

A Data-driven Variable Reduction Approach for Transmission-constrained Unit Commitment of Large-scale Systems

Yuzhou Zhou, *Member, IEEE*, Qiaozhu Zhai, *Member, IEEE*, Lei Wu, *Fellow, IEEE*, and Moammad Shahidehpour, *Fellow, IEEE*

Abstract—This paper presents a data-driven variable reduction approach to accelerate the computation of large-scale transmission-constrained unit commitment (TCUC). Lagrangian relaxation (LR) and mixed-integer linear programming (MILP) are popular approaches to solving TCUC. However, with many binary unit commitment variables, LR suffers from slow convergence and MILP presents heavy computation burden. The proposed data-driven variable reduction approach consists of offline and online calculations to accelerate computational performance of the MILP-based large-scale TCUC problems. A database including multiple nodal net load intervals and the corresponding TCUC solutions is first built offline via the data-driven and all-scenario-feasible (ASF) approaches, which is then leveraged to efficiently solve new TCUC instances online. On/off statuses of considerable units can be fixed in the online calculation according to the database, which would reduce the computation burden while guaranteeing good solution quality for new TCUC instances. A feasibility proposition is proposed to promptly check the feasibility of the new TCUC instances with fixed binary variables, which can be used to dynamically tune parameters of binary variable fixing strategies and guarantee the existence of feasible UC solutions even when system structure changes. Numerical tests illustrate the efficiency of the proposed approach.

Index Terms—Unit commitment, accelerated algorithm, data driven, variable reduction.

I. INTRODUCTION

UNIT commitment (UC), as a key building block in power system planning and operation and electricity market clearing problems, has long-standing challenges in computational performance and solution quality [1]. As a matter of

Manuscript received: June 14, 2021; revised: August 9, 2021; accepted: September 30, 2021. Date of CrossCheck: September 30, 2021. Date of online publication: February 11, 2022.

This work was supported in part by the National Natural Science Foundation of China (No. 61773309).

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

Y. Zhou and Q. Zhai are with Systems Engineering Institute, MOEKLINNS Lab, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yzzhou@sei.xjtu.edu.cn; qzzhai@sei.xjtu.edu.cn).

L. Wu (corresponding author) is with Electrical and Computer Engineering Department, Stevens Institute of Technology, Hoboken, U.S. (e-mail: lei.wu@stevens.edu).

M. Shahidehpour is with Electrical and Computer Engineering Department, Illinois Institute of Technology, Chicago, U.S. (e-mail: ms@iit.edu).

DOI: 10.35833/MPCE.2021.000382

fact, regional transmission organizations (RTOs) follow stringent market clearing timelines, e.g., the 1800 s time limit of MISO [2] to compute UC solutions and inform market participants of the results. However, as a non-deterministic polynomial-time hard (NP-hard) mixed-integer optimization problem with multitudes of binary/continuous variables and equality/inequality constraints, UC is recognized as the most complex math problem [3] in power system operation. To this end, fast computation approaches for UC problem are in urgent needs, especially in recognizing the increasing scale and complexity of modern power systems.

Lagrangian relaxation (LR) [4]–[7] and mixed-integer linear programming (MILP) [8]–[12] are among the most popular approaches to solve UC problems. LR solves the dual problem of UC, and uses the optimal dual solution to heuristically construct a feasible solution to the original UC problem. Indeed, constructing feasible solutions and accelerating convergence are two core challenges of LR, and many efforts have been taken on these aspects. A systematic approach is presented in [4] to construct feasible solutions to security-constrained UC (SCUC) problems based on analytical feasibility conditions within the LR framework. Relaxation-based neighborhood search and improved relaxation inducement are proposed in [5] to reduce the computation time of UC. A surrogate LR approach is developed in [6] to accelerate the convergence by penalizing constraint violations via quadratic functions. A variable reduction approach is applied to large-scale UC problems in [7], for improving the computational efficiency and ensuring solution quality.

With the recent advance of commercial MILP solvers such as CPLEX and Gurobi, the MILP-based approach becomes the mainstream approach in the power industry. In this direction, advanced modeling and decomposition techniques have been explored to resolve the computational concerns of MILP-based approaches for large-scale UC problems. Tight formulations and dynamic global cutting plane approaches are presented in [8] to accelerate transmission-constrained UC (TCUC). Reference [9] proposes a multi-cut outer approximation approach to decompose the UC problem into an MILP master problem (MP) and several nonlinear programming subproblems (SPs). An efficient MILP approximation approach is proposed in [10] for the hydro-thermal UC problem based on the variable separation and piecewise



linear techniques. A tight and compact MILP formulation for start-up and shut-down power trajectories of thermal units is presented in [11]. Reference [12] proposes a strengthened MILP formulation for gas turbines in the UC problem, in which several new families of strong valid inequalities are developed to reduce the computational time. Clustered unit commitment (CUC) model [13] groups identical or similar plants into clusters to replace the binary commitment variables within a cluster by a single integer variable. And [14] presents a set of constraints to correctly represent the units' hidden flexibility, thus replicating the result of the individual UC. By incorporating critical region preparation in gap time, [15] transfers most computation burden from the online decision stage to the gap time via a customized technique based on multi-parametric programming.

In observing the recent advance in data-driven approaches and the volume of available data, data-driven approaches are explored to solve the UC problem [16]-[20]. Reference [16] presents an integrated framework combining a data-driven approach and a variable-aggregation approach to improve the computational performance of SCUC and achieve good solution quality. In [17], a data-driven approach leveraging historical information is proposed to screen out and eliminate non-binding transmission constraints in the TCUC problem. A number of machine learning techniques is proposed in [18] to extract the information from previously solved instances, which can be used to improve the computational performance of MILP solvers when solving similar instances in the future. In [19], a dataset containing large samples of diverse environment and grid conditions along with their respective UC solutions is created to accelerate the computational procedure. The learning-based approach in [20] learns the map from the parameter to the optimal integer solution and the optimal basis, and then is discretized to accelerate the solving of UC problems online.

From the above literature review, we note that the LR-based approaches [4]-[7] suffer from the convergence issues, e.g., zigzagging effect, and feasibility issues. The computational time of the MILP-based approaches [8]-[12] highly depends on the number of nodes in the branch-and-bound (BAB) tree to be explored as well as time spent on each node (for generating cutting planes and solving the corresponding nodal relaxation problem). Although valuable efforts in [4]-[12] are proposed to improve LR-based and MILP-based approaches, these critical issues are not fully resolved. Besides, the CUC model [13], [14] is limited by the number of identical or similar plants, and when there are no identical or similar plants, the CUC problem will degenerate into the original UC problem. Model-driven approaches [15] are limited by the robust framework and the scalability cannot be guaranteed.

In comparison, the data-driven approaches show potentials to fundamentally mitigate computational complexity of the UC problem. However, in the existing data-driven studies, [16] and [17] focus on handling network security constraints, rather than directly addressing the UC solutions. Reference [18] offers initial UC solutions to accelerating the

solving process, which may be infeasible due to initial on/off statuses and other constraints. Similarly, the feasibility of UC solutions involved in the offline database is also not discussed in [19]. In [20], the division of discrete polyhedral regions is more susceptible to these differences in parameters of offline and online TCUC instances and the feasibility can only be guaranteed by a sufficiently large number of learning samples.

This paper discusses a new data-driven variable reduction approach to efficiently reduce the number of binary variables for large-scale TCUC problems and render the fast calculation. The main idea is depicted in Fig. 1.

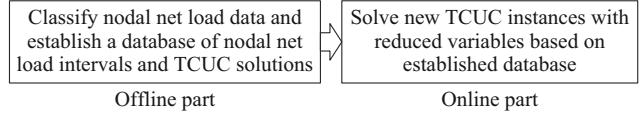


Fig. 1. Main idea of proposed data-driven variable reduction approach for fast TCUC.

The key features are summarized as follows.

1) The proposed data-driven variable reduction approach for fast TCUC includes two parts of offline and online calculations. Specifically, a database including multiple nodal net load, i.e., the load minus renewable energy output of each node, intervals and corresponding TCUC solutions is established offline. This knowledge is then leveraged to assist in fixing certain on/off binary variables in new TCUC instances, reducing the online computational burden of TCUC.

2) Different from the existing data-driven approaches such as [18] and [19], the TCUC solution for each nodal net load interval in the offline database is solved by an all-scenario-feasible (ASF) UC approach, ensuring solution feasibility and optimality for any load realization contained in this interval. In turn, this helps the solution quality of the online calculation of TCUC.

