Data-Driven Representative Day Selection for Investment Decisions: A Cost-Oriented Approach

Mingyang Sun, *Member, IEEE*, Fei Teng, *Member, IEEE*, Xi Zhang, *Student Member, IEEE*, Goran Strbac, *Member, IEEE*, and Danny Pudjianto, *Member, IEEE*

***Abstract*—Power system investment planning problems become intractable due to the vast variability that characterizes system operation and the increasing complexity of the optimization model to capture the characteristics of renewable energy sources. In this context, making optimal investment decisions by considering every operating period is unrealistic and inefficient. The conventional solution to address this computational issue is to select a limited number of representative operating periods by clustering the input demand-generation patterns while preserving the key statistical features of the original population. However, for an investment model that contains highly complex non-linear relationship between input data and optimal investment decisions, selecting representative periods by relying on only input data becomes inefficient. This paper proposes a novel investment cost-oriented representative day selection framework for large scale multi-spacial investment problems, which performs clustering directly based on the investment decisions for each generation technology at each location associated with each individual day. Additionally, dimensionality reduction is performed to ensure that the proposed method is feasible for large-scale power systems and high-resolution input data. The superior performance of the proposed method is demonstrated through a series of case studies with different levels of modeling complexity.**

***Index Terms*—Clustering, dimensionality reduction, investment planning, renewable energy sources, representative days.**

I. Introduction

DECARBONIZATION of electricity systems will significantly increase the penetration of renewable energy sources (RES). The variability, uncertainty and limited inertia capability of RES lead to fundamental challenges for system control, operation and planning. In particular, the key characteristics of intermittent RES need to be well accommodated in system planning models to achieve optimal investment decisions regarding future low-carbon power systems [1].

Power systems planning is generally modeled as a linear programming (LP) problem aiming to make the optimal investment decisions that minimize the total cost, which consists of operational cost and investment cost. Specifically, investment decisions are normally represented by binary decision variables at coarse time intervals (e.g., yearly) whereas an embedded system operational model makes short-term operational decisions on an hourly time scale [2]. Recently, more advanced planning models with significantly increased complexity have been proposed to fully reflect the challenges of increased penetration of RES through detailed modeling of inter-temporal constraints such as minimum up/down time for generators and the ancillary service requirements in hourly or even sub-hourly time scale [3]. It is important to highlight that the increasing complexity of investment planning models directly results in significant computational burden and may even lead to a problem that cannot be analytically solved if all operating periods are considered. To this end, it becomes imperative to select a subset of representative periods from a vast number of operating periods for consideration in the investment problem to attain optimal or near-optimal investment decisions.

Alternative approaches have been developed for selecting representative periods. Traditionally, heuristic selection is applied by experts to manually determine the representative operating conditions that describe the most relevant scenarios based on the variations in load and RES availability. For example, in [4], 17 time slices are selected from a 2-year period of data (i.e., 16 time slices representative of different seasons and one slice of summer super-peak time) to capture diurnal and seasonal variability in load and generation resources. However, heuristic selection approaches suffer from a lack of consistent selection criteria to use to select the representative periods or to validate the effectiveness of the selected periods [5], [6]. Moreover, the increased penetration of RES leads to more diverse patterns of system operation conditions and makes manual heuristic selection approaches inadequate.

A series of clustering-based methods have recently been proposed in the literature to capture statistical characteristics and correlations among the load and RES data. In [7], the k-means clustering technique is employed to model interspatial correlation between load and wind power generation for investment problems. The authors in [8] use k-means clustering to select representative operating points for wind generation investment. However, solving the generation investment planning problem based solely on the selected representative operating points cannot consider intertemporal operating constraints because of the break in the chronological sequence. Therefore, it is imperative to select representative time slices with longer periods (e.g., days or weeks) that simultaneously capture the correlations among

Manuscript received June 5, 2018; revised October 11, 2018, December 2, 2018, and December 24, 2018; accepted January 9, 2019. Date of publication January 11, 2019; date of current version June 18, 2019. This work was supported by UK Energy Research Centre Phase 3 project (EP/L024756/1). Paper no. TPWRS-00861-2018. *(Corresponding author: Fei Teng.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: mingyang.sun11@imperial.ac.uk; f.teng@imperial.ac.uk; x.zhang14@imperial.ac.uk; g.strbac@imperial.ac.uk; d.pudjianto@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2019.2892619

load, wind and solar; the temporal autocorrelation within each variable; and the interspatial correlations among different locations. To this end, an optimization-based representative day selection method is presented in [5] for capturing the implications of integrating RES in generation investment planning problems, along with proposed representativeness evaluation metrics. The authors of [2] propose two hierarchical clustering-based selection methods for long-term capacity expansion models by considering intertemporal operating constraints. Meanwhile, this method can capture important statistical features of the input operating conditions (i.e., temporal autocorrelations and spatial correlations).

All of the above selection approaches are based on the operating conditions in the input domain (e.g., demand and RES availabilities), with the benefit of straightforward implementation. However, this domain may not be the most efficient domain to perform clustering since the long-term investment decisions are highly non-linear with respect to the input variables. To this end, reference [9] recognizes that power flow patterns are key drivers for investing in new transmission lines. A moment-matching algorithm is applied to cluster operating points based on the optimal power flow (OPF) patterns. Numerical experiments have demonstrated that the OPF-based method indeed results in a more effective reduction in the number of scenarios required for obtaining the optimal transmission investment decisions. Moreover, in [10], an operational state aggregation approach is proposed to select representative conditions according to the line benefit. Note that the requirement of solving a relaxed transmission network expansion planning (TNEP) problem increases the computational cost before the clustering procedure. Additionally, the expected power transfers of the network corridors are considered as the clustering variables in [11], with a special focus on critical situations. The authors in [12] recently proposed an objective-based scenario selection framework that considers the transmission investment decisions of each individual scenario as the clustering variables.

In this context, little research has been conducted to investigate the alternative clustering domains that may lead to more effective selection of representative days (or even longer operating periods) for the investment planning problem with intertemporal operating constraints. In the research field of electricity trading, the authors in [13] and [14] proposed a novel idea, for the first time, to perform scenario reduction on the transformed space (i.e., outcome space), instead of the input space. More specifically, the space of the objective function value is considered as the transformed space. In the context of the investment planning problem, the objective function value is composed of the investment cost and the operation cost. Since the investment cost obtained in the expected value problem has been fixed as proposed in [13] and [14], the cost that is used to differentiate the scenarios are actually the operation costs, which are still not the most straightforward "objective" of the investment problem. Additionally, using a single cost value may neglect the information about the interspatial correlations between generation technologies at various locations.

Inspired by the work reported in [13] and [14] for an electricity trading problem, in this paper, an investment cost-oriented representative day selection framework is proposed for power system investment problems to cluster the operating days based on the costs of investment decisions that are driven by each individual day. Hierarchical clustering with Wards linkage is employed to group the operating periods based on the costs associated with investment decisions. Subsequently, the medoid of each constructed cluster is selected as the representative day. To address the curse of dimensionality of large-scale systems, dimensionality reduction is applied before the clustering procedure. The performance of the proposed method is demonstrated based on a four-location generation investment planning model with different levels of modeling complexity. The key contributions of this paper can be summarized as follows:

i) A novel investment cost-oriented framework, including four main stages of *Run System Investment Planning Per Day*, *Dimensionality Reduction*, *Perform Clustering* and *Representative Day Selection*, is proposed to select representative days for large scale multi-spacial power system investment planning problem with intertemporal operating constraints. In particular, clustering is performed based on a more effective domain - the costs of investment decisions for each generation technology at each location associated with each individual day.
ii) To ensure scalability of the proposed framework for large-scale systems, a nonlinear dimensionality reduction technique, Laplacian Eigenmaps, is implemented prior to clustering to address the curse of dimensionality.
iii) A comprehensive analysis is performed to demonstrate the superior performance of the proposed investment cost-oriented method based on a series of investment planning models with different levels of complexity. The key drivers for the increased benefit of the proposed approach are identified: 1) modelling of inter-temporal constraints and ancillary service, 2) including the RES capacities as decision variables, 3) utilizing high time-resolution input data.

