Semi-Supervised, Non-Intrusive Disaggregation of Nodal Load Profiles with Significant Behind-the-Meter Solar Generation

Daniel Moscovitz, Zhenyu Zhao, Liang Du, *Senior Member, IEEE,* Xiaoyuan Fan, *Senior Member, IEEE*

***Abstract*—It is of imperative interests for regional transmission organizations (RTOs) to effectively extract actual load profiles at transmission nodes with significant behind-the-meter solar generation, which remains a gap in the existing technology paradigm. This paper proposes an explicit yet efficient linear estimator to disaggregate actual load profiles at transmission buses with significant behind-the-meter (BTM) solar generations. The proposed estimator is based on disaggregating (i.e., extracting) at locations close to transmission buses under consideration. To overcome the lack of ground truth and validate the performance of the proposed algorithms, we first propose semi-supervised mechanisms with parameter tuning as well as unsupervised clustering and leverage the unique characteristics of zero-crossing points in BTM solar peaking behaviors, which we refer to as Zone-to-Node (Z2N) methods. Next, we further propose a bi-level Node-to-Node (N2N) framework that improves the overall disaggregation performances compared to Z2N. Numerical results are presented using real-world data at PJM Interconnection.**

***Index Terms*—behind-the-meter solar, load disaggregation, load modeling**

I. Introduction

High penetration of renewable energy resources, especially the widespread installation of solar generation, have imposed opportunities and challenges to transmission grid operators, such as regional transmission organization (RTOs) on system operation, planning, and control. Due to its inherently volatile nature, high penetration of solar generation can impose unforeseen, significant challenges to both supply (e.g., recent 2020 San Fernando disturbance in California [1] and 2021 Odessa disturbances in Texas [2]) and demand (e.g., recent near-zero net system demand by the California ISO on April 16, 2023), which justifies the imminent needs of granular nodal load and solar generation profiles across transmission networks.

From transmission grid operators' view, solar penetration can be categorized according to its impacts on the generation capacity (e.g., front-of-the-meter (FOM)) and net demands (e.g., behind-the-meter (BTM)). FOM solar resources have direct access to transmission networks. As of the end of 2022, there were over 10,000 projects in the U.S. representing 1,260 GW of renewable generation capacity (including 947 GW of solar), 680 GW of storage capacity, and 457 GW of co-located hybrid capacity (mostly solar-plus-storage) waiting for transmission access [3]. Moreover, BTM resources represent small-scale installations physically located "behind" utility meters (feeders) or substation telemetry (nodes), i.e., without sub-metering of individual installations and thus unobservable to system operators. For instance, Australia has reported a total capacity of 8.03 GW BTM solar as of April 2019 [4] and California Energy Commission has estimated over 40,000 GWh annual BTM solar energy by 2030.

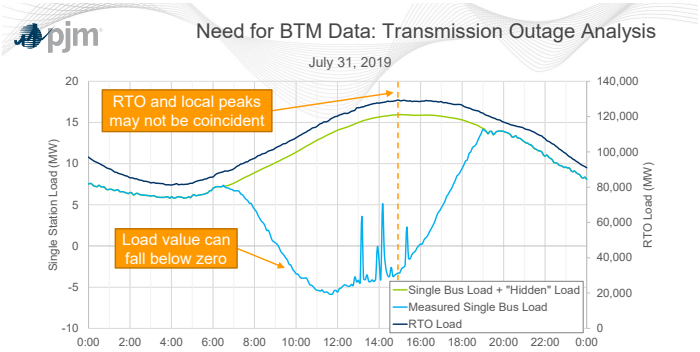


Fig. 1. A real-world example at PJM that needs actual load profiles: the peak load time of the entire RTO (the top line) could be one hour after the peak load time of a transmission bus (the middle line), whose actual load curve is unknown due to BTM solar (the bottom line) and self-managed loads [5]

.

Compared to distribution networks, the impact of significant BTM solar on transmission grids has generally been overlooked. As a real-world example, Fig. 1 shows a post-outage transmission analysis by PJM Interconnection (PJM), in which system operators need to know the actual peak time at a specific transmission bus (also interchangeably called transmission node) for resource planning and dispatching. However, the metered single-bus (i.e., nodal) load profile (the bottom line, which is available to grid operators) is the aggregation of significant BTM solar (the bottom line) and actual demands (the middle line, which is desired but unknown) and could be of negative values for hours. Therefore, it is of critical practice for grid operators to "disaggregate" actual load profiles at transmission nodes from such aggregated nodal measurements for system reliability under high penetration of BTM solar.

However, it is not practical nor economical to install metering devices throughout the transmission network. Instead, extracting and analyzing power consumption profiles from

Manuscript received October 21, 2022; revised April 2, 2023 and September 7, 2023; accepted on November 12, 2023. Date of publication XXX, 2023; date of current version XXX, 2023. L. Du was supported in part by the National Science Foundation (NSF) under Award 2238414. Paper no. TPWRS-01598-2022. (Corresponding author: Liang Du.)

D. Moscovitz is with PJM Interconnection, Audubon, PA 19403, USA. E-mail: daniel.moscovitz@pjm.com

Z. Zhao and L. Du are with the Department Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122, USA. E-mail: {z.zhao, ldu}@temple.edu.

X. Fan is with the Pacific Northwest National Laboratory, Richland, WA 99354 USA. E-mail: xiaoyuan.fan@pnnl.gov.

aggregated, metered waveforms without the installation of any extra metering devices, i.e., *non-intrusive* load disaggregation [6]–[8], has the technical potential to alleviate the aforementioned challenges. In the existing literature, with the proliferation of advanced metering infrastructure (AMI), a significant amount of recent efforts have been dedicated to non-intrusively disaggregate load and BTM solar profiles in distribution networks from aggregated AMI readings (e.g., smart meters). Specifically, the state-of-the-art largely aims at extracting BTM solar generation (instead of load) profiles from metered (aggregated) power consumption waveforms,

To name but a few, [9] proposes to estimate the unobserved amount of solar generations, which are modeled as a function of local global horizontal irradiance (GHI), from composite power flow measurements at the point of common coupling. However, GHI measurements are not widely available, which limits its applicability. Moreover, the correlation between monthly nocturnal and diurnal demand profiles and the similarity among solar generation profiles are exploited to disaggregate customer-level BTM solar using low-resolution but widely available hourly AMI readings [10], in which a maximum likelihood estimation based technique is proposed to utilize hourly typical solar exemplars based on constructed joint probability density functions of monthly nocturnal and diurnal demands. Similarly, [11] uses solar and demand examplers to find the optimal BTM solar generation based on a game-theoretic model. Finally, [12] proposes a mixed hidden Markov model to model general load consumption behaviors and estimate key parameters of BTM solar generations.

Besides aforementioned model-driven disaggregation techniques, recent advances in machine learning have also been extensively explored. Bayesian-dictionary-learning-based approaches are proposed in [13] to estimate the consumption of each load from aggregate measurements without knowing the ON/OFF status of each load. Moreover, [14] proposes an adaptive machine learning framework to transform available data (e.g., weather, location of PV, and net loads by smart meters or transformers) to estimate total (actual) load profiles and learn specific load patterns. Similarly, a disaggregation algorithm based on online training with multiple measurements include disparate active and reactive power flow measurements, complex bus voltage measurements, and residential smart meter measurements at feeder level is developed in [15].

To our best knowledge, this work is the first effort to investigate non-intrusive disaggregation of daily nodal load profiles at transmission nodes with significant BTM solar, which is yet to be addressed. Specifically, the objective of this work is to disaggregate the actual nodal load profiles based on the zonal load and proxy solar irradiance profiles, which differs from the existing literature in the following aspects.

- The state-of-the-art is generally designed for distribution networks and assumes the existence of actual BTM solar measurements for supervised-learning-type validation. However, such actual BTM solar or load profiles (i.e., the ground truth) generally do not exist at transmission nodes, which excludes the utilization of conventional supervised learning techniques. Therefore, the main technical challenge that has not addressed in the literature is: *without the actual nodal load profiles as the ground truth, how to evaluate the disaggregated profiles*?
- Although the RTO's zonal demand profiles are accurate, available solar profiles are only *proxy* (i.e., measured from a nearby location, not the considered location), which would impose nonlinear differences between these proxy solar profiles and the actual BTM solar generation [16].
- Nodal load profiles at transmission nodes are correlated in a spatio-temporal manner, which has not been addressed or utilized in the state-of-the-practice.
- Metered power outputs do not always linearly align with the proxy solar irradiance, even at the same location as illustrated by a de-identified solar farm (in PJM's service territory) in Fig. 2. These nonlinear errors have generally been ignored in the literature.

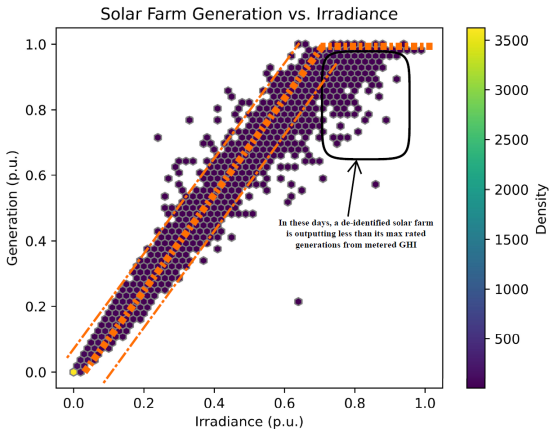


Fig. 2: A real-world example: GHI and solar generation output are not always aligned (data shown over one summer). Before reaching its max rated capacity, output of this de-identified solar farm generally follows a linear relationship with metered GHI. However, when GHI is sufficiently strong, solar output is often lower than its rated capacity (often due to partial equipment issues).