3) In the online calculation, several customized schemes are proposed to achieve the adaptability to the differences between online and offline TCUC instances such as initial on/off statuses and system parameter changes. Specifically, a feasibility check proposition is proposed to strategically adjust the number of on/off statuses for generators to be fixed, guaranteeing the feasibility against the new nodal net load instances and system parameter changes.

4) Numerical tests on the IEEE 6-bus, IEEE 118-bus, and Polish 2383-bus systems show that, with the comprehensive offline database established according to sufficient and high-quality historical and simulated nodal net load data, considerable on/off statuses of generators in the new TCUC instances can be strategically fixed via the customized schemes. Therefore, the computational efficiency with acceptable solution quality is improved.

The rest of this paper is organized as follows. Section II presents the framework of the proposed data-driven variable reduction approach for fast TCUC. The offline database is established in Section III, and the online fast TCUC calculation is introduced in Section IV. Numerical results are analyzed in Section V and Section VI concludes this paper.

II. FRAMEWORK OF PROPOSED DATA-DRIVEN VARIABLE REDUCTION APPROACH FOR FAST TCUC

This paper aims at accelerating the computational performance of TCUC for large-scale power systems. The main idea is to first use the data-driven approach for exploring the relationship between nodal net load profiles and feasibility/optimality of UC solutions, which is then leveraged to guide binary variable fixing strategies and render fast online calculation of new TCUC instances. The flowchart of proposed data-driven variable reduction approach for fast TCUC is shown in Fig. 2.

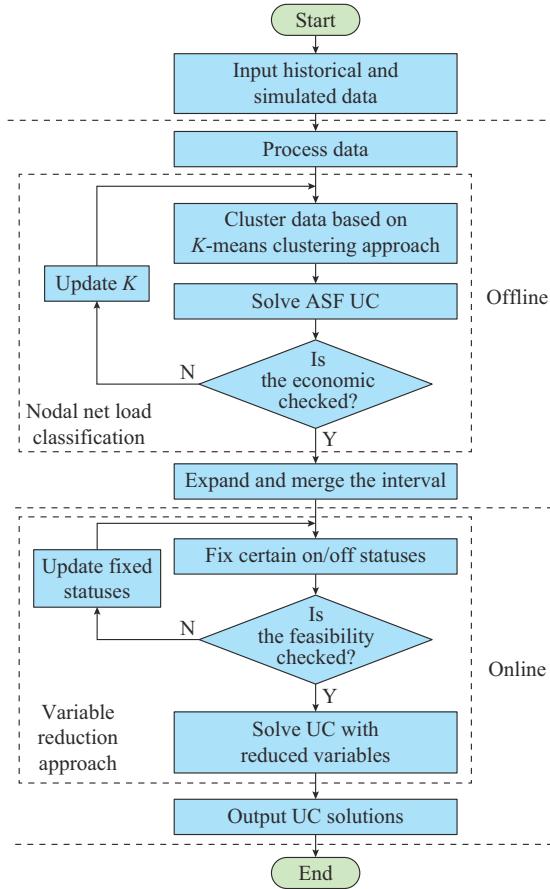


Fig. 2. Flowchart of proposed data-driven variable reduction approach for fast TCUC.

The first step is an offline procedure, which leverages historical and simulated nodal net load data, K -means clustering approach, and ASF UC approach to establish a database including a suite of classified nodal net load intervals and the corresponding unique TCUC solutions. Each record in the database describes the nodal net load intervals, and the corresponding unique TCUC solution supports the system operations when individual nodal net load varies within the intervals. The second step is an online procedure, which compares nodal net load profiles of new TCUC instances against individual classified nodal net load intervals to strategically fix on/off statuses of certain units. It will reduce the number of binary variables and promptly derive TCUC solutions. The establishment of offline database and the calculation of fast online TCUC will be detailed in Sections III and IV.

Two aspects of the proposed data-driven variable reduction approach for fast TCUC are discussed as follows.

1) The offline database is established based on the nodal net loads. In the actual operation of complex power systems, the TCUC solutions would depend on many features such as loads, generation bids, and system structure. However, if all features are considered to build the offline database, the database would be considerable large, impacting its establishment and application. Therefore, since nodal net loads contain the information on loads, renewable energy outputs, and partial system structure, i.e., locations of loads and renewables, and have a great influence on scheduling results, the offline database is established by leveraging nodal net loads, which is essentially a tradeoff between data scale and calculation efficacy.

2) The proposed data-driven variable reduction approach for fast TCUC could handle the differences in parameters of offline and online TCUC instances such as initial on/off statuses, generation bids, and system structures. In the online calculation of the proposed approach, several customized schemes are proposed to achieve the adaptability for such differences. Specifically, the online calculation includes a feasibility check pass, which can dynamically tune the number of on/off statuses for the generators to be fixed, guaranteeing the feasibility against the new nodal net load instances and changes of system parameters. In the extreme situation, when the online system parameters significantly differ from the offline database, only a limited number of on/off statuses will be fixed to guarantee the solution feasibility.

III. ESTABLISHMENT OF OFFLINE DATABASE

We first leverage multiple historical and simulated nodal net load data, with respect to the fixed grid structure of a given power system, to establish the offline database. To guarantee the solution quality of online TCUC calculation, the established offline database shall follow two principles.

1) Feasibility: each classified nodal net load interval corresponds to a unique UC solution. This solution remains feasible for any load realization contained in this interval.

2) Optimality: for any nodal net load profile contained in a classified nodal net load interval, the objective value with respect to the corresponding unique UC solution is within a prespecified threshold ε of the true optimal TCUC solution of the nodal net load profile.

A. Generation and Preprocessing of Nodal Net Load Data

As the database covering a wider range of nodal net load intervals would provide more information for the online TCUC calculation, besides historical nodal net load data, we also generate additional nodal net load data to establish the database offline. The simulated data intend to compensate the historical data for covering representative nodal net load situations. In this paper, the simulated nodal net load data are generated by the proper forecasting approaches [21]-[23], and their time resolution is consistent with the TCUC problem to be studied. Collected historical nodal net load data may contain certain obvious errors, which may influence the classification accuracy. To this end, the de-noising prepro-

cess of the data is adopted to remove these faulty data for improving the accuracy of the database.

B. Classification of Nodal Net Load

The proposed classification algorithm of nodal net load includes three steps. ① Classify net load data samples into K clusters, i.e., K nodal net load intervals. ② Check TCUC feasibility and optimality of each classified nodal net load interval. If both feasibility and optimality of a nodal net load interval are achieved, this classified nodal net load interval is completed. Otherwise, all nodal net load data in this interval will be reclassified. ③ Update K and go back to ① until all nodal net load data have been classified. Finally, multiple nodal net load intervals and their corresponding TCUC solutions that can simultaneously guarantee feasibility and optimality are obtained. The detailed nodal net load classification algorithm is described in Algorithm 1.

Algorithm 1: nodal net load classification

Input and initialization:

Step 1: Collect historical and simulated nodal net load data into set Ω , set threshold ε , and initialize the solution set \mathcal{C} as null

Step 2: Compute the optimal objective value of UC problem (1)-(3) for each nodal net load data in set Ω

Step 3: Analyze nodal net load data and their UC solutions to set a proper initial classification number K

Classify data:

Step 4: Classify nodal net load data in set Ω into K categories via the K -means clustering approach

Step 5: Calculate nodal net load interval of each category

Assess feasibility and optimality:

Step 6: Set $n=0$

Step 7: for $k=1:K$

Step 8: Calculate the unique UC solution corresponding to the nodal net load interval of category k by ASF UC approach (8)-(13)

Step 9: If a feasible UC solution is obtained, go to Step 10; otherwise, set $n=n+1$ and go to Step 7

Step 10: Compute objective values of the ED problem (4), (5) for all nodal net load data in this category

Step 11: Calculate the difference between ED objective from Step 10 and UC objective from Step 2 for all nodal net load data in this category

Step 12: If the differences of all nodal net load data are smaller than ε , save the UC solution and nodal net load interval of this category to the solution set \mathcal{C} , and remove all nodal net load data in this category from set Ω ; otherwise, set $n=n+1$

Step 13: End for

Step 14: If $n>0$, set $K=n+1$ and go to Step 4; otherwise, output \mathcal{C}

The convergence of the proposed classification algorithm is naturally guaranteed, as K will be updated only when the feasibility/optimality check of TCUC in ② fails. That is, in the worst case, each category includes a single sample of historical/simulated nodal net load data. Note that the final classified number of nodal net load categories is determined by continuously updating iterations of Algorithm 1. And the threshold ε could be set by system operators to weigh eco-

nomic and computational efficiency.

As shown in Algorithm 1, three approaches are involved to achieve the interval classification of nodal net load: ① UC and economic dispatch (ED) models; ② K -means clustering approach; and ③ ASF UC approach and acceleration algorithm, which are detailed as follows.

