On representative day selection for capacity expansion planning of power systems under extreme events

Can Li[a], Antonio J. Conejo[b,c], John D. Siirola[d], Ignacio E. Grossmann[a,*]

[a] *Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA*
[b] *Department of Integrated Systems Engineering, The Ohio State University, 1971 Neil Avenue, Columbus, OH 43210, USA*
[c] *Department of Electrical and Computer Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210, USA*
[d] *Center for Computing Research, Sandia National Laboratories, P.O. 5800, Albuquerque, NM, 87185, USA*

Abstract

Capacity expansion planning (CEP) of power systems determines the optimal future generation mix and/or transmission lines. Due to the increasing penetration of renewables, CEP has to capture the hourly variations of renewable generator outputs and load demand. Since CEP problems typically involve planning horizons of several years, solving the fullspace models where the operating decisions corresponding to all the days is intractable. Therefore, some "representative days" are selected as a surrogate to the fullspace model. We present an input-based and a cost-based approach in combination with the k-means and the k-medoids clustering algorithms for representative day selection. The mathematical properties of the proposed algorithms are analyzed, including an approach to calculate the "optimality gap" of the investment decisions obtained from the representative day model to the fullspace model, and the relationship between the clustering error and the optimality gap. To capture the extreme events, two novel approaches, i.e., a "load shedding cost" approach and a "highest cost" approach, are proposed to identify the "extreme days". We conclude with a case study based on the Electric Reliability Council of Texas (ERCOT)

[*] Corresponding author
Email addresses: canl1@andrew.cmu.edu (Can Li), conejo.1@osu.edu (Antonio J. Conejo), jdsiiro@sandia.gov (John D. Siirola), grossmann@cmu.edu (Ignacio E. Grossmann)

region, which compares the different approaches and the effects of adding the extreme days.

Keywords: Power Systems, Capacity Expansion Planning, Renewable Units, Representative Day Selection, Clustering Algorithm, Extreme Events

1. Introduction

Capacity expansion planning (CEP) models [1, 2, 3] that have been extensively used by power system operators and planners aim to determine the location, size, and type of the generating units and/or transmission lines that should be installed in order to meet the electricity demand within a given geographical region. These investment decisions in generating units and transmission lines are long-term decisions usually made on a yearly basis. However, due to the increase in the penetration of generation for renewable resources, CEP models developed recently [4, 5] incorporate the hourly operating decisions in order to capture the high variations of renewable generation. Operating decisions including unit commitment and ramping decisions, and economic dispatch need to be included in CEP models, which leads to large-scale mixed-integer linear programming (MILP) problems.

One of the challenges in CEP models that include operating decisions is the temporal complexity. Modeling each hour in a 10-30 year planning horizon will lead to MILP problems with billions of variables, which is intractable with the current commercial solvers. Different simplifications have been proposed to make the CEP models tractable. Mallapragada et al. [6] proposes a time-slice model where the authors average the load and the capacity factor data of each of the four seasons (spring, summer, fall, winter) into time slices representing morning (7 am-2 pm), afternoon (2-6 pm), evening (6-11 pm), and night (11 pm-7 am). This approach does not link consecutive periods and thus fails to characterize the chronology of the operating decisions. In most of the literature, a "representative days" approach is used [7, 8, 9, 10, 11, 6, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. A dataset is given that consists of the historical load data and capacity factor data for solar and wind generators. The historical data is then scaled to account for load growth. Based on this scaled dataset, k representative days are selected in each year of the planning problem to represent the whole planning horizon where $k \ll 365$. A clustering algorithm, such as k-means clustering or k-medoids clustering, is used to divide the whole historical dataset into k clusters based on some of the characteristics of the historical days. The centroid or the medoid of each cluster is selected as the representative day. Several issues could arise in the representative day selection procedure. We summarize these issues below and discuss how the existing literature addresses them.

I *What type of data should be used for clustering?*

Most papers perform clustering based on the input parameters to the CEP models [7, 8, 9, 10, 11, 6, 12, 13, 14, 15, 16, 18, 19, 20, 21], i.e., the load data and capacity factors corresponding to each historical day. More specifically, each input data point to the clustering algorithm is a concatenation of the load and the capacity factor time series corresponding to a given historical day. As a result, the temporal correlations of different time series and the hourly chronology of each time series can be preserved. Since the numerical values of the load and the capacity factors can vary significantly, several normalization approaches have been proposed [18]. Besides clustering based on input parameters, [22, 17] propose to perform clustering based on the optimal objective value or the optimal solution of each historical day. In these approaches, a CEP model has to be solved to optimality for each historical day. The case study in [17] shows that this cost-based approach has generally a superior performance than the input-based approach.