The rest of this paper is organized as follows: Section II presents the representative day selection problem and its main challenges. Section III introduces the proposed framework and the related technical details. Section IV conducts numerical experiments on different investment models. Finally, the main conclusions are given in Section V.

II. Problem Statement

Given the multidimensional operating condition data $X = \{\vec{x}_d, d = 1, \ldots, D\}$ of load, wind and solar availability, where D is the total number of periods, the power system investment planning problem aims to obtain the optimal investment decisions Γ^* that can minimize the total cost $C_{tot}(X, \Gamma^*)$ which is composed of system operational cost C_{op} and investment cost C_{inv} with the consideration of carbon target constraint and a series of intertemporal operating constraints, such as ramp-up and ramp-down constraints. Then, the minimum total cost can be defined as

$$C^* = C_{tot}(X, \Gamma^*) = C_{inv}(\Gamma^*) + C_{op}(X, \Gamma^*) \tag{1}$$

With the increasing complexity of the investment planning model, solving the optimization problem based on the whole dataset X becomes intractable. One effective solution is to select a subset of representative days $X^\dagger$ and their corresponding probabilities $\Psi^\dagger$ such that the investment decisions $\Gamma^\dagger$ obtained by solving the investment planning problem based on $X^\dagger$ can result in

$$C_{inv}(\Gamma^\dagger) + C_{op}(X, \Gamma^\dagger) = C_{tot}(X, \Gamma^\dagger) \approx C^* \tag{2}$$

where $C_{op}(X, \Gamma^\dagger)$ is the system operation cost based on the full operating data X with the investment decisions $\Gamma^\dagger$. In other words, representative period selection aims to regain tractability of the investment planning problem.

Although clustering-based methods have been demonstrated to be an effective approach for selecting representative periods $X^\dagger$, a series of key questions related to representative period selection can be summarized as follows: *How is the optimal clustering domain determined? Among the various clustering techniques, which is the most appropriate for representative period selection? After clustering, how is the representative day of each cluster selected?* The previous work [12] investigated the aforementioned questions with the focus on the selection of representative operating snapshots with the application of TNEP. However, for the power system investment planning problem with intertemporal operating constraints (e.g., the generation investment planning problem), it is imperative to select longer representative periods, such as days, which leads to the additional challenges as following:

Problem 1 (P-1): For the multivariate operating condition data, selecting representative operating snapshots only requires clustering to be performed based on a two-dimensional dataset, including $d1$-the variable (e.g., demand) and $d2$-the object (operating snapshots in this case). However, for representative period selection, clustering a three-dimensional dataset that includes $d1$-the variables, $d2$-the time steps within a period (e.g., 48 half-hourly data points for each day), and $d3$-the operating periods (e.g., days) becomes more difficult.

Problem 2 (P-2): The increased complexity of system investment planning models leads to higher nonlinearity between the input operating conditions and the output investment decisions. Therefore, capturing only the important statistical characteristics of the original input data (e.g., correlation, variability, and distribution) cannot guarantee the optimal investment decisions.

Problem 3 (P-3): The longer the operating periods that we need to select, the higher the dimension of the data that will be clustered in terms of $d2$ (i.e.,the time steps), thus resulting in the curse of dimensionality. In addition, for large-scale systems, a more significant dimensionality problem related to $d1$ could be encountered.

Note that, regarding the operating periods, this paper will focus on selecting representative days with the application of the generation investment planning problem. Nevertheless, the proposed method can be readily expanded to a longer period selection version (e.g., representative week or month selection) by changing the considered operating periods ($d3$) for other long-term investment planning problems.

In addition, beyond the intraday storage, the proposed method faces the challenge of handling the interday storage in the investment model because the selection procedure has not considered the continuity between the representative periods. Recently, the authors in [15] and [16] provide novel solutions to maintain the chronology of the input time series for dealing with interperiod storage. In particular, in [16], the Representative Periods with Transition Matrix and Cluster Indices (RP-TM&CI) model can guarantee some continuity between representative days, which can be employed to link the representative days selected using the proposed cost-oriented approach to capture the interday information. Additionally, a novel chronological time-period clustering method proposed in [15] is developed based on a hierarchical clustering method and performed in the input domain so that the selected representative periods can maintain the chronology of the input time series throughout the planning horizon. To this end, combined with the chronological time-period clustering method, the proposed cost-oriented approach could be further developed to handle both types of storage (i.e., intraday and interday), which will be investigated in our future work.

III. Proposed Cost-Oriented Representative Day Selection Framework

To address the aforementioned challenges, a novel representative day selection framework is proposed in this paper, as shown in Fig. 1. The proposed framework consists of four main steps: ***Step 1-Clustering Domain Transformation***: Run system investment planning for each individual day to obtain the clustering variables, consisting of the costs of the investment decisions per day. ***Step 2-Dimensionality Reduction***: Perform dimensionality reduction on clustering variables to extract effective features. ***Step 3-Cluster Assignment***: Perform clustering on the extracted features. ***Step 4-Representative Day Selection***: Select a representative day of each cluster in the original domain of the input data. A detailed description of each step will be presented in the following subsections.

A. Clustering Domain Transformation (Step 1)

To develop an efficient clustering-based representative period selection approach, the first question that needs to be properly answered is: *'What is the most effective domain in which to perform clustering such that the selected representative periods can lead to near-optimal or even optimal investment decisions?'* Mathematically, this question can be defined as follows. Let $X = \{\vec{x}_d, d = 1, \ldots, D\} \in \mathbb{R}^{D \times [N_B \times (N_G+1)] \times N_d}$ denote the input multivariate operating condition data of the investment planning model, where N_B, N_G, and N_d represent the numbers of network buses, generation technologies, and data points within each operating period, respectively. Note that regarding the dimension of X, $[N_B \times (N_G + 1)] = N_B \times N_G + N_B \times 1$ is the sum of the number of technologies multiply by the number of locations $N_B \times N_G$ and the number of loads $N_B \times 1$. The key challenge related to **P-1** and **P-2** is transforming the data from the input domain into a more effective domain with a nonlinear mapping $f : X \rightarrow \Gamma$ such that the clusters