The main innovations and contributions are listed as follows:

1) To alleviate the lack of ground truth, we propose explicit but yet effective mechanisms in a *semi-supervised* manner. Based on observations from real-world data and industry best-practices, we divide daily nodal net injections into a nighttime portion (i.e., with no BTM solar) and a daytime portion with significant solar BTM. Consequently, we propose an ordinary least-squares (OLS) based disaggregator as the first benchmark, which does not depend on the metered GHI or solar power outputs. Instead, it only assumes similar patterns between actual nodal/zonal load profiles in daytime and nighttime, which is validated by real-world data.
2) To further reduce aforementioned errors caused by proxy solar and seasonal patterns, we enhance the OLS-based disaggregator by 1) first classifying daily profiles in an unsupervised manner by the self-organizing map (SOM) and then extending from temporal OLS to in-class OLS disaggregation; and 2) parameter tuning by leveraging the unique characteristics of sunrise and sunset times (which are pre-defined based on GHI thresholds determined from real-world data based on customized applications). We refer such enhanced mechanisms as

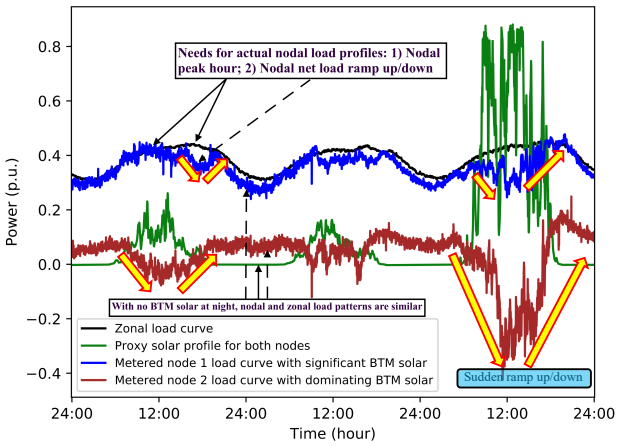


Fig. 3: Illustration of 1) minute-level, normalized, and anonymized PJM data [17] and 2) needs of actual nodal load profiles in daily RTO operations.

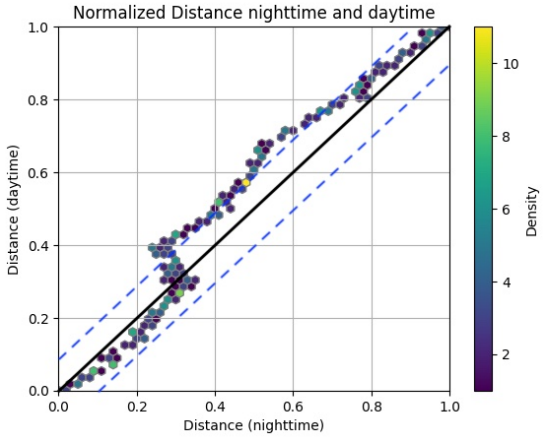


Fig. 4: Validation of the proposed affine relationship between daytime nodal/zonal and nighttime nodal/zonal actual load profile differences

.

Zone-to-Node (Z2N) methods.

3) To fill the gap in evaluating the performance of the proposed nodal profile disaggregator's performance without full ground truth, we utilize three distance functions based on spatial area difference, the Kullback–Leibler divergence, and the Kantorovich–Rubinstein metric to measure 1) the relative amount of daily solar and 2) the total difference between the disaggregated and metered nighttime nodal profiles. Furthermore, with the quantitative capability to evaluate disaggregation outcomes, we can further ensemble the three proposed disaggregation techniques by the voting mechanism.
4) Finally, we also investigated challenges to Z2N methods in cases where BTM dominates and the metered nodal load profile is negative and consequently propose a bi-level Node-to-Node (N2N) architecture to further enhance the efficiency and robustness of the proposed nodal load disaggregator.

The remaining of this paper is organized as follows. Section II formulates the proposed non-intrusive disaggregation problem for nodal load profiles and presents observations from real-world data that are aligned with industry best practice. Sections III proposes the proposed semi-supervised scheme with OLS disaggregation algorithms with three performance enhancement techniques: 1) parameter tuning using sunset/sunrise events; 2) unsupervised clustering based on SOM and consequently in-class OLS, and 3) a bi-level N2N architecture to alleviate errors when metered nodal load profile is negative. Moreover, Section V adopts three distance functions to evaluate the performance of the proposed disaggregation algorithms and discusses numerical validation results. Finally, Section VI draws conclusions and future work.

II. Problem Formulation

A. Problem Motivation, Definition, and Illustration

The example shown in Fig. 1 can be further illustrated in Fig. 3, which plots pre-processed, normalized, and anonymized PJM load profiles at two nearby transmission nodes (same substation) with similar, short distances to a nearby solar farm (i.e., FOM). A sample plot of the weekly data between July 1 and July 3, 2020 presents the total demand of an anonymized RTO zone (in black), a node with significant BTM solar (in blue), and another node with dominating BTM solar generation (in green). The reported solar irradiance profile at a nearby solar farm is also shown (in brown).

It can be observed that on both illustrated nodes, the actual nodal load profiles are unknown due to BTM solar. On the third day with strong solar irradiance, node 2 presented negative demands. Consequently, the actual peak demand hours at both nodes are unknown. Furthermore, high penetrations of BTM solar could cause unexpected, abrupt ramping up/down in metered nodal demands. Recent advancements in transmission grids operation and control by RTOs have investigated multi-scale issues in the existing energy/ancillary service co-optimization scheme with system-wide constraints and proposed granular solutions such as nodal reserves [18] and nodal reliability [19], for which actual nodal load and BTM solar profiles are necessary since nodal profiles in general do not follow system-wide profiles as shown above.

Moreover, since there generally do not exist telemetry or data sources for actual nodal load profiles, available data sources for estimating (i.e., disaggregating) actual nodal load profiles include only zonal demand profiles, proxy solar profiles, and other nearby nodes, which motivate the proposed techniques in this work.

It can be observed in Fig. 3 that, without BTM solar at nighttime, both nodes follow similar patterns as the zonal load profile in terms of their curve shapes (e.g., peak hours, ramping rates, and difference in peak p.u. values). In order to utilize nighttime data (i.e., no BTM solar and thus can be considered as the "ground truth" for verification purposes), an important hypothesis is that for each node, its actual load profile during daytime would statistically "align with" the learned pattern on how the nighttime bus load profiles follow the zonal load. To validate this assumption, Fig. 4 shows real-world data representing 493 selected, de-identified nodes within PJM service territory. The horizontal and vertical axes denote each transmission node's normalized, total daily difference (i.e., area differences) between nighttime nodal and

zonal load profiles as well as between daytime nodal and zonal load profiles, respectively. Note that the presented data (from 2018) in Fig. 4 was selected based on the following criteria:

1) There was no BTM solar on a specific day so that we can compare the differences between nodal and zonal load profiles at daytime and nighttime;
2) Data is recent as RTO load patterns change rapidly on a yearly basis with the ever-growing penetration of distributed energy resources, electrified transportation, and self-managed loads.

Consequently, it can be observed in Fig. 4 that the proposed assumption is general valid as most of the data points are bounded with in the range of $\pm 5\%$ error, which is widely acceptable in RTO operations.

B. Problem Formulation

Given a daily metered (aggregated) nodal load profile in the format of real power $\mathbf{x} = (x_1, \ldots, x_T)$, i.e., a timed sequence of T power consumption data points indexed by t, the problem to investigate in this paper is to determine its corresponding actual (disaggregated) nodal load profile $\mathbf{p} = (p_1, \ldots, p_T)$, i.e., "load disaggregation". The metered nodal load profiles are called aggregated as, at the considered transmission buses, each metered nodal load profile consists an unknown level of BTM solar generation $\mathbf{s}$. In other words, the metered nodal load profile $\mathbf{x}$ available to RTO can be considered as a sum of the BTM solar $\mathbf{s}$ and the actual nodal load $\mathbf{p}$, which needs to be "disaggregated", both of which are unknown, i.e.,

$$\begin{aligned} &\mathbf{x} = \mathbf{p} - \mathbf{s} \\ \text{or} \quad &x(t) = p(t) - s(t), \quad \forall t = 1, \ldots, T \end{aligned} \tag{1}$$

In this paper we follow the convention and assign positive values as net injections into the grid. In addition to the availability of real-world $\mathbf{x}$, to achieve feasible solutions, we further assume the availability of two extra sources of data: the zonal load profile $\mathbf{z} = (z_1, \ldots, z_T)$ (e.g., the total demand of an RTO's regional service territory or a specific RTO zone) and a proxy solar irradiance profile $\mathbf{r} = (r_1, \ldots, r_T)$ (e.g., metered solar irradiance at a nearby location to the transmission bus under consideration). Revisiting the illustrative example shown in Fig. 3, the green and blue, black, and red lines represent two different $\mathbf{x}$ curves at two transmission buses with significant and even dominating BTM solar, their corresponding zonal $\mathbf{z}$, and their proxy solar $\mathbf{r}$.