II *Which clustering algorithm should be used?*

Different clustering algorithms have been used including k-means clustering [7, 8, 10, 13, 18, 19, 11, 6, 14], k-medoids clustering [8, 18, 9, 14], hierarchical clustering [8, 12, 13, 17, 18, 11], DBA clustering [18], k-shape clustering [18], self-organizing map (SOM) clustering [20]. Most of the papers that we surveyed conclude that there is no clear winner among all the different clustering algorithms [8, 18].

III *How to choose the appropriate number of clusters?*

There is no standard method to choose the number of clusters *a priori*. [8, 18, 6, 17, 10, 12] adopt a trial-and-error approach to choose the appropriate number of representative days by gradually increasing the number of representative days and observe if the optimal objective value and the optimal solution stabilize. Other methods are solely based on the metrics on the input time series without solving any optimization problem. For example, [7] use an "elbow method" that determines how well the k clusters can characterize the variance of the input data. [23, 20] use metrics like average normalized root mean square error (NRMSE) for the time series data.

IV *Should additional "extreme days" be included and how to find these "extreme days"?*

Including k representative days in the CEP model alone may not be enough since the representative days are the centroids or the medoids of the clusters, and therefore, fail to capture extreme events such as the day with the peak load. To guarantee the feasibility of the investment decisions, several works have been done on "extreme days" selection. Most of these works select extreme days based on the values of the input data [8, 14, 13, 20]. [14, 13, 20] propose to select extreme days that have extreme values in the input data such as the days with the peak load, the peak net load, and/or the peak ramp-up and append these extreme days as additional representative days. [8, 10] propose to modify the clustering process to include the extreme values of the clusters. Besides identifying extreme days based on input data, [19] proposes to select extreme days based on the slack variables of the optimization problem itself.

V *How to estimate a bound for the error of the considered approach?*

A CEP model with representative days can be seen as a surrogate of the CEP model with all the days. It is relevant to know how far away the optimal solution of the surrogate is from the fullspace CEP model. [17] assume that the "ground truth model", i.e., the model with a sufficiently large number of days, is solvable and compare the investment decisions obtained from the representative day model with the solution of the "ground truth model". [15] proposes to use a sample average approximation approach [24] to calculate a statistical lower bound for the fullspace model. [18] proves that for some linear programming energy system problem, the surrogate problem is a relaxation of the fullspace problem under some assumptions. However, the properties proved in [18] are restricted to LP problems with uncertain data in the objective and right hand side coefficients, which does not apply to a general CEP model.

This paper aims to provide additional developments on the five issues discussed above. The works that are closest to this paper are [18, 17, 19]. The contributions of this paper are outlined below.

- The procedures of the input-based approach and the cost-based approach are presented with extensions to address general CEP problems.
- The theoretical properties of the cost-based approach and the input-based approach are analyzed including a method to estimate the "optimality gap" of the representative day approach with respect to the full day approach. Several relationships between the clustering error and the "optimality gap" are provided.
- Two extreme day selection methods are proposed.
- A case study is used to compare the effectiveness of different algorithms. The effects of adding extreme days are also analyzed.

2. Background: clustering algorithms

As discussed in the previous section, different clustering algorithms have been used for representative day selection. There is no clear winner among all the clustering algorithms reported in the literature [8, 18]. In this paper, k-means clustering and k medoids clustering are considered. In order to make this paper self-contained, the necessary background for these two clustering algorithms is provided in this section.

2.1. k-means clustering

Given a set of observations $(x_1, x_2, \dots, x_n)$, where each observation is a D-dimensional real vector, k-means clustering aims to partition the n observations into k $(\leq n)$ sets so as to minimize the within-cluster variances. Suppose $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$, denotes the partition of set $\{1, \dots, n\}$ into k subsets, k-means clustering is to solve the following optimization problem to find the optimal partition $\mathbf{S}^*$,

$$\mathbf{S}^* = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{1}$$

where μ_i is the mean of the points in S_i. The k-means clustering is an NP-hard problem. One approach to solve it is to use heuristics, such as the Lloyd–Forgy algorithm [25]. Implementation of this heuristic method is available in the well-known Python package scikit-learn [26].

Although the heuristics can provide good feasible solutions to Problem (1), they cannot guarantee that the global optimal solution is found. An alternative approach to solve (1) is to formulate the problem as a mixed-integer nonlinear program (MINLP) shown in Problem (2). Binary variable y_{il} represents whether the ith data point x_i belongs to the lth cluster for all $i \in \{1, \dots, n\}, l \in \{1, \dots, k\}$. Continuous variable c_{lj} denotes jth coordinate of the center of the lth cluster. Continuous variable d_i denotes the square of the Euclidean distance from the point x_i to the center of the cluster it belongs to. In Equation (2b), d_i is greater than or equal to the sum of the squared distances of each coordinate if point i belongs to cluster l, i.e., $y_{il} = 1$. The parameter M_i is a big-M parameter such that the inequality holds when $y_{il} = 0$. Equation (2c) forces each point i to be assigned to one of the clusters.