1) UC and ED Models

The UC model J^{UC} (1)-(3) and ED model J^{ED} (4), (5) are used to calculate the optimal reference of classification results in the proposed classification algorithm of nodal net load.

$$J^{UC} = \min_{z_p} (S(z) + C(p)) \quad (1)$$

s.t.

$$Ap + Bz + Ed \leq F \quad (2)$$

$$z \in X \quad (3)$$

$$J^{ED} = \min_p (S(z^*) + C(p)) \quad (4)$$

s.t.

$$Ap + Bz^* + Ed \leq F \quad (5)$$

where z and p are the UC and dispatch decision variables, respectively; $S(\cdot)$ and $C(\cdot)$ are the start-up/shut-down cost and fuel cost, respectively; matrices A , B , E and vector F are constant coefficients; d is the nodal net load vector, i.e., demands minus renewable energy outputs of individual nodes; X is the set of z ; and z^* is a known UC solution derived from (1)-(3).

The objective (1) is to minimize the total operation cost. Prevailing constraints coupling z and p are described as in (2) such as the load balance, generation capacity, ramping, and network constraints. In (3), the set X describes constraints related to z such as the minimum on/off time limits and on/off time requirements.

2) K -means Clustering Approach

K -means clustering approach is applied to classify nodal net load data into multiple categories according to their similarities. Each classified nodal net load category will be used later to identify a unique UC solution. The objective of the K -means clustering approach is to minimize the sum of squared errors over all K clusters, as shown in (6).

$$\min \sum_{k \in \mathcal{K}} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (6)$$

where k is the index of clusters, $k \in \mathcal{K}$; μ_k is the mean vector of the k^{th} cluster; and x_i and c_k are the vector of samples and sample set of the k^{th} cluster, respectively. It is well known that, even for $K=2$, (6) is an NP-hard problem [24]. To this end, the K -means clustering approach is a greedy algorithm that seeks for a local optimum. However, the K -means clustering approach could converge to the global optimum when clusters are well separated [25].

We emphasize several implementation issues of using the K -means to conduct the classification of nodal net load data, while additional details are in [24] and [25].

1) Initial number of clusters K : as shown in Algorithm 1, the number of clusters K is updated iteratively, and the final classification results are guided by the feasibility and optimality.

mality principles stated at the beginning of Section III. Thus, the initial number of clusters K shall have limited influence on the final classification results of Algorithm 1, and it can be set as a small integer value based on the experience.

2) Algorithm stability: although the K -means clustering approach is usually unstable, the stability of Algorithm 1 can be guaranteed because the classification results of Algorithm 1 are guided by the feasibility and optimality principles. Besides, the interval expansion and merging will further help mitigate the potential impacts of classification samples and algorithms on the offline database results. Thus, the stability of Algorithm 1 could be effectively guaranteed.

3) Comparison with other clustering algorithms: K -means clustering approach, as the most widely applied approach in the industry, is applied in this paper because of its simple principle and fast convergence speed. It is also noted that, the proposed classification algorithm of nodal net load can be implemented via other clustering approaches such as hierarchical clustering [26] and density-based spatial clustering of applications with noise [27], in which the depth of hierarchy and cluster radius, instead of the number of clusters, will be used as the input parameter. These approaches may present better stability performance than K -means clustering approach purely from the perspective of clustering, i.e., in terms of *Step 4* of Algorithm 1.

3) ASF UC Approach

For each classified nodal net load interval, the ASF UC approach [28]-[30] is applied to calculate a unique UC solution that meets the feasibility and optimality principles for any load profile contained in this nodal net load interval.

Considering a clustered nodal net load category containing M samples of nodal net load profiles, the maximum and minimum hourly nodal net load levels are used to build the nodal net load intervals $\underline{\mathbf{d}}_{i,t}$ and $\bar{\mathbf{d}}_{i,t}$ for this load category, i.e., $\underline{\mathbf{d}}_{i,t} = \min_m (\mathbf{d}_{m,i,t})$ and $\bar{\mathbf{d}}_{i,t} = \max_m (\mathbf{d}_{m,i,t})$ ($m = 1, 2, \dots, M, \forall t, i$), where t and i are the indices of time periods and nodal net loads, respectively. Constraint (7) describes the box set for the ASF UC approach, ensuring that the unique UC solution \mathcal{D} is feasible for any realization of nodal net load containing the nodal net load interval.

$$\mathcal{D} = \{\mathbf{d}_{i,t} \mid \underline{\mathbf{d}}_{i,t} \leq \mathbf{d}_{i,t} \leq \bar{\mathbf{d}}_{i,t}\} \quad (7)$$

In the ASF UC approach, the key idea is to build a scenario set \mathcal{S} , composed of selective vertex scenarios, and adopt nonanticipative constraints to guarantee the feasibility of the UC solution against the box set (7). As for the selective vertex scenarios, if only one node is associated with the variability, i.e., containing loads and/or renewable energy, the scenarios with the minimum and maximum nodal net load levels constitute the selective vertex scenarios. While for the multi-node uncertainty situation, the selective vertex scenarios shall include the combinations of these minimum and maximum nodal net load levels. The detailed procedure to design vertex scenarios and the proof for the UC solution feasibility are given in [28]-[30].

The ASF UC formulation is established as in (8)-(12) and (3). The objective (8) is to minimize the weighted total costs of all scenarios in set \mathcal{S} . Constraint (9) includes power bal-

ance limits, transmission capacity limits, generation capacity limits, and so on. Nonanticipative constraints (10)-(12) guarantee the feasibility of UC solutions.

$$J^{ASFUC} = \min_{\mathbf{z}, \mathbf{p}^{\max}, \mathbf{p}^{\min}} (\mathbf{S}(\mathbf{z}) + \beta^s \mathbf{C}(\mathbf{p}^s, \mathbf{z})) \quad (8)$$

s.t.

$$\mathbf{A}\mathbf{p}^s + \mathbf{B}\mathbf{z} + \mathbf{E}\mathbf{d}^s \leq \mathbf{F} \quad \forall s \in \mathcal{S} \quad (9)$$

$$\mathbf{p}_t^{\min} \leq \mathbf{p}_t^s \leq \mathbf{p}_t^{\max} \quad \forall t \in \mathcal{T}, s \in \mathcal{S} \quad (10)$$

$$\mathbf{p}_t^{\max} - \mathbf{p}_{t-1}^{\min} \leq \delta^+(\mathbf{z}_t, \mathbf{z}_{t-1}) \quad \forall t \in \mathcal{T} \quad (11)$$

$$\mathbf{p}_t^{\min} - \mathbf{p}_{t-1}^{\max} \geq -\delta^-(\mathbf{z}_t, \mathbf{z}_{t-1}) \quad \forall t \in \mathcal{T} \quad (12)$$

where the superscript s denotes the scenario s ; the superscripts max and min denote the maximum and minimum of allowable outputs of thermal units in the actual operation; the subscripts t and $t-1$ denote the time periods; β^s is the weighting factor of scenario s ; \mathbf{p}^{\min} and \mathbf{p}^{\max} are the auxiliary variables; and δ^+ and δ^- are the ramp-up and ramp-down capabilities, respectively.

In order to accelerate the calculation of large-scale ASF UC models, a time-decoupled decomposition (TDD) algorithm [31] is introduced. As shown in Fig. 3, the TDD algorithm presents a two-level structure; and the superscript * denotes the calculated valued of each variable. The master problem (MP) represents the ASF UC formulation (8)-(12) and (3) with respect to a few representative scenario set \mathcal{S}^{rep} from \mathcal{S} ; the SP checks the feasibility of this UC decision against the entire box set (7) and adds new scenarios with the largest violation into \mathcal{S}^{rep} of MP. The MP and SP are solved iteratively until the stopping criterion is satisfied, i.e., the maximum violation is no larger than tolerance σ . Complete descriptions of MP and SP could be found in [31]. It is noteworthy that the setup of the initial \mathcal{S}^{rep} would influence the convergence of the overall algorithm of nodal net load interval classification, and certain scenarios such as those with the maximum and minimum nodal net loads could be included in \mathcal{S}^{rep} to reduce the number of iterations.