Considering that there does not exist any technique or methodology for load disaggregation at the transmission buses level, without the ground-truth, this paper adopts linearized formulations to represent the similarity between zonal and nodal load profiles as an affine relationship with a constant load component and an unknown stochastic difference that are small compared to the normalized load values, which has been justified in distribution networks [16], [20] and by RTO best practices. Formally, the following formulations are made:

1) The actual nodal load profile $\mathbf{p}$ follows the zonal load profile data $\mathbf{z}$ in an affine manner subject to a constant weight C_l, a constant load component p_c, and an unknown, stochastic load mismatch error ϵ, i.e.,

$$\begin{aligned} &\mathbf{p} = C_l \cdot \mathbf{z} + p_c + \epsilon \\ \text{or} \quad &p(t) = C_l z(t) + p_c + \epsilon(t), \quad \forall t = 1, \ldots, T \end{aligned} \tag{2}$$

2) Adjacent buses have similar solar irradiance profiles that are close to the proxy solar generation profile ϕ, and the BTM solar generation $\mathbf{s}$ is an affine function of the proxy solar generation profile ϕ subject to a constant C_s and an unknown, stochastic mismatch error ζ, i.e.,

$$\begin{aligned} &\mathbf{s} = C_s \cdot \phi + \zeta \\ \text{or} \quad &s(t) = C_s \phi(t) + \zeta(t), \quad \forall t = 1, \ldots, T \end{aligned} \tag{3}$$

Consider a single transmission bus, we would like to disaggregate its metered (aggregated) nodal load profile $x(t)$ into its actual nodal load profile $p(t)$ and BTM solar $s(t)$, i.e.,

$$\begin{aligned} &x(t) = p(t) - s(t), \\ \text{s.t.} \quad &p(t) = C_l z(t) + p_c + \epsilon(t) \\ &s(t) = C_s \phi(t) + \zeta(t) \end{aligned} \tag{4}$$

With the proposed assumption discussed in Section II.A and illustrated in Fig. 4, i.e.,

$$\|\mathbf{p}_{\text{zonal,nighttime}}, \mathbf{p}_{\text{nodal,nighttime}}\| \approx \|\mathbf{p}_{\text{zonal,daytime}}, \hat{\mathbf{p}}_{\text{nodal,daytime}}\| \tag{5}$$

where $\hat{\mathbf{p}}$ denotes disaggregated nodal load profiles from $\mathbf{x}$ and $\|\cdot, \cdot\|$ denotes any distance function. The proposed nodal load disaggregation problem can be formulated by minimizing the difference between nighttime metered and disaggregated nodal load profiles:

$$\arg\min \|\mathbf{p}_{\text{nodal,nighttime}}, \hat{\mathbf{p}}_{\text{nodal,nighttime}}\| \tag{6}$$

In this formulation, profiles $\mathbf{x}, \mathbf{z}$, and $\boldsymbol{\phi}$ are known, and thus the problem to solve can be considered as linearized regression. Note that the proposed formulation only depends on metered data such as zonal load profile and night-time nodal profile, while proxy solar profiles are only utilized to determine sunset and sunrise times to segment the daytime and nighttime halves. This is a major distinction compared to the existing literature. The learning process is carried out on the daytime halves with $\phi(t)$ and $z(t)$ are considered as known as well as best fitting parameters C_s, C_l, and p_c to be learned. Note that the two transposition errors ϵ and ζ can be combined as one without affecting the results [16].

III. Proposed Disaggregation Algorithms

A. Baseline: Ordinary Least-Squares

As discussed above, the lack of actual load profiles at transmission buses (i.e., without BTM solar) excludes the possibility of adopting existing supervised learning techniques. Based on the aforementioned observations made from real-world data and operational domain knowledge, this paper proposes to divide each daily load profiles into two parts:

1) Nighttime portion of the load profile: with no solar generations and thus not impacted by BTM solar, i.e., closely follow the zonal load pattern;
2) Daytime portion of the load profile: with significant solar BTM generations, the actual load profile (though

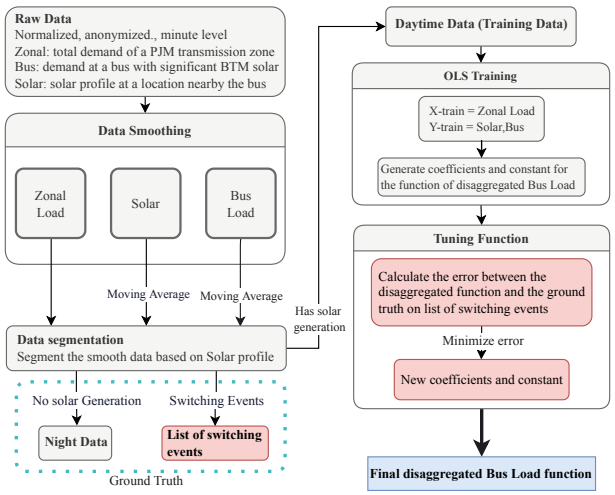


Fig. 5: Overview of the proposed OLS-based disaggregation.

unknown) would be statistically similar to the learned pattern on how the nighttime nodal load profiles follow the zonal load.

An overview of the proposed OLS-based disaggregation is shown in Fig. 5. Firstly, raw data is processed with simple moving average (SMA) applied to the $\mathbf{x}$ and its corresponding proxy solar ϕ, which could effectively reduce the impact of fluctuations in the raw data (especially in solar) on the accurate estimation of the parameters of large multivariate time series [21]. Secondly, prepared daily profiles are divided into day-time and night-time portions for training and validation, respectively. Finally, the multi-equation disaggregation model can be estimated equation-by-equation by ordinary least squares (OLS) [22], which is a type of linear least squares technique for estimating unknown parameters and provides minimum-variance, mean-unbiased estimation when the errors have the same finite variances but serially uncorrelated. Note that such assumptions are reasonable approximations in the proposed problem since the demand profiles on different nodes are uncorrelated but the solar irradiance profiles in a local region can be considered to be homoscedastic.

The proposed OLS-based architecture shown in Fig. 5 is considered to be semi-supervised [23], which generally refer to learning methods that are based on a small amount of labeled data with a large amount of unlabeled data during training, which align with the proposed problem in this paper. Specifically, the daytime profiles can be considered as "unlabeled" while the nighttime profiles can be considered as "labeled" as they are actual load profiles without BTM solar and utilized as (partial) ground truth. Compared to the conventional semi-supervised literature, the disaggregation problem considered in this paper is more balanced in terms of the ratio of labelled and unlabelled data, which would avoid many known issues caused by unbalanced data [24].

B. First Performance Enhancement via Parameter Tuning at Zero-crossing Points

Another important observation is that, though either the daytime or nighttime portions are complete in information (e.g., daytime portions lack ground-truth and nighttime portions lack BTM solar), two special data points that can be considered as possessing full information (i.e., both ground-truth and BTM solar since they are at the intersection of both) are the sunrise and sunset (i.e., zero-crossing) points, with sunrise time denotes the event from zero BTM solar at nighttime to positive BTM solar at daytime and the sunset time represents the event from positive BTM solar at daytime to zero BTM solar at nighttime. These two zero-crossing points can be considered as boundary conditions and are utilized in this work to further tune model parameters.

Specifically, for any consecutive N days, revisit Eqn. (2) and denote the disaggregated load profile of the n^{th} day by $p^n(t) = C_l^n z^n(t) + p_c^n + \epsilon^n(t), \forall t = 1, \ldots, T, n = 1, \ldots, N$ and the the sunrise and sunset times of the n^{th} day by tr^n and ts^n, respectively. To reduce the regression error, the proposed parameter tuning is implemented by introducing two auxiliary parameters a and b into $p^n(t)$ and consequently minimizing the sum of differences between $\mathbf{p}$ and $\mathbf{x}$ at sunset and sunrise times, i.e., $\forall t = 1, \ldots, T$ and $\forall n = 1, \ldots, N$,

$$p^n(t) = \quad (C_l^n + a) z^n(t) + p_c^n + \epsilon^n(t) + b \tag{7a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} p^n(\text{tr}^n) = \sum_{n=1}^{N} x^n(\text{tr}^n) \tag{7b}$$

$$\sum_{n=1}^{N} p^n(\text{ts}^n) = \sum_{n=1}^{N} x^n(\text{ts}^n) \tag{7c}$$

Note that parameters a and b are unknown parameters that can be uniquely calculated by two boundary conditions Eqns. (7b) and (7c) with respect to sunrise time $\texttt{tr}$ and sunset time $\texttt{ts}$ in a linear manner, with $\texttt{tr}$ and $\texttt{ts}$ determined based on thresholds calculated using proxy solar profiles, which are the cross-zero points of the proxy solar profile. Note that $\texttt{tr}$ and $\texttt{ts}$ can be considered as boundary conditions between daytime and nighttime halves, and thus difference between p and x at $\texttt{tr}$ and $\texttt{ts}$ are (theoretically) zero, which are interpreted in Eqns (7b) and (7c). Parameters a and b can then be solved in a straightforward manner from these two simultaneous equations without the need of utilizing recurrent techniques.