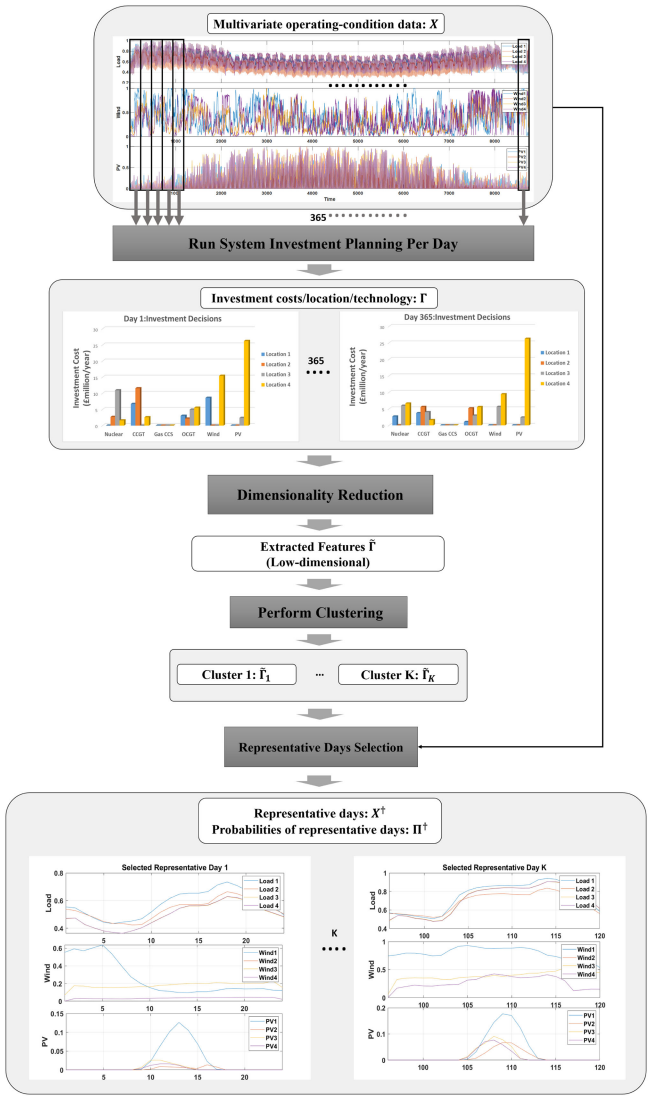


Fig. 1. The proposed cost-oriented representative day selection framework.

$[X^1, \ldots, X^K] \subset X$ constructed based on the cluster labels $y = Clustering(\Gamma)$ can finally result in the optimal investment decisions Γ^* by using the representative periods $X^\dagger$ selected from $[X^1, \ldots, X^K]$. Note that K and $Clustering(\cdot)$ are the number of clusters and the clustering procedure, respectively.

In the literature, most of the current work focuses on obtaining the cluster labels y based on the domain of input data X while attempting to make the selected representative periods $X^\dagger$ exhibit similar statistical characteristics to X (e.g., [2], [5]). In particular, the average and standard deviation of original dataset X are the general target moments that are required to retain in $X^\dagger$. In addition, authors in [5] attempt to preserve the annual load and average RES capacity factors, the distribution of each time series, the correlation between the different time series, and the variability. Nevertheless, clustering the operating periods based on the information in the input domain may not lead to an efficient operating period reduction for the following reasons:

R1: Operating periods with significantly distinguished patterns regarding their statistical factors in the input domain may drive identical or similar investment decisions;

R2: Operating periods with similar patterns regarding their statistical factors in the input domain may drive completely different investment decisions.

R3: The capacities of RES are decision variables in the generation investment planning problems. Without knowing the installed capacity of RES, directly clustering the input data, which usually includes the normalized load and the RES availability factor [2], may lead to less effective selection because the ratio of the bases (i.e., maximum values) between the loads and the RES capacities, which use normalized input data, cannot be determined before solving the investment problem. In other words, using the same normalized input data with different ratios of bases may result in different investment decisions.

R4: The impacts of ancillary service on system operation cannot be considered if clustering is performed in the input domain. In other words, the input data cannot depict the actual requirements of ancillary services for each time step, which can only be determined after solving the investment problem.

To this end, it becomes imperative to tackle the aforementioned issues by transforming the clustering domain (i.e., where we perform clustering based on) from the input space to a more effective space. For the transmission investment problem, the line benefit and the optimal power flow pattern have been demonstrated as effective clustering domains for selecting representative operating points in [9] and [10], respectively. However, they are still not the most straightforward drivers for the final optimal investment decision. In addition, for the periods with similar total or investment costs, a single cost value may consist of different system compositions. Therefore, clustering based on a single cost value may neglect the inter-spatial correlations between different generation technologies at various locations and thus resulting in a less effective clustering. Motivated by the fact that the overall optimal investment decision is fundamentally related to the investment decision for each operating period, in this research, a cost-oriented approach is proposed to transform the clustering domain from input periods to their corresponding investment costs of each generation technology at each location via running system investment planning for each period and then assigning the clusters based on the cluster label obtained via grouping the capital costs.

To highlight the motivations of the proposed cost-oriented approach and explain the aforementioned issues of the input-based method, Fig. 2 presents an example of the generation investment problem that aims to optimize the number of CCGTs (50 MW each) and wind generators (50 MW each) given the input data of load and wind availability factor, where the load data needs to be normalized to make the calculated distance used in the clustering methods place the same weight [2]. Let $L_{max} = 150$ MW denote the magnitude of the loads, the three

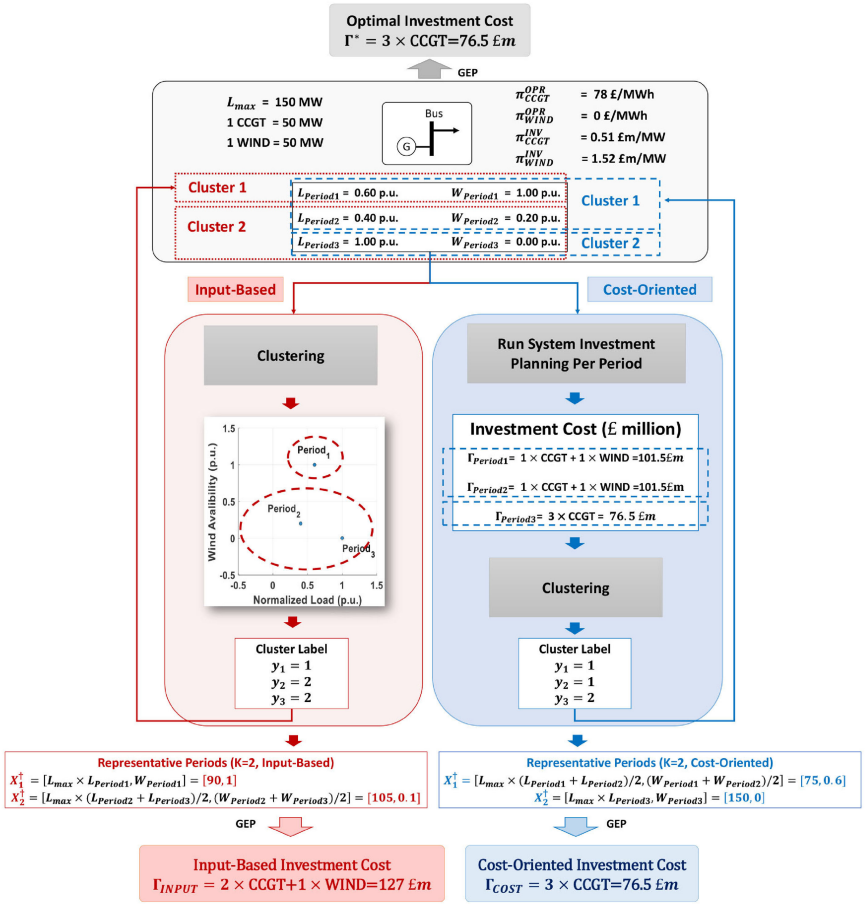


Fig. 2. An example of the input-based and cost-oriented representative period selection processes.