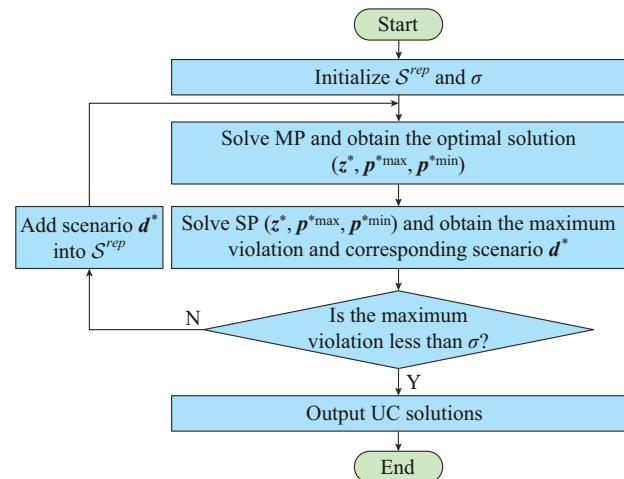


Fig. 3. Framework of TDD algorithm.

The ASF UC approach presents two major advantages: ① ASF UC can guarantee the feasibility of UC solutions against the box set (7), and meet the actual operation needs

on the robustness and nonanticipativity of UC solutions against nodal net load variabilities. Indeed, it is clearly pointed out in [28]-[30] that traditional approaches such as two-stage robust optimization cannot simultaneously guarantee the robustness and nonanticipativity of UC solutions, inducing potential infeasibility in actual operations; and ② instead of the “min-max” or “min-max-min” structure of the two-stage robust optimization, ASF UC approach presents a single-level MILP structure that can effectively mitigate computational complexity and solution over-conservativeness. Noted that the multi-stage stochastic programming and robust optimization approaches [32] could achieve similar effects as the ASF UC approach, but with higher computational burden. The proposed classification algorithm of nodal net load presents good adaptability and computational efficiency.

$$SP(z^*, p^{\max}, p^{\min}): V^{\text{SP}} = \max_{d \in \mathcal{D}} \min_{p, u^1, u^2} \sum_{m,t} (u_{m,t}^1 + u_{m,t}^2) \quad (13)$$

s.t.

$$Ap + Bz + E(d + u^1 - u^2) \leq F \quad (14)$$

$$p_t^{\min} \leq p_t \leq p_t^{\max} \quad \forall t \in \mathcal{T} \quad (15)$$

$$\begin{cases} u^1 \geq 0 \\ u^2 \geq 0 \end{cases} \quad (16)$$

where $m = 1, 2, \dots, M$; V^{SP} is the violation gap between MP and SP; and u^1 and u^2 are the non-negative slack variables.

C. Expanding and Merging Feasible Nodal Net Load Intervals

Some classified nodal net load intervals out of Algorithm 1 may not be the optimal (i.e., the largest) ones that the corresponding UC solution can handle, i.e., satisfying feasibility and optimality principles, due to the limited samples of nodal net load data. To this end, the interval expansion and merging processes are further conducted offline to refine the derived intervals. The interval expansion strategies could also alleviate the potential impacts of the instability issue of the K -means clustering approach on the final database established in this paper.

I) Interval Expansion

The unique UC solution of a certain classified nodal net load interval could potentially handle a larger interval than the one identified as in (7). To this end, an interval expansion model is established to calculate the largest possible nodal net load interval, as shown in (17)-(20).

$$\max_{d_{i,t}^{\max}, d_{i,t}^{\min}, p_t^{\max}, p_t^{\min}} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}} [(d_{i,t}^{\max} - \bar{d}_{i,t}) + (d_{i,t}^{\min} - \bar{d}_{i,t}^{\min})] \quad (17)$$

s.t.

$$Ap^s + Bz^* + Ed^s (d_{i,t}^{\max}, d_{i,t}^{\min}) \leq F \quad \forall s \in \mathcal{S} \quad (18)$$

(10)-(12) with respect to known z^*

$$d_{i,t}^{\min} \leq d_{i,t} \leq \underline{d}_{i,t} \leq \bar{d}_{i,t} \leq d_{i,t}^{\max} \leq \bar{d}_{i,t}^{\max} \quad \forall i \in \mathcal{I}, t \in \mathcal{T} \quad (19)$$

$$S(z^*) + C(p^s) \leq (1 + \varepsilon\%) J^{\text{UC}}(d^s) \quad \forall s \in \mathcal{S}^{\text{typ}} \quad (20)$$

where $d_{i,t}^{\max}$ and $d_{i,t}^{\min}$ are the decision variables describing the newly expanded nodal net load interval; \bar{d} and \underline{d} are the original nodal net load intervals calculated from Algorithm 1; \mathcal{S}^{typ} is the scenario set; and \bar{d}^{\max} and \bar{d}^{\min} are the pre-speci-

fied upper and lower bounds on the interval expansion to guarantee solution boundedness, respectively.

The objective (17) is to maximize the difference between the newly expanded nodal net load intervals and the original ones, i.e., identifying the largest nodal net load interval that z^* can handle. Constraint (18) guarantees the feasibility of z^* against the expanded nodal net load intervals. The scenario set \mathcal{S} is the same as that described in the ASF UC model. Constraint (19) describes the relationship between the original and expanded nodal net load intervals. Constraint (20) restricts the economic efficiency of UC solution against the expanded nodal net load intervals, which is evaluated via set \mathcal{S}^{typ} . As too many scenarios would lead to higher computational burden of the interval expansion problem (17)-(20), \mathcal{S}^{typ} could only include typical scenarios such as those with the maximum and minimum nodal net loads.

2) Interval Merging

After expanding the nodal net load interval via (17)-(20), certain nodal net load intervals may essentially belong to the same category and thus can be merged. For instance, when one interval is fully contained in another, it does not need to be kept in the database. That is, for two nodal net load intervals $[d_{i,t}^{\min,1}, d_{i,t}^{\max,1}]$ and $[d_{i,t}^{\min,2}, d_{i,t}^{\max,2}]$ derived from (17)-(20), if $d_{i,t}^{\min,1} \leq d_{i,t}^{\min,2}$ and $d_{i,t}^{\max,1} \geq d_{i,t}^{\max,2}$, only $[d_{i,t}^{\min,1}, d_{i,t}^{\max,1}]$ and its UC solution remain in the database.

IV. ONLINE FAST TCUC CALCULATION WITH STRATEGIC VARIABLE FIXING SCHEMES

A. Overall Procedure of Online Fast TCUC Calculation

Essentially, the offline algorithm of nodal net load classification establishes a comprehensive database that describes the relationship of nodal net load intervals and corresponding TCUC solutions. Thus, the database can be used to guide binary variable fixing strategies, i.e., certain on/off statuses of thermal units, of new TCUC instances, reducing the number of binary variables and improving the efficiency of online calculations.

The online fast TCUC calculation includes three main steps as follows.

Step 1: identify a trail UC solution for the new TCUC instance based on the offline database. The deviation degree θ defined in (22) is used to identify the nodal net load interval that could potentially best cover the nodal net load profile d^{new} of the new TCUC instance. It quantifies the distance between d^{new} and the median value of classified nodal net load intervals. The UC solution corresponding to the nodal net load interval with the smallest deviation degree is taken as the trail solution of this new TCUC instance.

$$\theta = \sum_{i,t} \left| 0.5(\underline{d}_{i,t} + \bar{d}_{i,t}) - d_{i,t}^{\text{new}} \right| / \sum_{i,t} \left| 0.5(\underline{d}_{i,t} + \bar{d}_{i,t}) \right| \quad (21)$$

Step 2: strategically fix certain on/off statuses of the units based on the trail UC solution. Because parameters of the new TCUC instance, e.g., initial on/off statuses of units and transmission capacity limits, could differ from those used in building the database, the full trail UC solution may be infeasible to the new TCUC instance. To this end, we only fix

certain on/off decisions of units according to the trail UC solution, while others remain as decision variables to be optimized. The strategy for selecting on/off status variables to be fixed is detailed in Section IV-B, in order to reduce computational complexity while guaranteeing solution quality of the new TCUC instance.

Step 3: check the feasibility of the TCUC problem with the fixed on/off statuses. If it is feasible, solve this TCUC problem with the remaining binary variables. Otherwise, go back to *Step 2* and update the fixed on/off statuses of units. The checking approach is detailed in Section IV-C.

B. Fixing Strategies for On/off Statuses of Units

Units are split into two categories based on their on/off statuses in the trail UC solution: ① the first one satisfies the initial minimum on/off time constraints; and ② the second one violates these constraints.