An illustrative example of performance enhancement with the proposed parameter tuning at sunrise and sunset times is shown in Fig. 6. It can be observed that due to the lack of ground truth at daytime, at highlighted time periods the disaggregated nodal load profiles (red dashed line) deviates from the metered (aggregated) nodal load profile (blue solid line) and almost coincide with the zonal load (black solid line), which is caused by the inherent characteristics of OLS. As a clear comparison, it can be observed that with the proposed parameter tuning at sunrise and sunset times, such deviation have been corrected and the disaggregated nodal load profile retains similar patterns as the zonal load and remain close to the metered (aggregated) nodal load profiles.

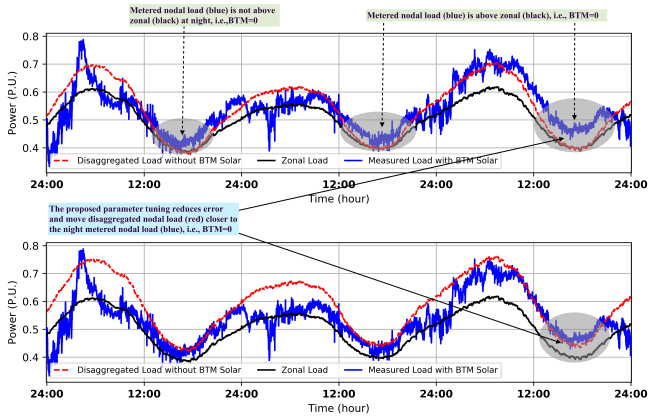


Fig. 6: Illustration of performance enhancement with the proposed parameter tuning at sunrise and sunset times.

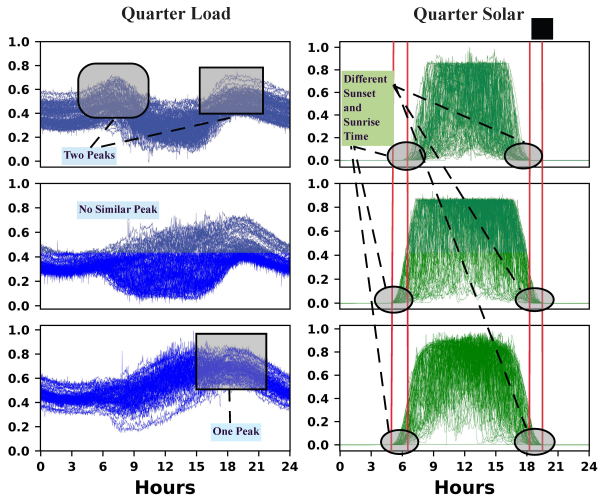


Fig. 7: Comparison of seasonal load and solar profiles: (Top) Q1 of 2020, (Middle) Q2 of 2020, and (Bottom) Q3 of 2020.

C. Second Performance Enhancement by Unsupervised Classification and In-class OLS

The proposed OLS-based architecture shown in Fig. 5 is based on the moving averaging pre-process and parameter regression utilizing data from consecutive days. However, a notable issue is that consecutive days at difference seasons might differ significantly in their daily solar irradiance and demand profiles, which in turn could inherently cause error to the consecutive-days-based OLS disaggregation. Therefore, we propose to first classify daily profiles into clusters in an unsupervised manner and then, for any day under consideration, perform OLS disaggregation using daily data from within its corresponding cluster. Note that the classification is unsupervised as there is no prior knowledge in how many clusters there are and/or what features to adopt. Consequently, the classification is purely inherently data-driven without any labels, i.e., unsupervised.

1) Feature Selection: In the existing literature of load modeling and/or forecasting, available data is typically split by different seasons, into weekdays and weekends, or into holidays and normal days. Most existing literature cluster data into consecutive days, and limited efforts have been devoted to cluster data on a daily base. Reference [25] adopts the daily bases clustering for solar prediction based on clustered GHI data, which is a single variable prediction problem utilizing only the GHI data.

This work performs unsupervised classification in three steps: 1) ensemble features from different aspects; 2) feature selection by computation; and 3) feature selection by performance. To start with features, daily zonal load and solar generation profiles of the first three quarters of 2020 are plotted in Fig. 7 to seek potential features from seasonal, quarterly, or monthly patterns in either the zonal load or solar irradiance. The following observations can be made:

- In Q2 and Q3, both sunrise and sunset occur at around the same time. Moreover, sunrise time in Q2 and Q3 is generally earlier than Q1, while the sunset time in Q2 and Q3 is generally later than Q1;
- The peak value of daily solar is similar in Q1 and Q2, which is lower than Q3's peak daily solar;
- In Q2 and Q3, the variance in daily solar profiles is generally smaller than Q1;
- In Q1, zonal loads have two peaks (morning and evening), which is consistent through out the season. Similarly, zonal load profiles in Q3 are also consistent with one evening peak. It can be observed that Q2 has combined features of Q1 and Q3, with a possible interpretation that Q2 is seasonal transition from Q1 to Q3.

To summarize, the observations made from seasonal load and solar patterns can be interpreted as two challenges: (1) diversity of daily profiles in the same season and (2) similarity among daily profiles from different seasons. Consequently, classification by only seasons is insufficient for the purpose of performance enhancement in the proposed OLS-based disaggregation algorithm. Instead, this work ensembles the following 14 features that are potentially suitable for clustering daily load and solar profiles by their inherent characteristics:

- Solar related features: (1) peak daily solar (in p.u.), (2) total daily solar generation (in p.u.), (3) sunrise time, (4) sunset time, (5) peak daily solar time, (6) solar generation at noon (in p.u.);
- Load related: (7) peak load (in p.u.), (8) minimum load (in p.u.), (9) total daily load (in p.u.), (10) peak load time, (11) minimum load time, (12) load at noon (in p.u.);
- Crest factors: (13) load crest factor (peak load/minimum load), (14) efficiency crest factor (peak load/peak solar).

In order to quantitatively determine an optimal subset of the proposed features without bias, a linear-time feature selection algorithm called `Relief` [26] is adopted to estimate each feature's contributing proportion and reduce the number of features by selecting only statistically relevant features. Specifically, `Relief` is a filter-method feature selection approach that is widely used as the benchmark to score and rank (the weights of) individual features. The contributing proportions of the proposed 14 features are listed in Table I. It can be observed that the sunrise and sunset times present significantly higher proportions than other features, followed by the maximum and minimum load times, which aligns with the

TABLE I: Feature selection by seasonal patterns

Feature	Proportion
Peak daily solar (in p.u.)	0.0466
Total daily solar generation (in p.u.)	0.0453
Peak daily load (in p.u.)	0.0530
Minimum daily load (in p.u.)	0.0418
Total daily load (in p.u.)	0.0590
Load crest factor (peak load/minimum load)	0.0185
Efficiency factor (peak load/peak solar)	0.0315
Sunrise time	0.2423
Sunset time	0.2044
Peak daily load time	0.0724
Minimum daily load time	0.0916
Peak daily solar time	0.0214
Solar generation at noon (in p.u.)	0.0313
Load at noon (in p.u.)	0.0404

characteristics of the seasonal label as the zonal demand is closed related to activities caused by sunrise and sunset events.

2) Unsupervised Classification by the Self-organizing Map: As discussed above, statistically selecting features by their relevance to seasons (i.e., in a supervised manner) aligns with seasonal inherent characteristics but not necessarily addresses neither the diversity within each season nor the similarity among different seasons as shown in Fig. 7. Therefore, in this work we propose to utilize unsupervised clustering of the available daily zonal load and solar profiles with respect to different subgroups of the proposed features. Specifically, we propose to adopt the time adaptive extension of the self-organizing map (SOM) [27], also known as the Kohonen map. The SOM is an unsupervised artificial neural network (ANN) trained by competitive learning (rather than error-correction learning methods such as gradient descent). Trained with high-dimensional data, the output of an SOM is a low-dimensional (typically two-dimensional), discretized grid of neurons. Each nueron is assigned a feature vector of the same dimension as the input data. The advantages of the SOM are three-fold:

- Clustering by nature: an SOM does not require *a priori* information on the number of clusters; instead, an SOM represents as clusters with respect to the relative distances among feature vectors. In other words, input feature vectors in proximal clusters have more similar values than feature vectors in distal clusters;
- Dimension reduction: the training process of an SOM is to represent a high-dimensional input feature space by a low-dimensional mapped space;
- Outcome visualization: a trained SOM could visualize boundaries among clusters to illustrate variances within each cluster and distances among different clusters.

Necessary concepts are introduced as follows [7].