$$\min_{\mathbf{c},\mathbf{d},\mathbf{y}} \sum_{i=1}^{n} d_i \tag{2a}$$

$$d_i \geq \Big(\sum_{j=1}^{D} (x_{ij} - c_{lj})^2 \Big) - M_i(1 - y_{il}) \quad \forall i \in \{1, \dots, n\}, l \in \{1, \dots, k\} \tag{2b}$$

$$\sum_{l=1}^{k} y_{il} = 1 \quad \forall i \in \{1, \dots, n\} \tag{2c}$$

$$\mathbf{c}_l \in \mathbb{R}^D \quad \forall l \in \{1, \dots, k\} \tag{2d}$$

$$d_i \in \mathbb{R}_+ \quad \forall i \in \{1, \dots, n\} \tag{2e}$$

$$y_{il} \in \{0, 1\} \quad \forall i \in \{1, \dots, n\}, l \in \{1, \dots, k\} \tag{2f}$$

It is easy to observe that the size of (2) increases with the increase in the dimension of the data D, the number of data points n, and the number of clusters k. As a result, Problem (1) is difficult to solve for real-world CEP problems. In fact, we find that problem (2) is not solvable to global optimality for our case study. Therefore, the heuristic algorithm implemented in scikit-learn is used for k-means.

2.2. k-medoids clustering

The k-medoids problem is a clustering problem similar to k-means. Instead of minimizing the sum of within-cluster variances, k-medoids clustering seeks to minimize the distance from the "medoids" of each cluster. The difference between the medoid and the mean of the cluster is that the medoid has to be one of the actual data points within that cluster. In general, the k-medoids clustering problem is also NP-hard. There are heuristic algorithms for the k-medoids clustering [27] whose implementations are available in scikit-learn.

Alternatively, k-medoids can be solved to global optimality using mixed-integer linear programming (MILP). The MILP formulation is described in Problem (3). Binary variable y_i represents whether the ith data point is a medoid of a cluster (y_i=1) or not (y_i=0). Binary variable z_{ij} denotes whether the ith data point belongs to the cluster with the jth data point as its medoid. The objective is to minimize the sum of the distances from each point to its medoid. It is relevant to note that the distance between any two points i and j can be calculated *a priori*. Equation (3b) denotes that each point has to be assigned to exactly one point that is the medoid of the cluster it belongs to. Equation (3c) forces z_{ij} be zero if point j is not a medoid. Equation (3d) specify that the number of medoids is equal to k.

$$\min_{\mathbf{z},\mathbf{y}} \sum_{ij} d_{ij} z_{ij} \tag{3a}$$

$$\sum_{j=1}^{n} z_{ij} = 1 \quad \forall i = 1, 2, \dots, n \tag{3b}$$

$$z_{ij} \leq y_j \quad \forall i = 1, 2, \dots, n, j = 1, 2, \dots, n \tag{3c}$$

$$\sum_{i=1}^{n} y_i = k \tag{3d}$$

The number variables in (3) increases with the number of data points, but is independent of the number of clusters and the dimension of the data. We have found that (3) can be solved efficiently in CEP applications.

3. Representative day selection algorithm

To guarantee the fidelity of the CEP model, an accurate forecast of the hourly load and capacity factors of renewable generating units is needed. We assume that some historical load and capacity factor data are available and that those data are sufficient to characterize the possible capacity factor and load variations. The historical loads are scaled to consider the future load growth, i.e., some annual load growth rate is assumed. For the capacity factor data, no scaling is needed since we assume that the weather condition change is negligible across years. We further assume that the inter-day (not the intra-day) ramping constraints can be neglected, i.e., there is no linking constraints between two consecutive days. While this assumption may decrease the fidelity of the model, it is necessary for our representative day approach. One can always increase the length of the time block, e.g., using representative weeks [9] instead of representative days.

With these assumptions, a fullspace CEP problem with all the historical data used to represent each year of the planning horizon is defined in (4). Set $\mathcal{D}$ represents the set of days in the historical dataset. Each day has a weight of $\frac{365}{|\mathcal{D}|}$ in the objective function. Equation (4) is a succinct representation of the CEP problem. A detailed MILP formulation can be found in [28].

$$(FD) \quad OBJ_{FD} = \min \sum_{t \in \mathcal{T}} \Big(c_t^\top x_t + \sum_{d \in \mathcal{D}} \frac{365}{|\mathcal{D}|} f_t^\top y_{t,d} \Big) \tag{4a}$$

$$\text{s.t.} \quad A_{t,d} x_t + B_t y_{t,d} \le b_{t,d} \quad \forall t \in \mathcal{T}, d \in \mathcal{D} \tag{4b}$$

$$C_{t-1} x_{t-1} + D_t x_t \le g_t \quad t = 2, 3, \ldots, |\mathcal{T}| \tag{4c}$$

$$x_t \in X_t, \quad \forall t \in \mathcal{T}, \quad y_{t,d} \in Y_t, \quad \forall t \in \mathcal{T}, d \in \mathcal{D} \tag{4d}$$