operating periods considered in this case are:

$$Period_1 = [L_{Period_1}, W_{Period_1}] = [0.6, 1.0], \quad (3)$$

$$Period_2 = [L_{Period_2}, W_{Period_2}] = [0.4, 0.2], \quad (4)$$

$$Period_3 = [L_{Period_3}, W_{Period_3}] = [1.0, 0.0]. \quad (5)$$

As can be seen in the left part of Fig. 2, the input-based approach will result $Period_2$ and $Period_3$ being assigned to the same cluster because they are statistically located near each other. More specifically, in the input domain, the Euclidean distances between each pair of periods are $d(Period_1, Period_2) = 0.8246$, $d(Period_1, Period_3) = 1.0770$ and $d(Period_2, Period_3) = 0.6325$, respectively. However, given the operational costs of Wind (0/MWh) and CCGT (78/MWh), the optimal investment costs for each period are:

$$\Gamma_{Period1} = \Gamma_{Period2} = \pi^{INV}_{CCGT} \times 1CCGT + \pi^{INV}_{WIND} \times 1WIND = 101.5£m \quad (6)$$

$$\Gamma_{Period3} = \pi^{INV}_{CCGT} \times 3CCGT = 76.5£m \quad (7)$$

Note that the investment costs of CCGT and Wind are given in Table II. According to these results, point **R1** can be well explained by the fact that significantly different input data $Period_1$ and $Period_2$ can result in the same investment decisions (i.e., $\Gamma_1 = \Gamma_2$). However, regarding point **R2**, the operating periods with similar patterns in the input domain (i.e., $Period_2$ and $Period_3$) can drive completely different investment decisions (i.e., $\Gamma_2 = 101.5$ million and $\Gamma_3 = 76.5$ million). Beyond that, it can be seen that when the capacities of RES (i.e., wind) are decision variables, the investment decisions of wind are not only determined by its availability factor (i.e., **R3**): $\Gamma_{Period1}$ and $\Gamma_{Period2}$ have identical investment decision determined based on the total cost although their wind availability factors are significantly different (i.e., $W_{Period_1} = 1$ and $W_{Period_2} = 0.2$). In other words, before the investment problem is solved, the proportion of the demand that can be supplied by the available wind generation is unknown and thus, rendering the input-based method ineffective when the capacities of RES have not been determined. To this end, using the proposed cost-oriented approach can effectively avoid this issue by directly grouping the operating periods based on their investment costs: as can be seen, $Period_1$ and $Period_2$ with same investment decisions are successfully clustered in the same group although their distance in the input domain is larger than that of $Period_2$ and $Period_3$.

Finally, when the number of selected periods is two ($K = 2$), based on the representative periods selected using the input-based approach, the investment decision is to build two

CCGTs and one Wind; however, the proposed cost-oriented method can lead to the real optimal investment decision to invest in three CCGTs, which is identical to the result with all three operating periods. The above example clearly illustrates the rationale and demonstrates the superior performance of the proposed cost-oriented approach in a simple case. More comprehensive analysis with complex system configurations will be presented in Section IV.

For the proposed framework shown in Fig. 1, this step aims to transform the clustering variables from the input domain to the domain of investment cost. The input operating condition data is denoted by $X = [X^L, X^W, X^P] = \{\vec{x}_d, d = 1, \ldots, D\} \in \mathbb{R}^{D\times[N_B\times(N_G+1)]}$, where X^L, X^W, and X^P represent the datasets of electricity load, wind availability and solar availability, respectively. The first step of the proposed framework is to run system investment planning for each individual day d. Note that day d is assumed to repeat across the whole horizon and the output of this step is the dataset of investment cost results $\Gamma = [\Gamma_1, \ldots, \Gamma_D]^T \in \mathbb{R}^{D\times(N_B\times N_G)}$, where $\Gamma_d = \{\gamma_d^{b,g}, b = 1...N_B, g = 1...N_G\}$ represents the investment cost of day d for generation technology g at bus b. Although D one-day-based investment planning problems need to be solved in this step, it is not computationally demanding as they can be effectively solved in parallel.

B. Dimensionality Reduction (Step 2)

In this step, dimensionality reduction is conducted to resolve the issue of the curse of dimensionality. As presented in Section II, two main challenges of representative day selection related to dimensionality can be summarized as **P-1:** *input operating days have 3 dimensions*: *d1*-variables $N_B \times (N_G + 1)$, *d2*-data points within the day N_d, and *d3*-days D; **P-3:** *limited number of days for clustering*. Regarding **P-1**, note that the proposed clustering domain transformation step can contribute to solving the problem by performing clustering on a 2-d dataset $\Gamma \in \mathbb{R}^{D\times(N_B\times N_G)}$ rather than a 3-d dataset $X \in \mathbb{R}^{D\times[N_B\times(N_G+1)]\times N_d}$. Nevertheless, for **P-3**, the large number of buses and varieties of candidate generation technologies in large-scale systems will lead to the curse of dimensionality, which refers to the problem caused by the exponential increase in volume associated with adding more dimensions to Euclidean space [17]. As illustrated in [18], the high dimensionality problem of input features will lead to the ineffective clustering results because of the unreliable similarity metrics in high dimensional space. One of the intuitive solutions is to transform data from high dimensional feature space to lower dimensional space in which to perform clustering. To this end, it is imperative to perform dimensionality reduction on the domain of investment costs, which already has lower-dimensional clustering variables than the conventional input-based method.

In general, dimensionality reduction can be achieved using two types of methods: *feature extraction* and *feature selection*. In this paper, the considered clustering variables of the proposed cost-oriented method are the investment cost of each technology for each location. Therefore, it is ppropriate to extract important features from the clustering variables in an automatic or nearly automatic manner as it is challenging to manually determine the most influential variables on the final optimal decisions.

Conventional linear dimensionality reduction (DR) techniques (e.g., PCA and linear discriminant analysis) have been widely used but with a performance limitation due to the linear transformation. To this end, a series of nonlinear DR techniques have been proposed, such as kernel PCA, Kohonen self-organizing maps, data-driven high-dimensional scaling (DD-HDS), and Laplacian Eigenmaps (LEM) [19]. The limited number of data samples (i.e., days) in our case restricts the performance of the techniques that require a large amount of data, such as neural network (NN)-based approaches and DD-HDS. Therefore, we select a geometrically motivated algorithm, LEM, which has locality-preserving properties and a natural connection to clustering. The constructed lower-dimensional data can reflect the intrinsic geometric structure of the manifold. Mathematically, the dimensionality reduction procedure for LEM can be illustrated as follows.

Given the input dataset of investment costs for each day $\Gamma = \{\Gamma_d\}_{d=1}^D \in \mathbb{R}^{D\times(N_B\times N_G)}$ and the target dimension r, the first step is to construct the adjacency graph $Q = (N, E)$, where $\Gamma_i, i \in [1, \ldots, D]$ corresponds to one node in $n_i \in N$, the total number of nodes $|N| = D$. A pair of nodes n_i and n_j are connected by an edge if $\Gamma_i, i \in \{1, \ldots, D\}$ and $\Gamma_j, j \in \{1, \ldots, D\}$ are close to each other. The "closeness" is measured using the k-nearest neighbor (KNN) method.

Then, we determine the weights $W = \{w_{i,j}, i, j = 1, \ldots, D\}$ of the constructed edges $E = \{E_{i,j}, i, j = 1, \ldots, D\}$ using a simple method as follows:

$$w_{i,j} = \begin{cases} 1, \text{if } n_i \text{ and } n_j \text{ are connected via edge } E_{i,j} \\ 0, \text{if } n_i \text{ and } n_j \text{ are not connected} \end{cases} \tag{8}$$

The next step is to solve the generalized eigenvector problem: $L\alpha = \lambda D\alpha$, where $D = \{D_{i,i} = \sum_j w_{j,i}, \forall i, j \in 1, \ldots, D\}$ is the diagonal weight matrix and $L = D - W$ is the Laplacian matrix. Given the target reduced dimensions r, the solution vectors $A = [\alpha_0, \ldots, \alpha_{r-1}] \in \mathbb{R}^{D\times r}$, which are ordered based on their eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq ...\lambda_{r-1}$. Finally, the lower-dimensional output data $\widetilde{\Gamma} = \{\widetilde{\Gamma}_i, \forall i \in 1, \ldots, D\} \in \mathbb{R}^{D\times r}$, where we have

$$\widetilde{\Gamma}_i = (\alpha_0(i), \ldots, \alpha_r(i)) \in \mathbb{R}^r \tag{9}$$

C. Cluster Assignment (Step 3)

This step is to cluster the investment decisions in a lower-dimensional space. Based on the extracted features of investment costs $\widetilde{\Gamma}$, the clustering techniques can be applied to construct K groups $\widetilde{\Gamma}^k \subset \widetilde{\Gamma}$, for $k = 1, \ldots, K$, which aims to distinguish different investment costs. The output of the clustering procedure will be the set of cluster labels $y \in \mathbb{R}^D$, which can be employed to assign the input operating days $\vec{x}_d, d = 1, \ldots, D$ into different groups

$$X_{cls} = \{X_k\}_{k=1}^K \tag{10}$$

where $X_k \in \mathbb{R}^{N_k\times[N_B\times(N_G+1)]\times N_d}$, according to their individual investment costs.