2.a) Neurons and Their Assigned Values: Consider an SOM with K neurons with l-dimensional training data consists of feature vectors $\underline{\boldsymbol{x}}_q = [x_{q1}, x_{q2}, x_{q3}, \ldots, x_{ql}]$. Each neuron n_i is assigned with [28] : 1) a time-invariant topological position (i.e., an $x-y$ coordinate in the 2-D output grid); 2) a time-varying parametric weight (also called reference, model, or codebook) vector $\underline{\boldsymbol{m}}_i = [m_{i1}, m_{i2}, m_{i3}, \ldots, m_{il}]$ of the same dimension as the input data; 3) a predefined function which defines a neighborhood (e.g., a circle or a square in 2-D) centered at the neuron. All neurons compete to respond to the input data but only one neuron wins at each time. Denoted by the subscript c, the winning neuron is called the Best Matching Unit (BMU):

$$c = \arg\min_i \left\{ \left\| \underline{\boldsymbol{x}}_q - \underline{\boldsymbol{m}}_i \right\| \right\} \tag{8}$$

where $\|\cdot\|$ is a distance function, typically the Euclidean

$$\mathrm{d}_E\left(\underline{\boldsymbol{m}}_i, \underline{\boldsymbol{x}}_q\right) = \sqrt{\sum_{k=1}^{l} \left(m_{ik} - x_{qk}\right)^2} \tag{9}$$

2.b) The Batch SOM Training Algorithm There are two types of training algorithms for the SOM: sequential and batch. Sequential training takes one single input vector of data at each training step and then updates, and the batch training presents all input data to the neuron grid before any updates are made at each training step. The batch SOM training algorithm proceeds as follows [28]:

1. Initialize weight vectors $\underline{\boldsymbol{m}}_i$;
2. Partition the input data set into the Voronoi regions of the weight vectors, i.e., each input vector $\underline{\boldsymbol{x}}_q$ belongs to the region of its closest neuron n_i;
3. Update $\underline{\boldsymbol{m}}_i$ according to

$$\underline{\boldsymbol{m}}_i(t+1) = \frac{\sum_{q=1}^{N} h_{q,c}(t)\underline{\boldsymbol{x}}_q}{\sum_{q=1}^{K} h_{q,c}(t)} \tag{10}$$

   where c is the BMU of the input vector $\underline{x}_q$, $h_{q,c}(t)$ is the neighborhood function, and K is the number of neurons;
4. Return to step 2 and repeat until stopping criteria is met.

As a result, feature vectors with similar values are mapped to neurons positioned close to one another and form a cluster.

Numerical comparison has been conducted using many different combinations of features to train the SOM, and three examples are illustrated in Fig. 8. It can be observed that

- Clustering performance differs significantly among different feature groups;
- The middle SOM has the best performance in terms of relatively balanced numbers of points in each of the four clusters as well as clear boundaries among clusters, for which 4 selected features include peak daily solar, total daily solar, peak daily load, and minimum daily load;
- Cluster 0 contains the most number of input feature vectors, while cluster 3 contains the least;

Moreover, the values of total amount of daily solar irradiance descends from clusters 0 to 3, which indicates that cluster 0 has the strongest solar generation profile (while cluster 3 has the weakest), which corresponds to features mostly from Q2 and Q3 and aligns with seasonal patterns.

3) In-class OLS Disaggregation: The proposed feature selection process and SOM-based classification can be integrated into the semi-supervised, OLS-based disaggregation algorithm represented in Fig. 5. The overall procedure is shown in Fig. 9. Note that another important factor to address is how to quantitatively evaluate the performance of disaggregation outcomes without the ground truth, which will be presented in Section IV.

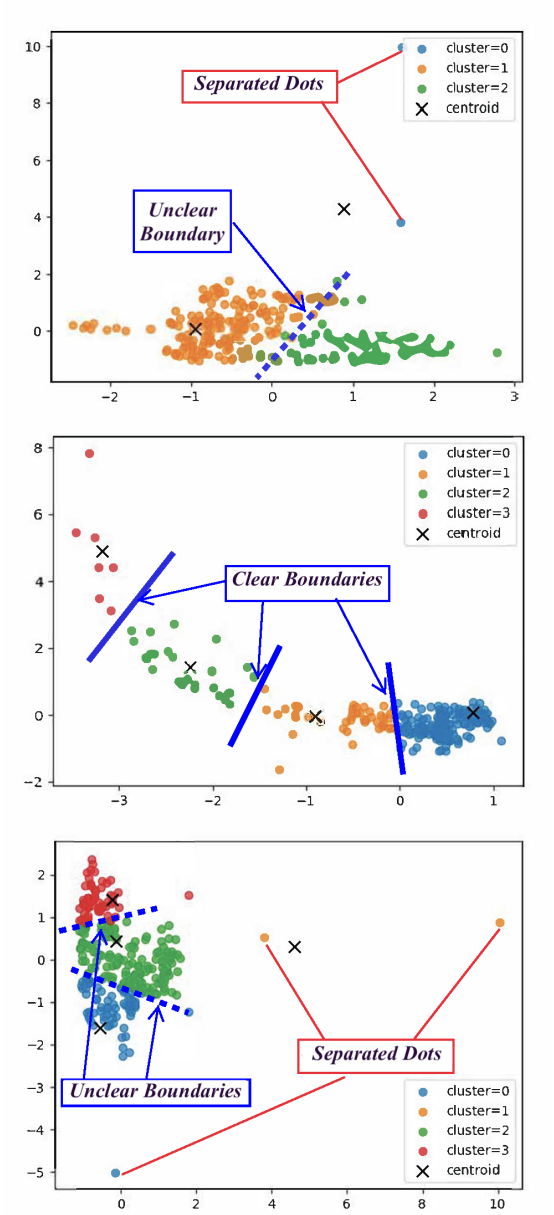


Fig. 8: Three examples of SOMs trained with different groups of features.

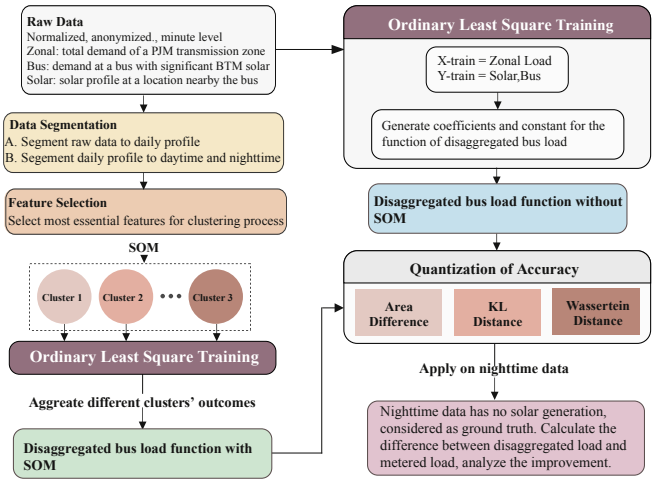


Fig. 9: Overview of the disaggregation procedure utilizing feature selection and unsupervised classification.

D. Third Performance Enhancement by Bi-level Node-to-Node Formulations

The Z2N disaggregation techniques presented above can be further extended to an N2N formulation, especially for cases in which metered (aggregated) nodal load profiles are negative due to dominating BTM solar. Specifically, the proposed N2N architecture can be considered as bi-level formulation.

First level: picking one bus without dominating BTM solar as a *reference bus*, whose metered (aggregated) nodal load profile $\mathbf{x}_{\text{bus}_{\text{ref}}}$ can be disaggregated into its corresponding actual nodal load $\mathbf{p}_{\text{bus}_{\text{ref}}}$ and its proxy BTM solar $\mathbf{s}_{\text{bus}_{\text{ref}}}$ by the proposed Z2N disaggregation techniques presented above;

Second level: aiming at disaggregating other N buses in the same region by applying node-to-node correlations between individual target buses and the selected reference bus. Specifically, the relation between the reference bus and the n^{th} bus can be formulated as

$$\mathbf{x}_{\text{bus}_n} = C_R \mathbf{x}_{\text{bus}_{\text{ref}}} + p_R \tag{11}$$

where coefficients C_R and p_R can be learned by nighttime portions (i.e., sunset to sunrise):

$$\min \quad \sigma_{\text{MSE}} := \sum_{t=T_{\text{ts}}}^{T_{\text{tr}}} (\mathbf{p}_{\text{bus}_n} - \mathbf{x}_{\text{bus}_n})^2. \tag{12}$$

Consequently, for daytime (i.e., sunrise to sunset), target node$_n$'s disaggregated nodal load profile follows the same relation:

$$\mathbf{p}_{\text{bus}_n} = C_R \mathbf{p}_{\text{bus}_{\text{ref}}} + p_R. \tag{13}$$

The proposed bi-level N2N architecture can be summarized by Algorithm 1. Moreover,

- In the first level (i.e., Z2N), daytime data is used for training, while the second layer uses nighttime data for training to obtain the N2N correlations;
- Unlike the first level, only the metered (aggregated) nodal load profiles are needed in the second level, i.e., proxy solar profile is not needed in the second level.