The units in the first category will be arranged in a descending order of their minimum on/off time limits. Then, we select the top units according to the pre-determined radio (PDR) and fix their on/off statuses based on the trail UC solution. PDR, as calculated in (22), determines the ratio of units whose statuses are to be fixed. PDR is a preset parameter that can be adjusted according to the needs of power system operators. Specifically, PDR^{\min} and PDR^{\max} are the pre-specified lower and upper bounds of PDR, respectively. ρ is the correlation coefficient, which could be set according to the requirements of system operators. Larger PDR^{\max} and PDR^{\min} mean more binary variables are fixed. The deviation degree θ represents the similarity between new UC instances and the database. PDR is negatively correlated with deviation degree. Furthermore, PDR could be further dynamically tuned as discussed in Section IV-C, ensuring that the binary variable fixing strategy will not compromise the feasibility of new TCUC instances.

$$PDR = \begin{cases} PDR^{\min} + \rho/\theta & \theta \geq \rho/(PDR^{\max} - PDR^{\min}) \\ PDR^{\max} & \text{otherwise} \end{cases} \quad (22)$$

The units in the second category are also arranged in a descending order of their minimum on/off time limits, and we select the top units according to PDR, i.e., the same as in (22), but with different values of PDR^{\min} and PDR^{\max} . The effects of PDR^{\min} and PDR^{\max} of this category are the same as those of the first category, and are set to be 0.1 and 0.025 in the numerical tests of this paper, respectively. Enough time periods are reserved for these units according to their initial on/off statuses, allowing them to flexibly adjust on/off statuses for satisfying the initial minimum on/off time requirements. The fixing strategy is depicted in Fig. 4, where G/O is the number of time periods that the unit shall be kept on/off as prescribed by its initial on/off statuses and the minimum up/down time limits; and τ^{on}/τ^{off} is the minimum up/down time limit of thermal units.

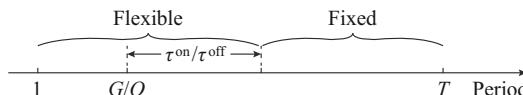


Fig. 4. Fixing strategy for on/off statuses.

The on/off statuses during the period $[1, G/O]$ will be determined according to the initial minimum on/off time requirements; the on/off statuses during the period $[G+1, G+\tau^{on}]$ or $[O+1, O+\tau^{off}]$ will be optimized to offer flexible adjustability of UC solutions; and on/off statuses of the remaining periods will be fixed based on the trail UC solution.

With the above fixing strategy, the TCUC model with a reduced number of binary variables can be solved to accelerate the calculation of the new TCUC instance. Particularly, when the initial minimum on/off time constraints are fully satisfied and the nodal net load profile of a new UC instance is completely contained in a certain classified nodal net load interval, i.e., $d_{i,t}^{new} \in [\underline{d}_{i,t}, \bar{d}_{i,t}]$, the offline UC solution for this nodal net load interval is a good feasible solution for the new TCUC instance, i.e., it is within ε difference from the true optimal solution. That is, a good-enough UC solution of the new UC instance could be directly obtained without solving the new MILP problem.

C. Feasibility Check and PDR Settings

Considering potential differences of system parameters between the offline database and online calculations, the settings of PDR may cause infeasibility to the new TCUC instances, if certain on/off status variables are fixed improperly. One solution is to run multiple instances in parallel with different PDR values, and adopt the solution with the best optimality gap.

In addition, a feasibility proposition is proposed to promptly check the feasibility of the new TCUC instances with fixed binary variables against the changes of system parameters. This fast feasibility-check ability allows us to efficiently tune PDR (or lower/upper bounds of PDR) online and dynamically adjust the on/off status variables to be fixed, guaranteeing the existence of a feasible UC solution, even under system structure changes.

We first define the whole on/off statuses of units as $z = [z^{fix}, z^{unfix}]$, where z^{fix} denotes the on/off statuses of units that are fixed according to the strategies discussed in Section IV-B; and z^{unfix} denotes the on/off statuses of units that remain unfixed. We relax the original TCUC problem and establish a constraint satisfaction problem (23)-(25).

$$A^{new} p + B^{new} [z^{fix}, z^{unfix}] + E^{new} d^{new} \leq F^{new} \quad (23)$$

$$0 \leq p \leq [z^{fix}, z^{unfix}] \bar{P}^{new} \quad (24)$$

$$z^{fix}, z^{unfix} \in X^{new} \quad (25)$$

where A^{new} , B^{new} , E^{new} , F^{new} , d^{new} , and X^{new} are the new TCUC instances; and \bar{P}^{new} is the upper bound of thermal output.

Proposition 1: the necessary condition for the existence of a feasible solution to the original TCUC problem (1)-(3) with new system parameters and nodal net loads is that the constraint satisfaction problem (23)-(25) with $z^{unfix} = 1$ is feasible.

Proof: with z^{fix} , the constraint satisfaction problem (23)-(25) is obtained by relaxing the original inequality $[z^{fix}, z^{unfix}] P^{new} \leq p \leq [z^{fix}, z^{unfix}] \bar{P}^{new}$ to constraint (24), where P^{new} is the lower bound of thermal output. Therefore, it is

obvious that the existence of a solution to the relaxation problem is a necessary condition for the existence of a solution to the original problem.

In the online calculation of a new TCUC instance with given system parameters such as network constraints, one can tune PDR iteratively to adjust z^{fix} to be fixed, guaranteeing that a feasible UC solution exists. Specifically, because fixing more on/off status variables would impact the feasibility of the new TCUC problem more significantly, we first set an initial PDR and check the feasibility according to Proposition 1. If the constraint satisfaction problem (23)-(25) is infeasible, we reduce PDR according to (26) and check the feasibility again until a feasible PDR is found.

$$PDR^{new} = PDR \cdot (1 - \omega\%) \quad (26)$$

where ω is the reduction ratio.

Specially, if the system parameter difference between the daily new UC instances is not significant, the PDR value from the previous day could be used. This tuning process with Proposition 1 can be efficiently conducted, as the constraint satisfaction problem (23)-(25) does not involve binary variables and thus can be solved quickly. Proposition 1 can also be used in the offline part to set a proper initial PDR value, reducing the updating iterations in the online part.

D. Online Update of Database

After solving the new TCUC instance online, the new nodal net load profile and its TCUC solution can also be used to dynamically update the database, providing more accurate information for future online TCUC calculations. Two updating strategies can be adopted as follows: ① when a new nodal net load profile is covered by the upper bound of one nodal net load interval and the lower bound of another, nodal net load samples in these two intervals and the new nodal net load data are reclassified to obtain new nodal net load intervals and UC solutions; and ② when a new nodal net load profile is not contained in any classified nodal net load interval of the database, net load data samples contained in low probability intervals and the new nodal net load profile are reclassified. Low probability intervals refer to the nodal net load intervals which contain significantly fewer net load data samples than others, i.e., smaller than a prespecified threshold of the overall number of net load data samples.

V. NUMERICAL TESTS

The proposed data-driven variable reduction approach for fast TCUC is tested on the IEEE 6-bus, IEEE 118-bus, and Polish 2383-bus systems. Based on the data of ERCOT, 1825 (i.e., 365×5) historical and simulated samples of nodal net loads are used to establish the database offline, and the nodal net load data of one year are used to test the efficiency of the proposed approach for online UC calculation. All tests are implemented via MATLAB R2014a and Gurobi 7.5.2 on a desktop with Intel i7-7700 3.60 GHz CPU and 16 GB RAM.

A. Characteristic Analysis on Classification Results and Interval Expansion

The characteristic analysis of the proposed approach is

conducted on the IEEE 6-bus system, with 3 thermal units and 3 nodal net loads [33]. Capacities of 3 thermal units are 220, 100, and 100 MW, respectively. Other system data can be found in [34].

1) Analysis on Classification Results

K is initially set to be 10 in the K -means clustering approach, and the threshold of low probability intervals is set as $\alpha = 0.5\%$. The final number of nodal net load categories is 37, in which 8 categories are low probability intervals, i.e., the number of net load data samples in each of these 8 categories is less than 10 (i.e., $365 \times 5 \times 0.5\%$). Figure 5 shows the results of four classified load category examples after the first K -means clustering calculation with $K = 10$, and Fig. 6 shows the final classification results via Algorithm 1. Note that Loads 1-3 in Figs. 5-7 are representative nodal net loads selected in this paper.

