1	8	
2	1		
3	1		
4	1		
5			1
6	4		
7			1
8			1
9			1
10			1
11	5		
Total	13	8	5

TABLE I
FEEDER F JUNE 2021 CLUSTER RESULTS GROUPED BY TRANSFORMERS.

C. Stability Over Time

Physically, meters do not change phase over time. Therefore, for the clustering (assignment of phase) to be meaningful, the result should not change over time.

To evaluate results from this research, we performed cluster analysis on two different time periods (June 2021 and July 2021) on Feeder F and D, then checked for inconsistent results. Any meter that changed phases (clusters) are considered time unstable. Note that the labels from the clustering (A B and C) are arbitrary, and so we use a cross tabulation of the two clustering results to see how meters are assigned in the clustering processes. Table III shows that the clustering from June to July is stable. All 13 meters assigned to Cluster A in June are also assigned in the same cluster in July; the same is true for Clusters B and C.

The same can be said about the stability of clustering Feeder D using our approach (Table IV).

	Cluster		
Transformer	A	B	C
1	1		
2	1		
3	1		
4	1		
5	2		
6	1		
7			1
8	1		
9	1		
10	4		
11	1		
12			1
13			1
14	1		
15		1	
16	1		
17	2		
18	1		
19	1		
20		1	
21		1	
22		3	
23		1	
24		2	
25	2		
26		1	
27	1		
28	1		
29		1	
30	1		
31	2		
32	1		
33	4		
34	1		
35	1		
36		1	
37	2		
38	3		
39		1	
Total	39	13	3

TABLE II
FEEDER D JUNE 2021 CLUSTER RESULTS GROUPED BY TRANSFORMERS.

VI. FUTURE WORK

While the above results look very promising, we have not applied this approach to a larger feeder (say with over 300 meters), or to a data set with multiple feeders. We suspect, due to the increased likelihood of data related issues, that the results may not be as "perfect" as we have seen so far.

To advance our research, the approach would be applied to a larger data set with multiple feeders. The same

		July			
		A	B	C	Total
June	A	13			13
	B		8		8
	C			5	5
	Total	13	8	5	26

TABLE III
CLUSTERING FEEDER F - JUNE AND JULY 2021

		July			
		A	B	C	Total
June	A	39			39
	B		13		13
	C			3	3
	Total	39	13	3	55

TABLE IV
CLUSTERING FEEDER D - JUNE AND JULY 2021

approach should also be applied to a data set with several months; clustering could be done month by month, or with several months combined. Considerations should also be given to use this approach to cluster a subset of the data set and, after the validation process as outlined above, using the cluster labels for the development of a supervised learning model for the classification of other meters.

VII. CONCLUSION

In this research, we have applied a novel method of approximating a time series with its Fourier series. We then used hierarchical clustering methods on the dimension-reduced data. The major application of this approach is in the phase identification of smart meters in a network environment.

Results from two small data sets using this approach show significant promise as they passed two important tests: same assignment for meters in the same transformer and stability of assignment over time. The application of this approach to a larger data set with multiple feeders would therefore be a worthwhile exercise.

REFERENCES

[1] A. ADELMALEK, *Get harmoniques of a real signal*, 2022. Last accessed 13 October 2022.

[2] A. H. AKHIJAHANI, S. HOJJATINEJAD, AND A. SAFDARIAN, *A milp model for phase identification in lv distribution feeders using smart meters data*, in 2019 Smart Grid Conference (SGC), IEEE, 2019, pp. 1–6.

[3] V. ARYA, D. SEETHARAM, S. KALYANARAMAN, K. DONTAS, C. PAVLOVSKI, S. HOY, AND J. R. KALAGNANAM, *Phase identification in smart grids*, in 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2011, pp. 25–30.

[4] L. BLAKELY, M. J. RENO, AND W.-C. FENG, *Spectral clustering for customer phase identification using ami voltage timeseries*, in 2019 IEEE Power and Energy Conference at Illinois (PECI), IEEE, 2019, pp. 1–7.

[5] B. FOGGO AND N. YU, *A comprehensive evaluation of supervised machine learning for the phase identification problem*, International Journal of Computer and Systems Engineering, 12 (2018), pp. 419–427.

[6] ———, *Improving supervised phase identification through the theory of information losses*, IEEE Transactions on Smart Grid, 11 (2019), pp. 2337–2346.

[7] A. HEIDARI-AKHIJAHANI, A. SAFDARIAN, AND F. AMINIFAR, *Phase identification of single-phase customers and pv panels via smart meter data*, IEEE Transactions on Smart Grid, 12 (2021), pp. 4543–4552.

[8] A. HOOGSTEYN, M. VANIN, A. KOIRALA, AND D. VAN HERTEM, *Low voltage customer phase identification methods based on smart meter data*, Electric Power Systems Research, 212 (2022), p. 108524.

[9] S. P. JAYADEV, A. RAJESWARAN, N. P. BHATT, AND R. PASUMARTHY, *A novel approach for phase identification in smart grids using graph theory and principal component analysis*, in 2016 American Control Conference (ACC), IEEE, 2016, pp. 5026–5031.

[10] H. P. LEE, M. ZHANG, M. BARAN, N. LU, P. REHM, E. MILLER, AND M. MAKDAD, *A novel data segmentation method for data-driven phase identification*, arXiv preprint arXiv:2111.10500, (2021).

[11] Y. MA, X. FAN, R. TANG, P. DUAN, Y. SUN, J. DU, AND Q. DUAN, *Phase identification of smart meters by spectral clustering*, in 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), IEEE, 2018, pp. 1–5.

[12] MATHWORKS, *linkage*, 2022. Last accessed 17 October 2022.

[13] ———, *mdscale*, 2022. Last accessed 17 October 2022.

[14] F. OLIVIER, A. SUTERA, P. GEURTS, R. FONTENEAU, AND D. ERNST, *Phase identification of smart meters by clustering voltage measurements*, in 2018 Power Systems Computation Conference (PSCC), IEEE, 2018, pp. 1–8.

[15] H. PEZESHKI AND P. WOLFS, *Correlation based method for phase identification in a three phase lv distribution network*, in 2012 22nd Australasian Universities Power Engineering Conference (AUPEC), 2012, pp. 1–7.

[16] T. A. SHORT, *Advanced metering for phase identification, transformer identification, and secondary modeling*, IEEE Transactions on Smart Grid, 4 (2012), pp. 651–658.

[17] W. WANG AND N. YU, *Advanced metering infrastructure data driven phase identification in smart grid*, 07 2017.

[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] M. H. WEN, R. ARGHANDEH, A. VON MEIER, K. POOLLA, AND V. O. LI, *Phase identification in distribution networks with micro-synchrophasors*, in 2015 IEEE Power & Energy Society General Meeting, IEEE, 2015, pp. 1–5.

[20] N. ZARAGOZA AND V. RAO, *Phase identification of power distribution systems using hierarchical clustering methods*, in 2021 North American Power Symposium (NAPS), 2021, pp. 1–6.

[21] ———, *Phase identification of power distribution systems using hierarchical clustering methods*, in 2021 North American Power Symposium (NAPS), IEEE, 2021, pp. 1–6.

[22] J. ZHU, M.-Y. CHOW, AND F. ZHANG, *Phase balancing using mixed-integer programming [distribution feeders]*, IEEE Transactions on Power Systems, 13 (1998), pp. 1487–1492.