E. Quantile Regression for Comparison

For effective comparison of performance enhancements by the proposed bi-level framework, another technique called *quantile regression* (QR) [29] is also adopted to be compared with OLS. It is widely acknowledged that in some cases QR is more robust than OLS as OLS minimizes the conditional mean while QR estimates the conditional quantiles (i.e., median). The error of the QR model is represented by $\mathbf{y} = f(\mathbf{x}, \beta_{\text{QR}}) + \epsilon$, where β_{QR} is estimated by:

$$\begin{aligned} \hat{\beta}_{\text{QR}} = \arg\min_{b} & \sum_{i:y_i \geq f(\mathbf{b},x_i)}^{n} q\,|y_i - f(\mathbf{b}, x_i)| \\ & + \sum_{i:y_i < f(\mathbf{b},x_i)}^{n} (1-q)\,|y_i - f(\mathbf{b}, x_i)| \end{aligned} \tag{14}$$

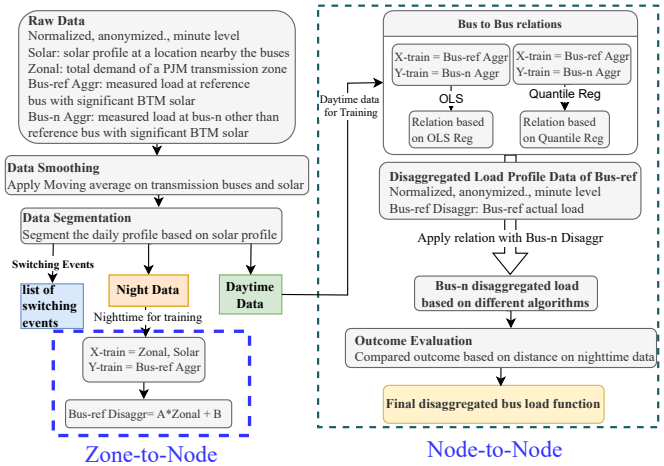


Fig. 10: Overview of the bi-level procedure utilizing Z2N and N2N for any transmission bus in the same region

In our model, we choose $q = 0.5$ to estimate the conditional median of the disaggregated nodal load profile at the target bus and simplify QR as

$$\hat{\beta}_{\mathrm{QR},q=0.5} = \arg\min_{b} \sum_{i=1}^{n} 0.5 \left| y_i - f\left(\mathbf{b}, x_i\right) \right|, \qquad (15)$$

which is equivalent to a Least Absolute Deviation (LAD) formulation [30]:

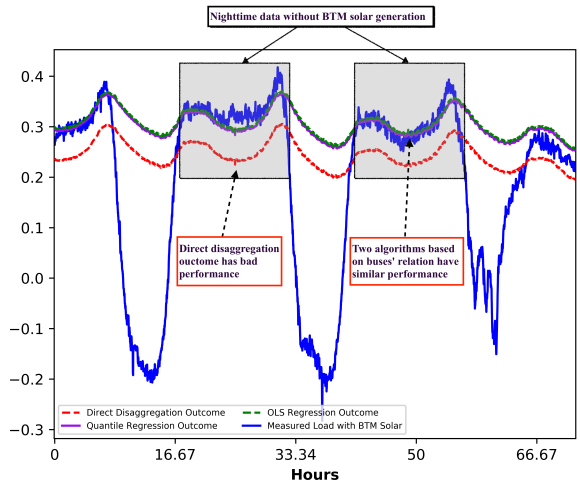
$$\hat{\beta}_{\mathrm{LAD}} = \arg\min_{b} \sum_{i=1}^{n} \left| y_i - f\left(\mathbf{b}, x_i\right) \right|. \qquad (16)$$

The proposed performance enhancement by the prposed bi-level N2N can be illustrated in Fig. 11, in which 2 illustrative examples are shown to represent different seasons and solar patterns. Each plot includes three-day's profiles, and it can be observed that the proposed Z2N OLS disaggregator does not perform sufficiently well when the metered (aggregated) nodal load profiles turn negative, while the proposed bi-level N2N can significantly enhance the disaggregation performance.

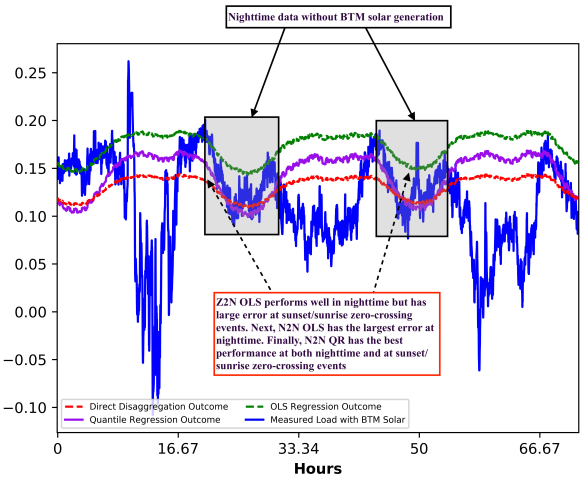
IV. Quantitative Evaluation of Outcomes

In the proposed semi-supervised architectures, the disaggregated nodal load profiles at night-time match their corresponding nodal injection profiles since there is no BTM solar at night, which could evaluate the performance of the proposed disaggregation algorithms without ground truth. In addition, this work proposes a novel comparative evaluation technique. Specifically, the ground truth can be considered as scenarios with zero BTM solar, and thus the main idea here is to first quantify the total daily solar generation and then calculate the difference between the disaggregated and observed load profiles. Consequently, the proposed disaggregation techniques are considered effective if days with less quantified daily solar implies smaller differences. For instance, for any two days A and B,

$$\text{(daily solar A} \leq \text{daily solar B)} \quad \Rightarrow$$
$$\text{(disaggregation difference A} \leq \text{disaggregation difference B)}.$$



(a) Jan 21-23, 2020



(b) May 19-21, 2020

Fig. 11: Performance enhancement by the proposed bi-level N2N architecture (compared to Z2N OLS) when metered (aggregated) nodal load profiles are negative. (Top) Jan 21-23, 2020 and (Bottom) May 19-21, 2020.

In this work, we adopt the area-based spatial differences between two time series of equal length [31], which is the accumulation of point-to-point Euclidean distances and an intuitive representation of similarity between two equal-length time series. The smaller the area difference is, the higher the similarity, vice versa. Specifically, the following two distance functions are also adopted.

A. Kullback–Leibler Distance

In probability theory and information theory, the Kullback–Leibler (KL) divergence is a statistical measure of the difference between a discrete (actual) probability distribution P from its reference probability distribution Q. In applications, P typically presents the observed data with Q being its modeled values. Formally, for discrete probability distributions P and Q defined on the same probability space $\mathcal{Y}$, the KL divergence from Q to P is defined by

$$D_{\mathrm{KL}}(P\|Q) = \sum_{y\in\mathcal{Y}} P(y) \log\left(\frac{P(y)}{Q(y)}\right), \qquad (17)$$

Algorithm 1: Bilevel Node-to-Node Disaggregation

Input $\mathbf{s}, \mathbf{z}, \lambda, \mathbf{x}_{\text{bus}_n}, n = 1, \ldots, N$
// Applying Moving Average
$x_{\text{MA}}(t) \leftarrow \frac{1}{\lambda}(x_{\text{bus}_n}(t) + \ldots + x_{\text{bus}_n}(t+\lambda-1))$
$s_{\text{MA}}(t) \leftarrow \frac{1}{\lambda}(s(t) + \ldots + s(t+\lambda-1))$
for $t = 0 : T-1$ **do**
  $x_{\text{MA}}(t+1) \leftarrow x_{\text{MA}}(t) - x_{\text{bus}_n}(t) + x_{\text{bus}_n}(t+\lambda)$
  $s_{\text{MA}}(t+1) \leftarrow s_{\text{MA}}(t) - s(t) + s(t+\lambda)$
  $t \leftarrow t+1$
end
// Z2N OLS at reference bus
Input $\mathbf{x}_{\text{ref}}(T_{\text{sunrise}} : T_{\text{sunset}}), \mathbf{s}, \mathbf{z}$
Calculate $C_l, p_c \leftarrow$ Min. daytime error
Update $C_l, p_c \leftarrow$ Min. switching points error
Output $\mathbf{p}_{\text{ref}} = C_l \mathbf{z} + p_c$
// N2N disaggregation w/ correlation
for $n = 1, \ldots, N, n \neq$ *ref* **do**
  Input $\mathbf{x}_{\text{ref}}(T_{\text{sunset}} : T_{\text{sunrise}}), \mathbf{x}_n(T_{\text{sunset}} : T_{\text{sunrise}})$
  $C_{R,\text{MSE}}, p_{R,\text{MSE}}, \sigma_{\text{MSE}} \leftarrow$ Min. nighttime MSE
  $C_{R,\text{LAE}}, p_{R,\text{LAE}}, \sigma_{\text{LAE}} \leftarrow$ Min. nighttime LAE
  if $\sigma_{MSE} \geq \sigma_{LAE}$ **then**
    $C_R, p_R \leftarrow C_{R,\text{LAE}}, p_{R,\text{LAE}}$;
  else
    $C_R, p_R \leftarrow C_{R,\text{MSE}}, p_{R,\text{MSE}}$;
  end
  $\mathbf{p}_n = C_R \mathbf{p}_{\text{ref}} + p_R$
end

which is the expected (over P) logarithmic difference between P and Q.

Note that although the KL distance is originally defined over two probability distribution, recent works have justified utilizing a symmetrical KL divergence for evaluating the similarity between two time series sequences [32], which is also adopted in this paper as

$$\begin{aligned} D_{\text{KLsym}} = (&D\left(\mathbf{p}_{\text{nodal,nighttime}} \| \hat{\mathbf{p}}_{\text{nodal,nighttime}}\right) \\ &+ D\left(\hat{\mathbf{p}}_{\text{nodal,nighttime}} \| \mathbf{p}_{\text{nodal,nighttime}}\right))/2 \end{aligned} \tag{18}$$

where $\hat{\mathbf{p}}$ and $\mathbf{p}$ denote the disaggregated and metered nodal load profiles, respectively.