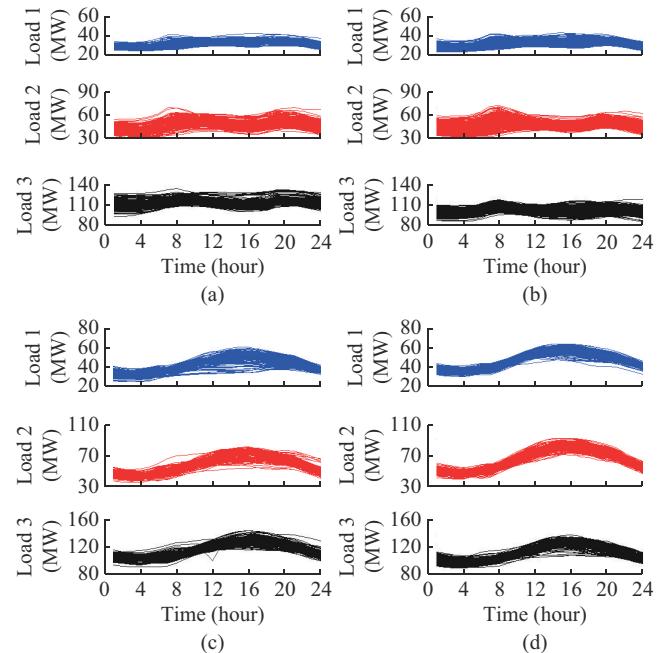


Fig. 5. Results of four classified load category examples after the first K -means clustering calculation. (a) Results of classification No. 1. (b) Results of classification No. 2. (c) Results of classification No. 3. (d) Results of classification No. 4.

Figures 5 and 6 show that the final classification results of the proposed classification algorithm of nodal net load are thinner than those of the first K -means clustering calculation. This can be clearly observed by comparing Fig. 5(b) and Fig. 6(b). The reason is that the K -means clustering approach is a clustering approach based on data density, which only reflects class spacing and data characteristics. In comparison, the proposed classification algorithm, besides embedding K -means clustering approach to reflect class spacing and data characteristics, also guarantees the feasibility and optimality of unique UC solutions for individual nodal net load intervals, thus deriving finer load categories. We further note that the net load data samples in Figs. 5 and 6 may not necessarily be one-to-one correspondence.

The numbers of net load data samples in Fig. 5(a)-(d) and

Fig. 6(a)-(d) are 136, 298, 145, 175, 197, 127, 87, and 82, respectively. One interesting observation is that although the nodal net load interval in Fig. 6(a) is thinner than that in Fig. 5(a), it indeed contains more samples. The reason is that the proposed classification algorithm does not simply divide the results of K -means, but reclassifies the samples with respect to UC characteristics. In addition, the final classification results indicate that the proposed classification algorithm of nodal net load effectively identifies 37 UC solutions, which are regarded as the most relevant optimal operation schemes of this system as implicated by the historical and simulated nodal net load data.

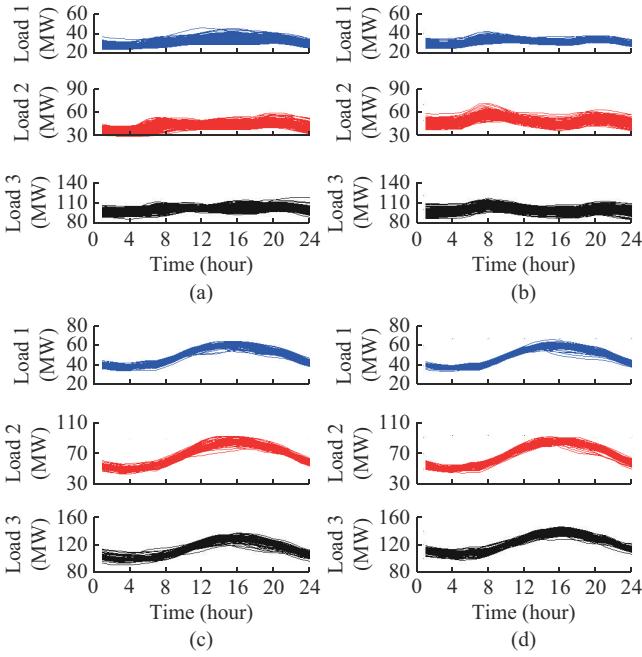


Fig. 6. Final classification results via Algorithm 1. (a) Results of classification No. 1. (b) Results of classification No. 2. (c) Results of classification No. 3. (d) Results of classification No. 4.

2) Analysis on Interval Expansion

The interval expansion is to find the maximum nodal net load intervals that the corresponding UC solutions could handle without the feasibility and optimality of comprising solution. Four expanded nodal net load intervals are shown in Fig. 7, and the numbers of net load data samples contained in these four intervals are 197, 127, 82, and 11, respectively.

The expanded nodal net load intervals in Fig. 7(a) and (b) are the same as the original ones; the expanded nodal net load intervals in Fig. 7(c) are slightly larger than the original ones; and the expanded nodal net load intervals in Fig. 7(d) are significantly larger than the original ones. Correspondingly, the number of net load data samples in the first nodal net load interval is the largest, while that in the fourth interval is the smallest. The reason is that more net load data samples would bring more information to help explore the maximum interval boundaries that the corresponding UC solutions can handle. Conversely, the intervals containing fewer net load data samples may be expanded more significantly.

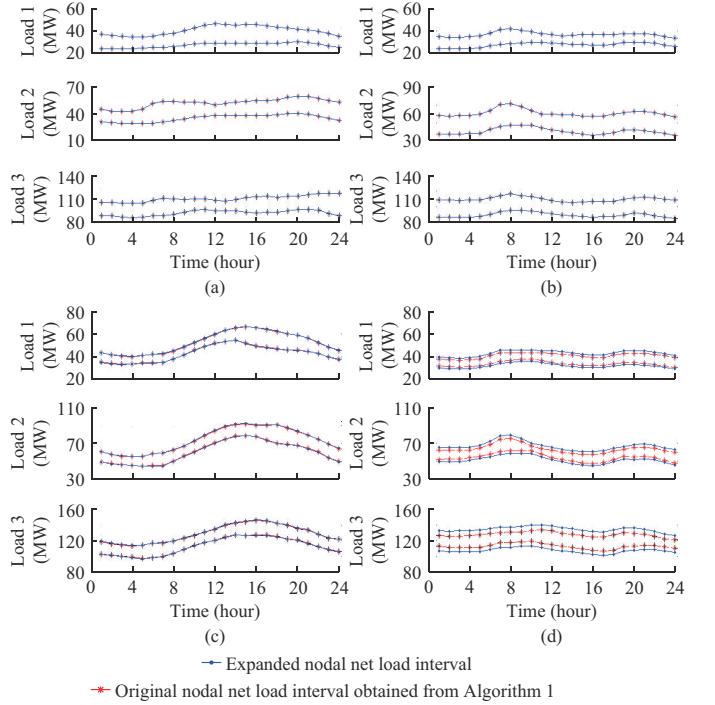


Fig. 7. Four expanded nodal net load intervals. (a) Interval of classification No. 1. (b) Interval of classification No. 2. (c) Interval of classification No. 3. (d) Interval of classification No. 4.

B. Validation of Efficiency and Effectiveness

The IEEE 118-bus and Polish 2383-bus systems with 1825 net load data samples [33] are used to test the effectiveness of the proposed data-driven variable reduction approach. The 118-bus system includes 54 thermal units and 91 uncertain nodal net loads [34], and the Polish 2383-bus system includes 179 thermal units, 182 renewable plants, and 1789 loads [35]. The initial upper and lower bounds of PDR are set to be 0.5 and 0.05 for the first-category units discussed in Section IV-B, and 0.1 and 0.025 for the second-category units, respectively.

1) Analysis on Computational Efficiency

The computational time for IEEE 118-bus and Polish 2383-bus systems is analyzed. Testing results are shown in Figs. 8 and 9 and Tables I and II.

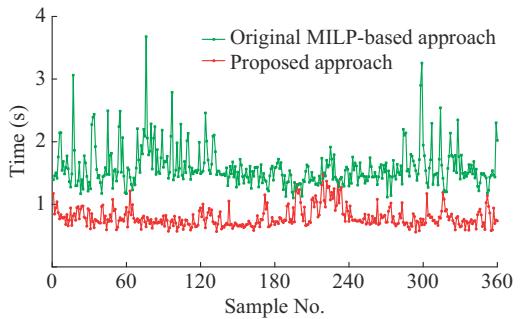


Fig. 8. Computational time of TCUC for IEEE 118-bus system.

Figures 8 and 9 show that computational time of the proposed data-driven variable reduction approach for fast TCUC are much shorter than the original MILP-based approach for both systems.

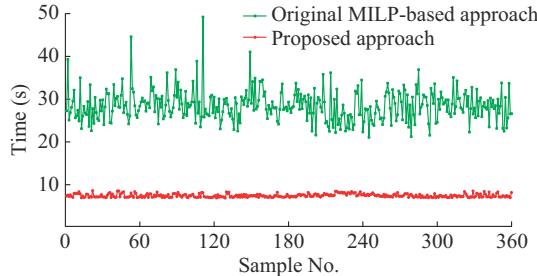


Fig. 9. Computational time of TCUC for Polish 2383-bus system.