As one of the most prevalent clustering techniques, hierarchical clustering can construct a hierarchy of clusters by employing a measure of similarity between groups of data points [20], [21]. In this research, we employ the agglomerative hierarchical clustering method (bottom-up approach) with Ward's linkage to establish different groups of investment decisions and identify the representative day from each cluster for the following reasons:

i) In terms of the shape of the constructed clusters, the hierarchical clustering method can handle nonspherical data;

ii) Regarding repeatability, the constructed hierarchical clusters have a deterministic nature because they are independent of the initial allocation of data points;

iii) No prior knowledge of the number of clusters is required for hierarchical clustering. In other words, we can terminate the agglomeration procedure at any number of clusters as required;

iv) In contrast to other types of linkages (e.g., single-linkage and complete-linkage), Ward's minimum variance criterion [22] aims to minimize the total within-cluster variance. For the proposed cost-oriented representative day selection algorithm, it is important to ensure that the variance of grouped investment costs in each cluster can be minimized to identify the operating conditions in the input domain that result in similar investment decisions.

In general, hierarchical clustering can be outlined in the following steps based on our clustering variables $\widetilde{\Gamma} \in \mathbb{R}^{D\times r}$. First, each data point of $\widetilde{\Gamma}$ is assigned to its own singleton group. Then we construct the similarity matrix

$$S = \{s_{i,j}, \forall i, j \in 1, \dots D\} \in \mathbb{R}^{D\times D} \tag{11}$$

for $\widetilde{\Gamma}$ based on the Euclidean distance. Consequently, each pair of clusters that are closest to each other will be merged to a higher level according to the calculated similarity. Note that, in this research, Ward's linkage criterion is employed to measure the intergroup similarity. For a pair of clusters k_1 and k_2, the distance measure d_{k_1,k_2} can be calculated as follows:

$$d_{k_1,k_2} = \|\widetilde{\Gamma}_c^{k_1} - \widetilde{\Gamma}_c^{k_2}\|_2 \sqrt{2n_{k_1}n_{k_2}/(n_{k_1}+n_{k_2})} \tag{12}$$

where n_{k_1} and n_{k_2} are the numbers of operating days in clusters k_1 and k_2, $\widetilde{\Gamma}_c^{k_1}$ and $\widetilde{\Gamma}_c^{k_2}$ represent the centroids of clusters k_1 and k_2, and $\|\cdot\|_2$ is Euclidean distance.

D. Representative Day Selection (Step 4)

Finally, each cluster needs to be represented by one operating period selected or created from the cluster. The most widely used representatives are the mean point as the average value or the medoid point as the period closest to the mean point. Due to the domain transformation in *Step 1*, the mean point cannot be linked back to any real operating period. Therefore, the medoid point $\vec{\gamma}_k^{\dagger} \in \mathbb{R}^{N_B\times N_G}$ of cluster k is selected in the domain of $\widetilde{\Gamma}$ and then transformed back to the input domain of X to obtain the representative operating condition data $\vec{x}_k^{\dagger} \in \mathbb{R}^{N_B\times(N_G+1)}$.

Given that each operating period has the same probability of occurrence, the weight of each cluster can be calculated as the number of operating periods that belong to the cluster. The final outputs of the proposed framework are the selected representative days and their corresponding weight defined as $X^{\dagger} = \{\vec{x}_k^{\dagger}, k = 1, \dots, K\}$ and $\Psi^{\dagger} = \{\psi_k^{\dagger}, k = 1, \dots, K\}$, respectively.

To summarize, the proposed cost-oriented representative day selection method is outlined in Algorithm 1.

Algorithm 1: Proposed Cost-Oriented Representative Day Selection Method for the Generation Investment Planning Problem.

Input: Multidimensional historical data of demand, wind availability and solar availability: $X = [X^L, X^W, X^P] = \{\vec{x}_d, d = 1, \dots, D\}$; Target dimension: r; Number of selected representative days: K; The tested system investment planning model: $Planning(\cdot)$.

Output: Set of selected representative days: $X^{\dagger} = \{\vec{x}_k^{\dagger}, k = 1, \dots, K\}$; Set of corresponding probabilities: $\Psi^{\dagger} = \{\psi_k^{\dagger}, k = 1, \dots, K\}$.

Step 1: Run system planning for each operating day and obtain the corresponding investment cost.

$\Gamma_d = Planning(\vec{x}_d)$, for $d = 1, \dots, D$.

Step 2: Given the input target dimension r, LEM is performed to reduce the dimensionality of Γ to r.

$\widetilde{\Gamma} = LEM(\Gamma)$.

Step 3: Given the number of selected representative days K, hierarchical clustering is performed to construct the groups of $\widetilde{\Gamma}$. Then map the constructed clusters from the cost domain of $\widetilde{\Gamma}^k$ to the input domain of X_k.

$[\widetilde{\Gamma}^k|_{k=1}^K, \Lambda_{D,k}|_{k=1}^K] = HierachicalClustering(\widetilde{\Gamma}, K)$.

$X_k = \{\vec{X}_d, \forall d \in \Lambda_{D,k}\}$, for $k = 1, \dots, K$.

Step 4: Determine the representative day and the corresponding probability for each cluster. Note that the output of function $medoid$ is the index of the day that indicates the medoid point of a dataset.

$idx_k^{med} = medoid(\widetilde{\Gamma}^k)$, for $k = 1, \dots, K$

$\vec{x}_k^{\dagger} = X_k(idx_k^{med})$, for $k = 1, \dots, K$

$X^{\dagger} = \{\vec{x}_k^{\dagger}, k = 1, \dots, K\}$

$\Psi^{\dagger} = \{\psi_k^{\dagger} = |\Lambda_{D,k}| / |\Lambda_D|, k = 1, \dots, K\}$.

IV. Simulation Study and Results Analysis

A. Test Model and System Description

To demonstrate the performance of the proposed representative day selection algorithm, the electricity investment model presented in [23], in which various constraints are taken into consideration including electricity balance constraints, generation operation constraints, network reinforcement constraints, power flow constraints, security constraints, ancillary service constraints and carbon constraints, while minimizing the total system cost, is applied in this paper. A simplified GB transmission system characterized by four key regions, including 1) Scotland, 2) North England & Wales, 3) Middle England & Wales, and 4) South England & Wales is employed for the

TABLE I
PARAMETER VALUES OF TEST MODEL

	Capital Cost (£m/MW)	RampUp/Down (%/h)	MinUp/Down (h)
Nuclear	4.34	0.10	10
CCGT	0.51	0.60	4
Gas-CCS	2.15	0.50	4
OCGT	0.32	1.00	1
Wind	1.52	-	-
PV	0.67	-	-

TABLE II
TEST MODELS WITH DIFFERENT COMPLEXITIES

	M1	M2	M3
RES	fixed	non-fixed	non-fixed
AS, Ramp, MinOn/Off	×	×	√

simulation. Different types of conventional generation (i.e. CCGT and OCGT) and various low-carbon generation (i.e., nuclear, gas CCS, wind, and PV) are taken into account. Table I summarizes the technical and economic data of each technology. In addition, all operational and investment data related to the electricity system are given in [24]. Hourly and half-hourly electricity demand data, wind and solar power generation output data in different regions are obtained from the Open Power System Data (OPSD) project [25]. Note that the collected demand data need to be scaled to the corresponding level according to the local population.