B. Wassertein Distance

The other distance function we propose to adopt for measuring the difference between two time series is the Wasserstein distance, which is also originally designed as a distance metric between two probability distributions and has been widely used in calculating ambiguity sets in power system applications [33], [34]. Specifically, the Wasserstein metric $D_W(P,Q) : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}$ between two distributions P and Q is defined as

$$D_W(P,Q) = \inf \left\{ \int_{\Xi \times \Xi} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \, \Pi(d\boldsymbol{w}_1, d\boldsymbol{w}_2) \right\}, \tag{19}$$

where $\mathcal{M}(\Xi)$ denotes the set of all probability distributions with support Ξ, Π is a joint distribution of (the two integral variables) $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ with marginal distributions P and Q, respectively, and $\|\cdot\|$ is a norm. Note that $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$ is the cost of moving a unit mass from $\boldsymbol{w}_1$ to $\boldsymbol{w}_2$ as defined in the optimal transport problem. Intuitively, the joint distribution Π can be viewed as a plan to transport probability mass from P to Q. Consequently, $D_W(P,Q)$ returns the lowest cost of transporting probability mass from P to Q so that $P = Q$.

In our case, each metered (aggregated) nodal load profile and its corresponding actual (disaggreagted) nodal load profile are considered as two samples of the same size and also with the same with support, which aligns with aforementioned observations and industry best practices that statistical characteristics of each nodal demand can be considered as stationary.

C. Numerical Results

Numerical results using aforementioned PJM data throughout the year of 2020 are listed in Table II. Moreover, to further show details of the performance enhancement by the proposed unsupervised clustering, in-class OLS, and parameter tuning, 4 cases are further shown in Fig 12, in which key periods with significant performance enhancement are amplified with details and compared with comments.

Specifically, in each case presented in Table II, we compare the disaggregation performances by 1) the baseline OLS-based algorithm and 2) the in-class OLS-based algorithm with parameter tuning. As discussed in Section V, to quantitatively measure disaggregation accuracy, the three distance metrics proposed in Section V are calculated and compared for all cases. It can be observed that

- In most cases, the proposed enhancements by unsupervised learning and by parameter tuning significantly improve the accuracy of the proposed OLS-based disaggregation algorithm in terms of (1) reducing differences between nighttime metered and actual (disaggregated) nodal load profiles and more importantly (b) correcting the disaggregated load profiles in daytime and keeping the them aligned with the metered (aggregated) daytime nodal load profiles, as shown in Fig. 12;
- In some cases, the baseline OLS could perform better than the in-class OLS due to two potential causes: (1) for days with low solar irradiance, the baseline OLS already perform very well with low differences, i.e., case (g); and (2) an untypical day (e.g., highly volatile solar) that disaggregation by linear regression has a high difference could cause undesirable errors in clustering.

Both cases could be improved with more data.

V. Conclusion

This paper introduces a cluster-based unsupervised structure to disaggregate actual nodal load profiles from metered (aggregated) nodal load profiles with BTM solar. To overcome the lack of "ground truth" and validate the performance of the proposed algorithms, we first proposed a semi-supervised scheme by dividing each daily profile into two portions: daytime (i.e., with BTM solar) and nighttime (i.e., without BTM solar) and developed a baseline OLS disaggregation algorithm. Three performance enhancement techniques have been later introduced: 1) parameter tuning using sunset/sunrise events;

TABLE II: Comparison of numerical results on selected weeks throughout the year of 2020

Scenarios in Fig. 12		Area Diff.	KL_{SYM} Distance	Wassertein Distance	Cluster	Area Diff. Improvement	KL_{SYM} Improvement	Wassertein Improvement	Month
(a)	Baseline OLS	61.47	51.37	0.025	2	14.59%	16.67%	12%	Jan
	In-class OLS + tuning	52.50	42.95	0.022					
(b)	Baseline OLS	166.38	136.40	0.067	2	5.11%	5.54%	4.47%	Jan
	In-class OLS + tuning	157.84	128.84	0.064					
(c)	Baseline OLS	113.75	95.49	0.045	1	4.41%	5.60%	4.44%	Jan
	In-class OLS + tuning	108.73	90.14	0.043					
(d)	Baseline OLS	123.75	91.97	0.088	0	18.42%	19.85%	18.18%	Mar
	In-class OLS + tuning	100.95	73.71	0.072					
(e)	Baseline OLS	28.15	19.52	0.023	0	21.92%	24.49%	21.73%	Apr
	In-class OLS + tuning	21.98	14.74	0.018					
(f)	Baseline OLS	35.09	29.08	0.023	2	13.25%	10.38%	13.04%	May
	In-class OLS + tuning	30.44	26.06	0.020					
(g)	Baseline OLS	10.65	8.92	0.08	2	-168.83%	-185.54%	162.50%	Jun
	In-class OLS + tuning	28.63	25.47	0.021					
(h)	Baseline OLS	37.76	32.61	0.025	2	24.18%	21.89%	16.00%	Sep
	In-class OLS + tuning	28.63	25.47	0.021					

2) unsupervised clustering based on SOM and consequently in-class OLS, and 3) a bi-level N2N architecture to alleviate errors when metered nodal load profile is negative. Moreover, three distance metrics to evaluate differences between time series data are adopted for quantitative performance analysis, including the area difference, the KL Distance, and the Wassertein distance. Numerical validations using real-world PJM data justified the applicability of the proposed methods.

Furthermore, all proposed techniques are linear and thus are computationally effective. The proposed methods are also practical and feasible since RTOs have the granularity of solar profiles (e.g., PJM have real-time solar forecast for every five miles and every five minutes), and the proposed nodal load disaggregator can be integrated into RTOs's EMS as an additional tool. Finally, for future work, the proposed nodal load profile disaggregation problem will be extended to be formulated by nonlinear techniques, such as machine learning models. Furthermore, the proposed nodal load profile disaggregation also enables the possibility of studying nodal reserves at RTOs.

REFERENCES

[1] North American Electric Reliability Corporation (NERC), "San Fernando Disturbance - Southern California Event: July 7, 2020 Joint NERC and WECC Staff Report," November 2020. [Online]. Available: https://www.nerc.com/pa/rrm/ea/Documents/San_Fernando_Disturbance_Report.pdf

[2] ——, "Odessa Disturbance - Texas Events: May 9, 2021 and June 26, 2021 Joint NERC and Texas RE Staff Report," September 2021. [Online]. Available: https://www.nerc.com/pa/rrm/ea/Documents/Odessa_Disturbance_Report.pdf

[3] J. Rand *et al.*, "Queued up: Characteristics of power plants seeking transmission interconnection as of the end of 2022 [slides]," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA, Tech. Rep., 2023.

[4] C. E. Tidemann, J. M. Bright, and N. A. Engerer, "Solar forecasting as an enablement tool for the distribution system operator (dso)," in *IEEE 46th Photovoltaic Specialists Conference*, 2019, pp. 1637–1644.

[5] D. Moscovitz, "Need for BTM Data: Transmission Outage Analysis," in *The 9th Annual Monitoring and Situational Awareness Technical Conference*. NERC, October 2021. [Online]. Available: https://www.nerc.com/pa/rrm/Resources/Documents/2021_MSA_Conference_Session2.pdf

[6] Y. Du, L. Du, B. Lu, R. Harley, and T. Habetler, "A review of identification and monitoring methods for electric loads in commercial and residential buildings," in *2010 IEEE Energy Conversion Congress and Exposition*. IEEE, 2010, pp. 4527–4533.

[7] L. Du, J. A. Restrepo, Y. Yang, R. G. Harley, and T. G. Habetler, "Nonintrusive, self-organizing, and probabilistic classification and identification of plugged-in electric loads," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1371–1380, 2013.

[8] S. Wang, L. Du, J. Ye, and D. Zhao, "A deep generative model for non-intrusive identification of ev charging profiles," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4916–4927, 2020.

[9] F. Sossan, L. Nespoli, V. Medici, and M. Paolone, "Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers," *IEEE Trans. Ind. Inform.*, vol. 14, no. 9, pp. 3904–3913, 2018.

[10] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Guo, "Disaggregating customer-level behind-the-meter pv generation using smart meter data and solar exemplars," *IEEE Trans. Power Systems*, vol. 36, no. 6, pp. 5417–5427, 2021.

[11] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Zhang, "A data-driven game-theoretic approach for behind-the-meter pv generation disaggregation," *IEEE Trans. Power Systems*, vol. 35, no. 4, pp. 3133–3144, 2020.

[12] F. Kabir *et al.*, "Joint estimation of behind-the-meter solar in a community," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 682–694, 2020.

[13] M. Yi and M. Wang, "Bayesian energy disaggregation at substations with uncertainty modeling," *IEEE Trans. Power Systems*, vol. 37, no. 1, pp. 764–775, 2022.

[14] R. Saeedi, S. K. Sadanandan, A. K. Srivastava, K. L. Davies, and A. H. Gebremedhin, "An adaptive machine learning framework for behind-the-meter load/pv disaggregation," *IEEE Trans. Industrial Informatics*, vol. 17, no. 10, pp. 7060–7069, 2021.