TABLE I
TIME-RELATED RESULTS OF IEEE 118-BUS SYSTEM

Approach	Average time (s)	STD of average time (s)	Reduction ratio (%)
Original MILP-based approach	1.59	0.32	
Proposed approach	0.79	0.16	50.31

TABLE II
TIME-RELATED RESULTS OF POLISH 2383-BUS SYSTEM

Approach	Average time (s)	STD of average time (s)	Reduction ratio (%)
Original MILP-based approach	28.44	3.59	
Proposed approach	7.48	0.41	73.70

Besides, the proposed approach performs more stably, i.e., standard deviation (STD) of computational time among different TCUC instances, than directly solving the original TCUC problem. Tables I and II report statistical results of computational time, where the reduction ratio represents reduced time of the proposed approach in percentage against computational time of solving the original TCUC problem. These tests clearly show that the proposed approach is superior to directly solve the original TCUC problem, in terms of both average time and maximum time.

In addition, the comparison between the IEEE 118-bus system and the Polish 2383-bus system indicates that computational benefits of the proposed approach are more significant on larger systems. The reason is that the locations of units with larger minimum on/off time limits could influence the computational performance, when certain on/off decisions of these units are fixed during the online calculation. Units with larger minimum on/off time limits are more concentrated on the IEEE 118-bus system. Thus, when on/off statuses of these units are first fixed, power flow constraints of certain transmission lines are very likely to be binding, which affects the computational efficiency. For instance, the computational time of the proposed approach for several new TCUC instances is even longer than the original UC problem, as shown in Fig. 8. The locations of certain units and capacities of certain transmission lines could be considered in the proposed approach to further improve the efficiency of fast TCUC in the future.

2) Analysis on Solution Quality

Table III analyzes the economic efficiency and reduced number of binary variables of the proposed approach. The maximum relative errors and average relative errors represent the maximum and average relative errors of the TCUC objective value, respectively. The average reduction number and reduction ratio represent the average number and reduction ratio of reduced binary variables by the proposed approach, respectively. With the average relative errors of only 0.47% and 0.67%, the proposed approach shows good economic efficiency on both systems. Moreover, significant binary variables are fixed by the proposed approach, i.e., with the average reductions of 57.33% and 58.82%, with guaranteed solution quality.

TABLE III
OBJECTIVES AND VARIABLE REDUCTION RELATED RESULTS

System	The maximum relative error (%)	Average relative error (%)	Average reduction number	Reduction ratio (%)
118-bus	1.32	0.47	743	57.33
2383-bus	2.40	0.67	2527	58.82

Figure 10 compares UC solutions of the proposed approach and the original MILP-based approach for the 118-bus system. It shows that only 27 out of 1296 (i.e., 54×24) on/off statuses are different. The proposed approach performs well in identifying the optimal UC solutions.

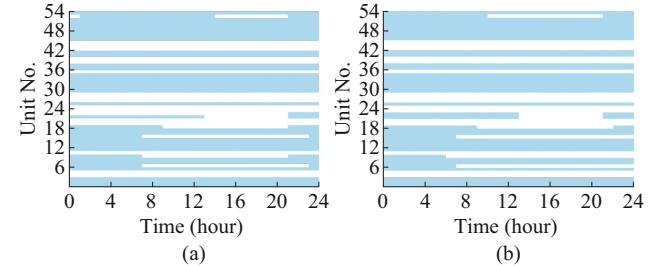


Fig. 10. UC solutions of IEEE 118-bus system. (a) Proposed approach. (b) Original MILP-based approach.

3) Influence of Generator Bidding Price Changes

Bidding price change is among the most significant factors that influence economic performance. Different levels of bidding price changes in the new TCUC instance against the values used in the database are compared to show the impacts on the proposed approach. The testing results are described in Table IV, which are the average values of 365 days. In Table IV, price change represents the deviation level of bidding prices between the offline database and online calculation.

Table IV shows that, with the increase in the change level, the average computational time is almost the same, and the economic efficiency gets worse. Because the on/off statuses of certain thermal units are fixed in the proposed approach, when the bidding prices become higher/lower, these thermal units can only reduce/increase power output levels to minimize the total cost. To this end, the economic efficiency

would become worse than the original optimal TCUC solutions. However, it is noted that with the 25% bidding price fluctuation, the average relative error in the total cost is still less than 1%, indicating acceptable economic performance of the proposed approach against the bidding price changes.

TABLE IV
INFLUENCE OF BIDDING PRICE CHANGE

Price change (%)	Average time (s)	The maximum time (s)	Average error (%)	The maximum error (%)
10	0.80	1.52	0.50	1.68
15	0.80	1.64	0.57	2.02
20	0.79	1.45	0.67	2.05
25	0.79	1.46	0.88	2.09
30	0.80	1.33	1.13	3.55

4) Influence of Data Scale

The data scale refers to the amount of historical and simulated data used to build the database offline, which could influence classification results and in turn computational performance of the online calculation. Five data scales, i.e., 1 year to 5 years, are compared to analyze the impact. The tests are implemented on the IEEE 118-bus system and results are reported in Table V. The results in Table V are the average values of 365 days.

TABLE V
INFLUENCE OF DATA SCALE

Data scale (year)	Average time (s)	Average error (%)
1	0.79	0.55
2	0.76	0.50
3	0.78	0.49
4	0.79	0.50
5	0.79	0.47

Table V shows that the average solution time and average relative error of 365 new TCUC instances are almost the same with different data scales. Thus, one-year data are enough to reach good performance for this system.

5) Comparison of Different Data-driven Approach

Since different data-driven approaches would bring varying computational performance, we further compare the proposed approach in this paper with the data-driven approaches in [17] and [18]. As the actual performance of data-driven approaches also highly depends on the sophisticated implementation strategies and parameter tuning skills, for the fair comparison, we directly cite the results reported in [17] and [18] instead of reproducing their studies. Table VI lists the relative errors, relative reduction time, infeasible rate, reduction variable, and system scale of references [17] and [18], which is benchmarked with the original MILP-based approach.

Table VI shows that, although the data-driven approach in [17] shows the best performance, it cannot guarantee 100% feasibility for new TCUC instances. Compared with the approach in [18], the proposed approach performs a 17.3% economic advantage while with an economic disadvantage of

0.57%. Besides, compared with the reduction variables, the approach in [17] is to reduce constraints rather than binary variables, so the reduction variable is empty. Since the proposed approach has higher computational efficiency with a lower reduction variable rate, it shows its greater potential in accelerating the computation. In a word, the proposed approach shows good computational efficiency and acceptable economic efficiency on the premise of guaranteeing 100% feasibility.

TABLE VI
COMPARISONS OF DIFFERENT ACCELERATION APPROACHES

Approach	Relative error (%)	Relative reduction time (%)	Infeasible rate (%)	Reduction variable (%)	System scale
Approach in [17]	0.02	95.4	0.06		2000-bus
Approach in [18]	0.10	56.4	0.00	68.7	2848-bus
Proposed approach	0.67	73.7	0.00	58.8	2383-bus

6) Influence of PDR Settings

The settings of PDR influence the fixed on/off statuses of thermal units, and then influence the computational efficiency and economic performance for new TCUC instances. To this end, the influence of PDR on computational performance is investigated in Table VII.

TABLE VII
COMPARISONS OF DIFFERENT PDR SETTINGS

PDR setting (%)	Relative error (%)	Relative reduction time (%)	Relative reduction number of binary variable (%)
30	0.32	59.26	30.00
40	0.49	68.26	40.00
50	0.67	73.30	50.00
60	0.67	72.76	60.00
70	0.89	75.10	69.94
80	1.04	76.21	79.63
90	1.13	76.43	88.69

In Table VII, PDR setting represents that the values of PDR^{\max} and PDR^{\min} are always set to be 5%. Table VII shows that relative errors, relative reduction time, and relative reduction number all increase with the value of PDR. When the PDR settings are larger than 50%, the change rate of relative reduction time becomes lower while damaging similar economic efficiency. Therefore, considering the calculation efficiency and economic benefits, the 50% setting is a better choice for this case.

VI. CONCLUSION

This paper presents a new data-driven variable reduction approach to accelerate the calculation of TCUC problems for large-scale systems. It includes an offline data-driven approach to build a database that explores the relationship of nodal net load intervals and corresponding UC solutions,

and an online approach to promptly calculate UC instances by strategically fixing on/off statuses of considerable units via the information of the database. By making full utilization of historical and simulated nodal net load information, it can essentially reduce the number of binary variables and improve the computational efficiency of TCUC problems.