As presented in Table II, three cases of generation investment planning with different levels of complexity are considered. Specifically, '**RES**' indicates whether the model considers wind and PV generation capacities as fixed values or as decision variables in the optimization model. In this work, M1 uses a fixed capacity of RES, and the others consider the capacity of RES as decision variables. In addition, M3 includes the ramp up/down constraints, minimum online/offline time constraints as well as ancillary service requirements, indicated by '**Ramp**', '**MinOn/Off**' and '**AS**', respectively. Note that for simplicity, we consider the planning from scratch and the investment problem is relaxed to a LP problem with continuous investment decisions. As demonstrated in [26], the relaxed LP problem can provide very similar decision as the original MILP problem in the case of clustered representation of generation units. The total number of continuous variables and constraints are 3,529,800 and 2,505,850, respectively. Additionally, the considered planning approach is static.

B. Tested Methods

In this paper, for each of the aforementioned models, '**COST**' refers to the proposed cost-oriented representative day selection method and '**INPUT**' denotes the state-of-the-art input-based method proposed in [2], which can be briefly described as follows:

Step 1: Reshape the input operating condition data. For day d, load data, wind availability data, and solar availability data are represented as $X_d^L = \{x_{b,d,1}^L, \ldots, x_{b,d,24}^L, \forall b = 1, \ldots, |B|\}$, $X_d^W = \{x_{b,d,1}^W, \ldots, x_{b,d,24}^W, \forall b = 1, \ldots, |B|\}$, and $X_d^P = \{x_{b,d,1}^P, \ldots, x_{b,d,24}^P, \forall b = 1, \ldots, |B|\}$, respectively;

TABLE III
GEP SOLUTION BENCHMARK: M1

	Operational Cost (£million/year)	Investment Cost (£million/year)
All days	2340.07	1230.86
	Total Cost (£million/year)	CPU Times(s)
All days	3570.94	899.62

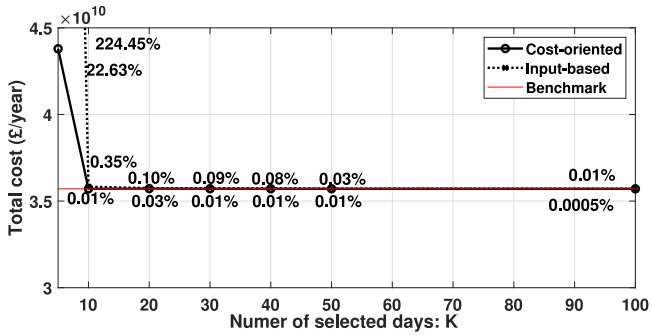


Fig. 3. Comparison of tested methods with different numbers of K based on GEP solutions of total cost (M1).

Step 2: Construct the clustering variables in the input domain $X_{input} = \{[X_d^L, X_d^W, X_d^P], \forall d = 1, \ldots, D\}$;

Step 3: Hierarchical clustering method is employed to group the days based on X_{input};

Step 4: The medoid point of each constructed cluster is selected as the representative day with corresponding probability.

Detailed information of the 'INPUT' method can be found in [2]. Note that the tested representative day selection methods and the investment planning optimization problem were implemented in MATLAB 2017a and FICO Xpress, respectively, and run on an Intel Xeon E5-2690 PC with 8 cores.

C. Performance Evaluation Across Different Models and Time Resolutions

1) M1: As shown in Table II, M1 is designed to be the simplest generation investment planning model that does not include any intertemporal constraints; at the same time, the RES capacities are assumed to be fixed. In particular, wind generators built in regions 1 and 4 (i.e., WIND1 and WIND4) are set to 18 GW and 10 GW, respectively. Additionally, solar generators built in regions 3 and 4 (i.e., PV3 and PV4) are set to 2 GW and 9 GW. First, Table III presents the benchmark solution of M1 when considering all days (i.e. $K = 365$).

The performance of the tested methods across different numbers of selected days is illustrated in Fig. 3, indicated by the percentage error of the total cost between the benchmark and the estimated results for $K = [5, 10, 20, 30, 40, 50, 100]$. As shown, both the input-based and cost-oriented methods can approach the benchmark solution after $K = 10$ with extremely low errors (i.e., $\leq 1\%$). Additionally, the proposed cost-oriented method exhibits slightly better performance than the input-based method for most numbers of K.

TABLE IV
CPU TIMES(S): M1

	K=5	K=10	K=20	K=30	K=40	K=50	K=100
COST	2.10	2.12	4.35	7.08	11.78	17.58	70.74
Input	1.78	2.22	3.68	6.84	11.69	15.75	65.19

TABLE V
GEP SOLUTION BENCHMARK: M2

	Operational Cost (£million/year)	Investment Cost (£million/year)
All days	1666.95	1712.44
	Total Cost (£million/year)	CPU Times(s)
All days	3379.39	4325.83

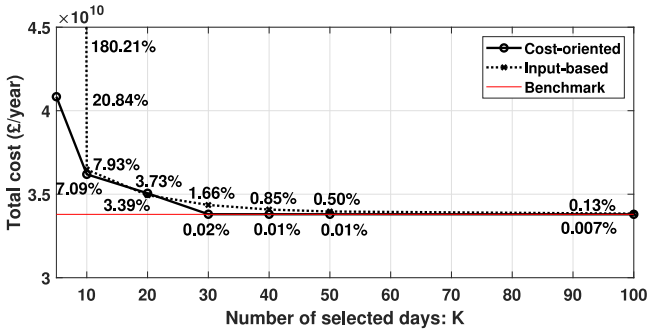


Fig. 4. Comparison of tested methods with different numbers of K based on generation investment planning solutions of total cost (M2).

Regarding the computational cost of M1, the simulation times of each number of selected days are presented in Table IV. Compared with the CPU time for all days, the day selection approach leads to an approximately 99.76% reduction in CPU time while obtaining very accurate investment decisions. It is imperative to note that the CPU time of the COST method includes the entire selection and optimization process under the assumption that the investment problem for each day can be solved in parallel.

2) M2: In the context of fixed capacities of RES, both input-based and cost-oriented methods present considerable performance in terms of the required number of selected days to approach the optimal investment decisions. However, when considering RES as decision variables that need to be optimized by solving the generation investment planning problem, it becomes inefficient to perform clustering based on RES availability data because the accurate proportion of RES during clustering cannot be predefined. In contrast, the proposed method overcomes this challenge by clustering based on the investment costs driven by each individual day.

To demonstrate the aforementioned points, the input-based and cost-oriented methods are conducted on M2 for different numbers of clusters. The results of the benchmark case that considers all the operating conditions for 365 days are presented in Table V. In addition, the estimated total costs and the percentage errors between the estimated and benchmark values are shown in Fig. 4. It can be seen that, for both methods, a sustained decline in the estimated total cost is observed with increasing number of selected days. However, the superior performance of the proposed cost-oriented method can be indicated regarding the required number of K to achieve the benchmark value. Specifically, for the input-based method, the calculated percentage error can be reduced from 180.21% to 0.13% when K increases from 5 to 100, which is still approximately 6.5 times greater than that of the cost-oriented method when $K = 30$. In other words, for the proposed cost-oriented method, the estimated total cost tends to converge to the benchmark solution after $K = 30$ with significantly low total error (i.e., $\leq 0.02\%$), whereas the input-based method can achieve a relatively accurate result when $K = 100$ but still with $e = 0.13\%$.