[15] G. Ledva and J. Mathieu, "Separating feeder demand into components using substation, feeder, and smart meter measurements," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3280–3290, 2020.

[16] E. C. Kara *et al.*, "Disaggregating solar generation from feeder-level measurements," *Sustainable Energy, Grids and Networks*, vol. 13, pp. 112–121, 2018.

[17] X. Fan, D. Moscovitz, L. Du, and W. Saad, "A data-driven democratized control architecture for regional transmission operators," in *IEEE PES Innovative Smart Grid Technologies Conf. (ISGT)*, 2022.

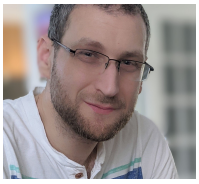
[18] F. Wang, A. Korad, and Y. Chen, "Reserve deliverability with application of short-term reserve product," in *FERC Technical Conf on Increasing Real-Time and Day-Ahead Market Efficienty through Improved Software*, 2019, pp. 2020–09.

[19] B. Lami and K. Bhattacharya, "Clustering technique applied to nodal reliability indices for optimal planning of energy resources," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4679–4690, 2016.

[20] M. Tabone, S. Kiliccote, and E. C. Kara, "Disaggregating solar generation behind individual meters in real time," in *Proceedings of the 5th Conference on Systems for Built Environments*, 2018, pp. 43–52.

[21] H. Spliid, "A fast estimation method for the vector autoregressive mov-

ing average model with exogenous variables," *Journal of the American Statistical Association*, vol. 78, no. 384, pp. 843–849, 1983.
[22] A. E. Clements, A. Hurn, and Z. Li, "Forecasting day-ahead electricity load using a multiple equation time series approach," *European Journal of Operational Research*, vol. 251, no. 2, pp. 522–530, 2016.
[23] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.
[24] S. R. Searle, *Linear models for unbalanced data*. John Wiley & Sons, 2006, vol. 639.
[25] C. Feng, M. Cui, B.-M. Hodge, S. Lu, H. F. Hamann, and J. Zhang, "Unsupervised clustering-based short-term solar forecasting," *IEEE Trans. Sustainable Energy*, vol. 10, no. 4, pp. 2174–2185, 2018.
[26] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.
[27] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
[28] L. Du, D. He, Y. Yang, J. Restrepo, B. Lu, R. Harley, and T. Habetler, "Self-organizing classification and identification of miscellaneous electric loads," in *IEEE Power & Energy Society General Meeting*, 2012.
[29] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
[30] P. Bloomfield and W. Steiger, "Least absolute deviations curve-fitting," *SIAM Journal on scientific and statistical computing*, vol. 1, no. 2, pp. 290–301, 1980.
[31] X. Wang, F. Yu, and W. Pedrycz, "An area-based shape distance measure of time series," *Applied Soft Computing*, vol. 48, pp. 650–659, 2016.
[32] D. García-García, E. Parrado Hernández, and F. Díaz-de María, "A new distance measure for model-based sequence clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1325–1331, 2009.
[33] A. Zhou, M. Yang, M. Wang, and Y. Zhang, "A linear programming approximation of distributionally robust chance-constrained dispatch with wasserstein distance," *IEEE Trans. Power Systems*, vol. 35, no. 5, pp. 3366–3377, 2020.
[34] X. Zheng and H. Chen, "Data-driven distributionally robust unit commitment with wasserstein metric: Tractable formulation and efficient solution method," *IEEE Trans. Power Systems*, vol. 35, no. 6, pp. 4940–4943, 2020.

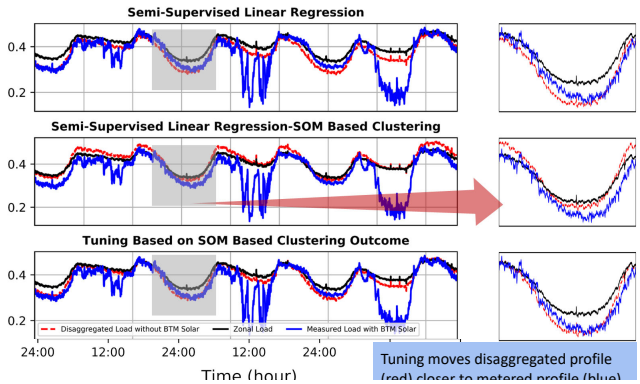


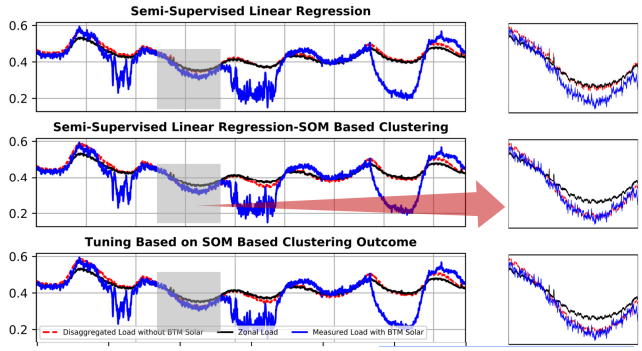
Daniel Moscovitz has worked for 20 years at PJM Interconnection with a focus on Transmission Network Applications and Energy Management System Support. He has earned a Bachelor of Science in Electrical Engineering from Bucknell University, Lewisburg, PA, a Master of of Science in Electrical from Drexel University, Philadelphia, PA, and is currently a Ph.D. candidate at Temple University, Philadelphia, PA. His research focuses on bulk electric grid optimization in the renewable energy supply driven future through improvements in behind the meter solar detection and transmission constrained economic dispatch of energy and reserves.

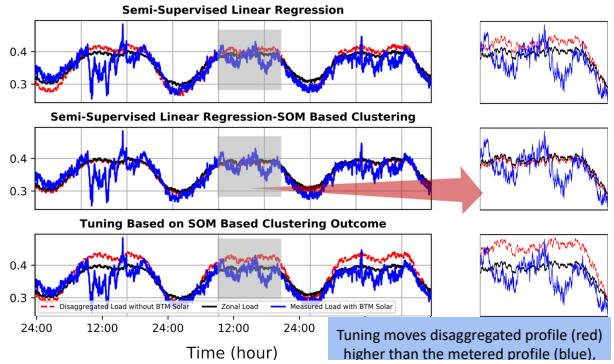
Zhenyu Zhao (S'23) received his B.Eng degree in Automation from Wuhan University of Technology, China in 2018 and his M.S. degree in Electrical Engineering from the George Washington University, Washington, DC in 2020. He is currently pursuing his Ph.D. degree at Temple University, Philadelphia, PA. He is an intern at PJM Interconnection (Audubon, PA) since summer 2023. His research interests include advanced data analytical and learning methods with applications to transmission systems.

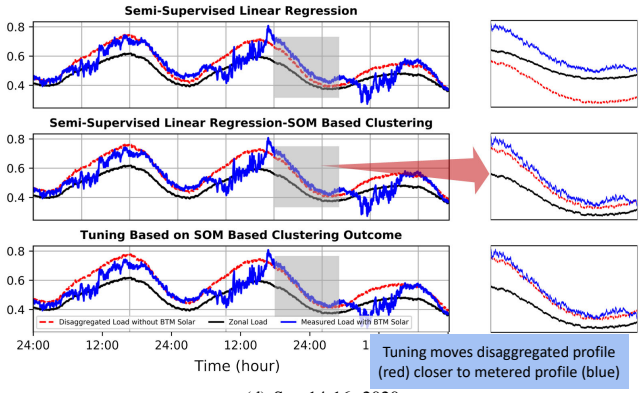
Liang Du (S'09–M'13–SM'18) received the Ph.D. degree in electrical engineering from Georgia Institute of Technology, Atlanta, GA in 2013. He was a Research Intern at Eaton Corp. Innovation Center, Mitsubishi Electric Research Labs, and Philips Research N.A. in 2011, 2012, and 2013, respectively. He was an Electrical Engineer with Schlumberger, Sugar Land, TX, from 2013 to 2017. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering at Temple University, Philadelphia. Dr. Du received the Ralph E. Powe Junior Faculty Enhancement Award from ORAU in 2018, Early-Career Fellowship from National Academies of Science in 2022, and CAREER award from National Science Foundation in 2023. He currently serve as an associate editor for IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, and IEEE TRANSACTIONS ON TRANSPORTATION ELECTRIFICATION.

Xiaoyuan Fan (S'12–M'16-SM'19) received the Ph.D. degree in electrical engineering from the University of Wyoming, Laramie, WY, in 2016, and M.S. and B.S. degrees in electrical engineering from Huazhong University of Sciences & Technology, Wuhan, China, in 2012 and 2009, respectively. He is currently a Senior Staff Engineer and Power Electronics Team Leader with the Pacific Northwest National Laboratory (PNNL), Richland, WA, USA. His research interests focus on data analytics for power system reliability, multi-discipline resilience analysis and high-performance computing. He is a Senior Member of IEEE, and serves as a volunteer reviewer of 20+ top-level journals and conferences in the area of power systems and signal processing.









(d) Sep 14-16, 2020

Fig. 12: Illustration of performance enhancement with the proposed parameter tuning and unsupervised classification.