Numerical tests illustrate the effectiveness of the proposed data-driven variable reduction approach, which could greatly reduce the computational time and present high computational stability. Future works could extend the fast solution approach to solve stochastic TCUC problems with heterogeneous uncertainty factors such as renewable energy and demand response assets. The fast solution approach may also be extended to include other key parameters relevant to the optimal TCUC solutions such as generation bids, for further improving the computational performance and solution quality.

REFERENCES

- [1] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*. New York: Wiley, 2002.
- [2] Y. Chen, A. Casto, F. Wang *et al.*, “Improving large scale day-ahead security constrained unit commitment performance,” *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4732-4743, Nov. 2016.
- [3] National Academies of Sciences, Engineering, and Medicine, *Analytic research Foundations for the Next-generation Electric Grid*. Washington D. C.: National Academies Press, 2016.
- [4] H. Wu, X. Guan, Q. Zhai *et al.*, “A systematic method for constructing feasible solution to SCUC problem with analytical feasibility, conditions,” *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 526-534, Feb. 2012.
- [5] Z. Ma, H. Zhong, Q. Xia *et al.*, “A unit commitment algorithm with relaxation-based neighborhood search and improved relaxation induce-ment,” *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3800-3809, Sept. 2020.
- [6] X. Sun, P. Luh, M. Bragin *et al.*, “A novel decomposition and coordination approach for large day-ahead unit commitment with combined cycle units,” *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5297-5308, Sept. 2018.
- [7] X. Li, Q. Zhai, J. Zhou *et al.*, “A variable reduction method for large-scale unit commitment,” *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 261-272, Jan. 2020.
- [8] L. Wu, “Accelerating NCUC via binary variable-based locally ideal formulation and dynamic global cuts,” *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4097-4107, Sept. 2016.
- [9] L. Yang, J. Jian, Z. Dong *et al.*, “Multi-cuts outer approximation method for unit commitment,” *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1587-1588, Mar. 2017.
- [10] Y. Chen, F. Liu, B. Liu *et al.*, “An efficient MILP approximation for the hydro-thermal unit commitment,” *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 3318-3319, Jul. 2016.
- [11] G. Morales-España, J. M. Latorre, and A. Ramos, “Tight and compact MILP formulation of start-up and shut-down ramping in unit commitment,” *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1288-1296, May 2013.
- [12] K. Pan, Y. Guan, J. Watson *et al.*, “Strengthened MILP formulation for certain gas turbine unit commitment problems,” *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1440-1448, Mar. 2016.
- [13] J. Meus, K. Poncelet, and E. Delarue, “Applicability of a clustered unit commitment model in power system modeling,” *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2195-2204, Mar. 2018.
- [14] G. Morales-España and D. A. Tejada-Arango, “Modeling the hidden flexibility of clustered unit commitment,” *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3294-3296, Jul. 2019.
- [15] W. Zheng, Z. Li, and H. Zhou, “Efficient robust look-ahead dispatch incorporating critical region preparation in gap time,” *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4840-4843, Sept. 2021.
- [16] Y. Yang, X. Lu, and L. Wu, “Integrated data-driven framework for fast SCUC calculation,” *IET Generation, Transmission & Distribution*, vol. 14, no. 24, pp. 5728-5738, Dec. 2020.
- [17] S. Pineda, J. M. Morales, and A. Jiménez-Cordero, “Data-driven screening of network constraints for unit commitment,” *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3695-3705, Sept. 2020.
- [18] Á. S. Xavier, F. Qiu, and S. Ahmed, “Learning to solve large-scale security-constrained unit commitment problems,” *Informs Journal on Computing*, vol. 33, no. 2, pp. 739-756, Feb. 2020.
- [19] G. Dalal, E. Gilboa, S. Mannor *et al.*, “Unit commitment using nearest neighbor as a short-term proxy,” in *Proceedings of 2018 Power Systems Computation Conference*, Dublin, Ireland, Jun. 2018, pp. 1-7.
- [20] M. Li, W. Wei, Y. Chen *et al.*, “Learning the optimal strategy of power system operation with varying renewable generations,” *IEEE Transactions on Sustainable Energy*. doi: 10.1109/TSTE.2021.3088951
- [21] Q. Tu, S. Miao, F. Yao *et al.*, “Forecasting scenario generation for multiple wind farms considering time-series characteristics and spatial-temporal correlation,” *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 837-848, Jul. 2021.
- [22] J. Wang, X. Chen, F. Zhang *et al.*, “Building load forecasting using deep neural network with efficient feature fusion,” *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 1, pp. 160-169, Jan. 2021.
- [23] Y. Wang, N. Zhang, Q. Chen *et al.*, “Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV,” *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3255-3264, May 2018.
- [24] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, Aug. 2010.
- [25] C. Si, S. Xu, C. Wan *et al.*, “Electric load clustering in smart grid: methodologies, applications, and future trends,” *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237-252, Mar. 2021.
- [26] G. Karypis, E. Han, and V. Kumar, “Chameleon: hierarchical clustering using dynamic modeling,” *Computer*, vol. 32, no. 8, pp. 68-75, Aug. 2002.
- [27] M. Ester, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of 1996 Process International Conference Knowledge on Discovery & Data Mining*, Portland, United States, Aug. 1996, pp. 226-231.
- [28] Y. Zhou, Q. Zhai, and L. Wu, “Optimal operation of regional microgrids with renewable and energy storage: solution robustness and nonanticipativity against uncertainties,” *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4218-4230, Nov. 2022.
- [29] Y. Zhou, Q. Zhai, M. Zhou *et al.*, “Generation scheduling of self-generation power plant in enterprise microgrid with wind power and gateway power bound limits,” *IEEE Transactions on Sustainable Energy*, vol. 11, no. 2, pp. 758-770, Apr. 2020.
- [30] Y. Zhou, Q. Zhai, and L. Wu, “Multistage transmission-constrained unit commitment with renewable energy and energy storage: implicit and explicit decision methods,” *IEEE Transactions on Sustainable Energy*, vol. 12, no. 2, pp. 1032-1043, Apr. 2021.
- [31] X. Li and Q. Zhai, “Multi-stage robust transmission constrained unit commitment: a decomposition framework with implicit decision rules,” *International Journal of Electrical Power & Energy*, vol. 108, pp. 372-381, Jan. 2019.
- [32] A. Lorea and X. Sun, “Multistage robust unit commitment with dynamic uncertainty sets and energy storage,” *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 1678-1688, Mar. 2016.
- [33] Y. Zhou. (2021, Mar.). Historical and simulated data. [Online]. Available: https://www.researchgate.net/publication/350373720_historical_and_simulated_data
- [34] AI-Roomi International. (2021, Mar.). Electric power system analysis & nature-inspired optimization algorithms, test system. [Online]. Available: <https://al-roomi.org/component/tags/tag/test-system>
- [35] R. Zimmerman, C. Murillo-Sánchez, and R. Thomas, “MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education,” *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12-19, Jan. 2011.

Yuzhou Zhou received the B.S. degree in automation and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2021, respectively. From 2019 to 2020, he was a Visiting Scholar at Stevens Institute of Technology, Hoboken, USA. His research interests include planning and scheduling of power system with renewable energy and energy storage.

Qiaozhu Zhai received the B.S. and M.S. degrees in applied mathematics and the Ph.D. degree in systems engineering from Xi'an Jiaotong University, Xi'an, China in 1993, 1996, and 2005, respectively. He is currently a Professor with the Systems Engineering Institute of Xi'an Jiaotong Universi-

ty. His research interests include optimization of large-scale system and integrated resource bidding and scheduling in electric power market.

Lei Wu received the B.S. degree in electrical engineering and the M.S. degree in system engineering from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2004, respectively, and the Ph.D. degree in electrical engineering from Illinois Institute of Technology (IIT), Chicago, USA, in 2008. From 2008 to 2010, he was a Senior Research Associate with the Robert W. Galvin Center for Electricity Innovation, IIT. He was a summer Visiting Faculty at New York Independent System Operator (NYISO) in 2012. He was a Professor with the Electrical and Computer Engineering Department, Clarkson University, Potsdam, USA, till 2018. He is currently a Professor with the Electrical and Computer Engineering Department, Stevens Institute of Technology, Hoboken, USA. His research interests include power system op-

eration and planning, energy economics, and community resilience microgrid.

Mohammad Shahidehpour received the Honorary Doctorate degree from the Polytechnic University of Bucharest, Bucharest, Romania. He is currently a University Distinguished Professor and a Bodine Chair Professor and the Director of the Robert W. Galvin Center for Electricity Innovation, Illinois Institute of Technology, USA. He is a Member of the U.S. National Academy of Engineering and a Fellow of American Association for the Advancement of Science and National Academy of Inventors. His research interests include distributed power generation, power generation economics, optimisation, and power generation dispatch, integer programming, power markets, and power generation scheduling.