TABLE VI
CPU TIMES(S): M2

	K=5	K=10	K=20	K=30	K=40	K=50	K=100
COST	11.46	14.26	25.40	40.26	60.72	92.92	368.01
INPUT	11.76	13.84	25.58	35.69	55.57	83.47	353.12

TABLE VII
GEP SOLUTION BENCHMARK: M3

	Operational Cost (£million/year)	Investment Cost (£million/year)
All days	2179.79	1320.61
	Total Cost (£million/year)	CPU Times(s)
All days	3500.40	95460.13

In addition, the computational times for different numbers of selected days are given in Table VI. Compared with the CPU time for all days (i.e., 4325.83 s), significant reductions in computational burden can be achieved by solving the generation investment planning problem based on a reduced number of representative days. For example, it only takes 40.26 seconds to obtain an accurate investment plan that can approach the optimal total cost with only approximately 0.02% error, achieving an approximately 99.7% reduction in computational cost.

3) M3: To further complete the investment planning model, ancillary services, ramp constraints and minimum online/offline time constraints are included in M3. In this case, the considered intertemporal operating constraints introduce difficulties in selecting the representative days that can retain the original temporal autocorrelations. Nevertheless, the proposed cost-oriented method can prevent the selection procedure from addressing this issue by directly considering the information extracted from their corresponding investment decisions for each day. The benchmark solutions and CPU times for M3 are shown in Table VII.

Under different day selection methods for M3, Fig. 5 summarizes the estimated total costs and percentage errors for different numbers of selected days K, ranging from 5 to 100. The results indicate that, for model M3 with intertemporal and ancillary service constraints, the proposed cost-oriented method exhibits a more outstanding performance than for M2. This result is evidenced by the fact that the differences in the total cost error between the cost-oriented and input-based methods are larger than those of M2 for most numbers of K. Additionally, it is constructive to highlight that, for COST, only 20 representative days

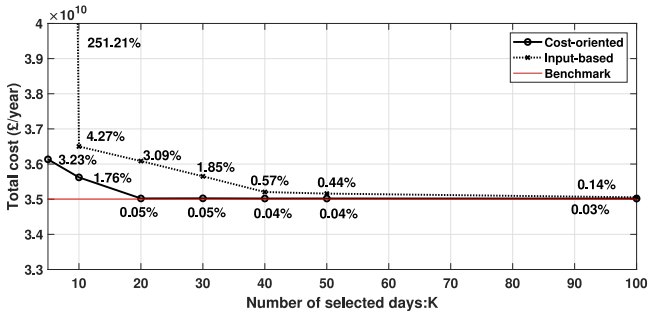


Fig. 5. Comparison of tested methods with different numbers of K based on generation investment planning solutions of total cost (M3).

TABLE VIII
INVESTMENT DECISIONS (MW) AND DECISIONS ERRORS (MW) FOR EACH GENERATION TECHNOLOGY AND NRMSE (%)

	Benchmark	COST	INPUT	ϵ_{COST}	ϵ_{INPUT}
CCGT	41.34	41.52	33.94	0.18	7.4
OCGT	28.13	28.08	34.32	0.05	6.19
WIND	24.37	22.78	39.61	1.59	15.24
PV	36.77	35.84	37.45	0.93	0.68
NRMSE	-	2.11%	22.59%	-	-

TABLE IX
CPU TIMES(S): M3

	K=5	K=10	K=20	K=30	K=40	K=50	K=100
COST	15.43	47.73	141.42	369.92	684.61	1098.32	8419.37
INPUT	14.98	48.31	136.39	357.62	659.24	997.45	8143.58

are required to be selected to achieve the benchmark solution with approximately 0.05% error. Nevertheless, the input-based method cannot achieve such a low error even when $K = 100$.

More specifically, considering the case of $K = 20$, the estimated investment decisions and corresponding decision errors for each technology as well as the normalized root mean square error (NRMSE) between the benchmark and the estimated total cost are shown in Table VIII. Note that the results of Nuclear and Gas-CCS are not presented in this table because they are not chosen to be built in this case. Regarding the investment decisions for each technology, OBJ presents significantly lower decision errors for most of the generation technologies, except for a slightly larger error (i.e., 0.93 MW -0.68 MW $= 0.25$ MW) for PV. Additionally, it is important to highlight that the superior overall performance of the proposed method can be indicated by the approximately ten times lower NRMSE value when using the COST method (i.e., NRMSE $= 2.11\%$) rather than the Input method (i.e., NRMSE $= 22.59\%$).

Table IX presents the CPU times for solving the generation investment planning model of M3 based on the selected days obtained via COST and INPUT, respectively, for different numbers of K. It is important to highlight that, with the increasing level of model complexity, solving the planning problem is more time consuming and the CPU times increase exponentially with an increasing number of K. However, it can be observed that when employing the proposed cost-oriented method, the planning problem only needs to be solved based on 20 representative days, which reduces the CPU times from 95,460.13 s in the full case to 141.42 s with less than 0.05% error. This result

TABLE X
RESULTS OF M3 + HALF-HOURLY DATA

	NRMSE of Investment Decisions(%)	Total Cost Error(%)	CPU Times (s)
All Days	-	-	2.98E5
COST	3.23%	0.17%	157.58
Input	23.35%	2.57%	211.24

demonstrates the increasing benefit of optimal representative day selection for more complex generation investment planning models. It is imperative to note that, as the proposed COST approach requires to perform system investment planning for each day within a year, if the investment problem is extremely complex so that it is not possible to perform this task within acceptable times for system planners with limited computational resources, the "per-day-investment-problem" can be somehow relaxed to make the clustering variable construction procedure tractable. In addition, system planners can also employ high performance computing techniques and cloud computing services (e.g., Amazon Web Services or Google Cloud) to solve the per-day-investment-problem in parallel with sufficient computational resources.

4) Performance Evaluation For Different Data Resolutions: The influx of high-resolution measurements renders it more challenging to select the representative days, particularly for input-based methods, due to the issues of high variability and dimensionality. Nevertheless, the proposed cost-oriented method does not need to address these issues because the selection procedure is performed in the domain of investment costs, whose dimensions and variabilities are not directly dependent on the input operating condition data. To demonstrate this point, based on M3, we employ higher-resolution data with a 30-minute time interval as the input data for M3. The previous numerical testing indicates that a minimum of 20 representative days is required to obtain an accurate solution (e.g., NRMSE$< 0.2\%$) for COST. Consequently, the calculated NRMSE of the estimated investment decisions, the total cost errors, and CPU times are shown in Table X for $K = 20$. The challenges of higher-resolution data can be illustrated by the increased total cost error and the calculated NRMSE for both the COST and Input methods. Nevertheless, the proposed COST method still presents the best performance for this high-resolution case. The extremely high computing time 2.98×10^5 when considering all 365 days emphasizes the importance of selecting a subset of representative days, particularly for the case of high-resolution input data.

5) Performance Evaluation Across Different r: In order to evaluate the effectiveness of the proposed dimensionality reduction stage, for the high-dimensional case with higher resolution input data (30 min), we compare the estimated total costs of the proposed cost-oriented approach with and without using LEM to perform dimensionality reduction in the context of $K = 10$, which exhibits higher error than that of $K = 20$. Note that the original dimension of the clustering variables in the cost domain is 15 (i.e., $r_{original} = 15$) because 9 variables of investment costs with all zeros have been removed from the original 24-dimensional dataset. Fig. 6 shows the total cost errors for $K = 10$ across different numbers of reduced dimensions $r = [1, 3, 6, 10, 15]$. It can be seen that the proposed Dimensionality

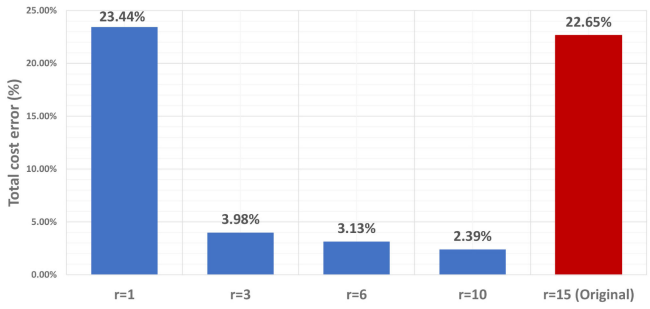


Fig. 6. Performance evaluation across different r (cost-oriented approach).

Reduction Stage can effectively enhance the performance of the proposed objective-based method with the estimated total cost error reduced from 22.65% ($r_{original} = 15$) to 2.39% ($r_{reduced} = 10$) even though only $K = 10$ representative days are considered.

V. Conclusion

This paper proposes a novel cost-oriented representative day selection method that includes four main stages: clustering domain transformation, dimensionality reduction, cluster assignment, and representative day selection. In the clustering domain transformation stage, we aim to obtain the dataset to perform clustering in the objective domain, which consists of the investment costs of each technology across different locations for each individual day. Dimensionality reduction aims to address the issue of high-dimensionality and to enable the clustering procedure to be performed in a more effective domain that is constructed with important features. Hierarchical clustering method with Ward's linkage criterion is employed to group the days based on the associated investment costs. Finally, the medoid point of each constructed cluster is selected as the representative day. The superior performance of the proposed method is demonstrated based on a GB electricity system. The tested generation investment planning problems with different levels of complexity are designed to illustrate the increasing advantages of the proposed method over the conventional input-based method. Finally, the effectiveness of the proposed dimensionality reduction stage is demonstrated through sensitivity analysis.

Future research could be devoted to further developing the proposed framework for selecting longer operating periods to deal with the investment models with interday or seasonal energy storage. Furthermore, it would be useful to investigate the expansion of the proposed framework for multi-stage investment problems. Beyond the generation investment problem, the development of the cost-oriented approach for generation and transmission investment problems with large-scale system is also of significant interest.

References

[1] B. Hua, R. Baldick, and J. Wang, "Representing operational flexibility in generation expansion planning through convex relaxation of unit commitment," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2272–2281, Mar. 2018.

[2] Y. Liu, R. Sioshansi, and A. J. Conejo, "Hierarchical clustering to find representative operating periods for capacity-expansion modeling," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3029–3039, May 2018.

[3] H. Saboori and R. Hemmati, "Considering carbon capture and storage in electricity generation expansion planning," *IEEE Trans. Sustain. Energy*, vol. 7, no. 4, pp. 1371–1378, Oct. 2016.

[4] K. Eurek *et al.*, "Regional energy deployment system (reeds) model documentation: Version 2016." (National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-6A20-67067, pp. 1–101, 2016.

[5] K. Poncelet, H. Hschle, E. Delarue, A. Virag, and W. Dhaeseleer, "Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 1936–1948, May 2017.

[6] F. J. Sisternes and M. D. Webster, "Optimal selection of sample weeks for approximating the net load in generation planning problems," eSD Working Paper, Massachusetts Institute of Technology. 2013. [Online]. Available: https://dspace.mit.edu/handle/1721.1/102959

[7] L. Baringo and A. Conejo, "Correlated wind-power production and electric load scenarios for investment decisions," *Appl. Energy*, vol. 101, pp. 475–482, Jan. 2013.

[8] R. Domnguez, A. J. Conejo, and M. Carrin, "Toward fully renewable electric energy systems," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 316–326, Jan. 2015.

[9] D. Z. Fitiwi, F. de Cuadra, L. Olmos, and M. Rivier, "A new approach of clustering operational states for power network expansion planning problems dealing with res (renewable energy source) generation operational variability and uncertainty," *Energy*, vol. 90, pp. 1360–1376, Oct. 2015.

[10] Q. Ploussard, L. Olmos, and A. Ramos, "An operational state aggregation technique for transmission expansion planning based on line benefits," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2744–2755, Jul. 2017.

[11] R. Alvarez, A. Moser, and C. A. Rahmann, "Novel methodology for selecting representative operating points for the TNEP," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2234–2242, May 2017.

[12] M. Sun, F. Teng, I. Konstantelos, and G. Strbac, "An objective-based scenario selection method for transmission network expansion planning with multivariate stochasticity in load and renewable energy sources," *Energy*, vol. 145, pp. 871–885, Feb. 2018.

[13] S. Pineda and A. J. Conejo, "Scenario reduction for risk-averse electricity trading," *IET Gener., Transmiss. Distrib.*, vol. 4, no. 6, pp. 694–705, Jun. 2010.

[14] J. M. Morales, S. Pineda, A. J. Conejo, and M. Carrion, "Scenario reduction for futures market trading in electricity markets," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 878–888, May 2009.

[15] S. Pineda and J. M. Morales, "Chronological time-period clustering for optimal capacity expansion planning with storage," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7162–7170, Nov. 2018.

[16] D. A. Tejada-Arango, M. Domeshek, S. Wogrin, and E. Centeno, "Enhanced representative days and system states modeling for energy storage investment analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6534–6544, Nov. 2018.

[17] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2011, pp. 257–258.

[18] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1753–1759.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[20] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.

[21] M. Sun, Y. Wang, G. Strbac, and C. Kang, "Probabilistic peak load estimation in smart cities using smart meter data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1608–1618, Feb. 2019.

[22] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *J. Am. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, Apr. 1963.

[23] D. Pudjianto, M. Aunedi, P. Djapic, and G. Strbac, "Whole-systems assessment of the value of energy storage in low-carbon electricity systems," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 1098–1109, Mar. 2014.

[24] X. Zhang, G. Strbac, N. Shah, F. Teng, and D. Pudjianto, "Whole-system assessment of the benefits of integrated electricity and heat system," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 1132–1145, Jan. 2019.

[25] OPSD. Data package time series. 2017. [Online]. Available: https://data.open-power-system-data.org/time_series/2017-07-09/
[26] L. Zhang, T. Capuder, and P. Mancarella, "Unified unit commitment formulation and fast multi-service lp model for flexibility evaluation in sustainable power systems," *IEEE Trans. Sustain. Energy*, vol. 7, no. 2, pp. 658–671, Apr. 2016.

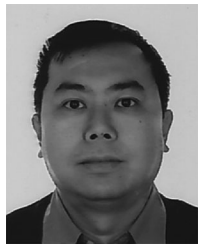
Mingyang Sun (M'16) received the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2017.

He is currently a Research Associate with Imperial College London. His current research interests include big data analytics and artificial intelligence in energy systems.

Fei Teng (M'15) received the bachelor's degree from Beihang University, Beijing, China, in 2009, and the Ph.D. degree from Imperial College London, London, U.K., in 2015. He is currently a Lecturer with the Control and Power Group, Imperial College London. His research interests include power system control and operation, system flexibility, and stochastic optimization.

Xi Zhang (S'17) received the bachelor's degree and master's degree in electrical engineering from Tsinghua University, Beijing, China, in 2012 and 2014, respectively. He is currently working toward the Ph.D. degree with Imperial College London, London, U.K. His research interests include whole-energy system planning, and multi-vector energy system modeling.

Goran Strbac (M'95) is a Professor of electrical energy systems with Imperial College London, London, U.K. His current research interests include electricity generation, transmission and distribution operation, planning and pricing, and integration of renewable, and distributed generation in electricity systems.



Danny Pudjianto (M'98) received the B.Sc. degree from Institut Teknologi 10 Nopember, Surabaya, Indonesia, in 1996, and the master's and Ph.D. degrees from The University of Manchester Institute of Science and Technology, Manchester, U.K., in 1999 and 2003, respectively. He is currently a Research Fellow with Imperial College London, London, U.K.