Variable x_t represents the investment decisions at year t. Variable $y_{t,d}$ represents the operating decisions corresponding to day d in year t. The objective (4a) is to minimize the total cost. Equation (4b) describes the operational decisions of each year t and each day d, such as power flow equations, unit commitment, and economic dispatch. Equations (4c) are investment-related constraints, such as the installation and retirement of generating units. Equations (4d) specify the domain of the variables. We note that the days in our dataset differ in the load and the capacity factors of the renewable generators. The parameters corresponding to the load appear on the right hand side of equation (4b) represented by $b_{t,d}$. The parameters corresponding to the capacity factors are part of the constraint matrices $A_{t,d}$. All the other parameters in the model, including c_t, f_t, B_t, C_t, D_t, and g_t, are only indexed by year t because they only change on a yearly basis.

Since problem (FD) includes all the historical data, it is best to solve (FD) directly to obtain planning decisions that are feasible for all the days in dataset $\mathcal{D}$. However, solving (FD) directly is prohibitive in practice since the number of variables in (FD) can easily exceed one billion [28] if we consider one or more years of historical data. Therefore, the model (RD) below is solved as a surrogate of the fullspace model where a set of representative days $\mathcal{K}$ is selected or constructed to approximate problem (FD). The number of representative days is denoted by the cardinality of set $\mathcal{K}$, i.e., $|\mathcal{K}|$. The weight of the kth representative day is denoted by w_k. Variable $y_{t,k}$ represents the operating decisions corresponding to the kth representative day in year t. Set $\tilde{Y}_t$ represents the LP relaxation of set Y_t, i.e., set $\tilde{Y}_t$ is obtained from set Y_t if all the integrality constraints regarding the $y_{t,k}$ variables are relaxed. The reason for relaxing the integrality constraints for $y_{t,k}$ is to make the model (RD) amenable to decomposition algorithms, such as the Benders decomposition in [28] or the nested Benders decomposition in [4]. Since all the integer variables are general integer variables instead of binary variables, the relaxation provides a very tight bound [28].

$$(RD) \quad OBJ_{RD} = \min \sum_{t \in \mathcal{T}} \left(c_t^\top x_t + \sum_{k \in \mathcal{K}} w_k f_t^\top y_{t,k} \right) \tag{5a}$$

$$\text{s.t.} \quad A_{t,k}x_t + B_t y_{t,k} \leq b_{t,k} \quad \forall t \in \mathcal{T}, k \in \mathcal{K} \tag{5b}$$

$$C_{t-1}x_{t-1} + D_t x_t \leq g_t \quad t = 2, 3, \ldots, |\mathcal{T}| \tag{5c}$$

$$x_t \in X_t, \quad \forall t \in \mathcal{T}, \quad y_{t,k} \in \tilde{Y}_t, \quad \forall t \in \mathcal{T}, k \in \mathcal{K} \tag{5d}$$

The main challenge that we face is to find out how the representative days should be selected in order to approximate problem (FD) as well as possible. Before diving into the representative day selection approaches, we introduce a quantitative metric to evaluate the solution quality of (RD). Suppose that the optimal investment decision of (RD) is $\mathbf{x}^{RD}$, then the "actual cost" of $\mathbf{x}^{RD}$ can be obtained by fixing the investment decisions at $\mathbf{x}^{RD}$ and solving the problem corresponding to each day in the full dataset $\mathcal{D}$ individually, which is equivalent to fixing the $\mathbf{x}$ variables in (FD) and solving the rest of the fullspace problem. We denote this objective as $OBJ_{FD}(\mathbf{x}^{RD})$. It is easy to see that

$$OBJ_{FD}(\mathbf{x}^{RD}) \geq OBJ_{FD}(\mathbf{x}^{FD}) = OBJ_{FD} \tag{6}$$

where $\mathbf{x}^{FD}$ represents the optimal investment decisions obtained with (FD).

3.1. Input-based approach

Most algorithms on representative day selection perform clustering directly on the input data to problem (FD). Let us first define the notations for this approach. Suppose we have a dataset that contains the historical loads and the capacity factors of solar and wind for each node $n \in \mathcal{N}$ within the considered geographical region. The set of historical days in this dataset is represented by set $\mathcal{D}$. The data corresponding to day $d \in \mathcal{D}$ is represented by vector H_d. H_d is a concatenation of the hourly time series data for the load, capacity factors of all the nodes $n \in \mathcal{N}$ in day d. This approach captures the correlations of the input data because the data corresponding to each day are concatenated. Before

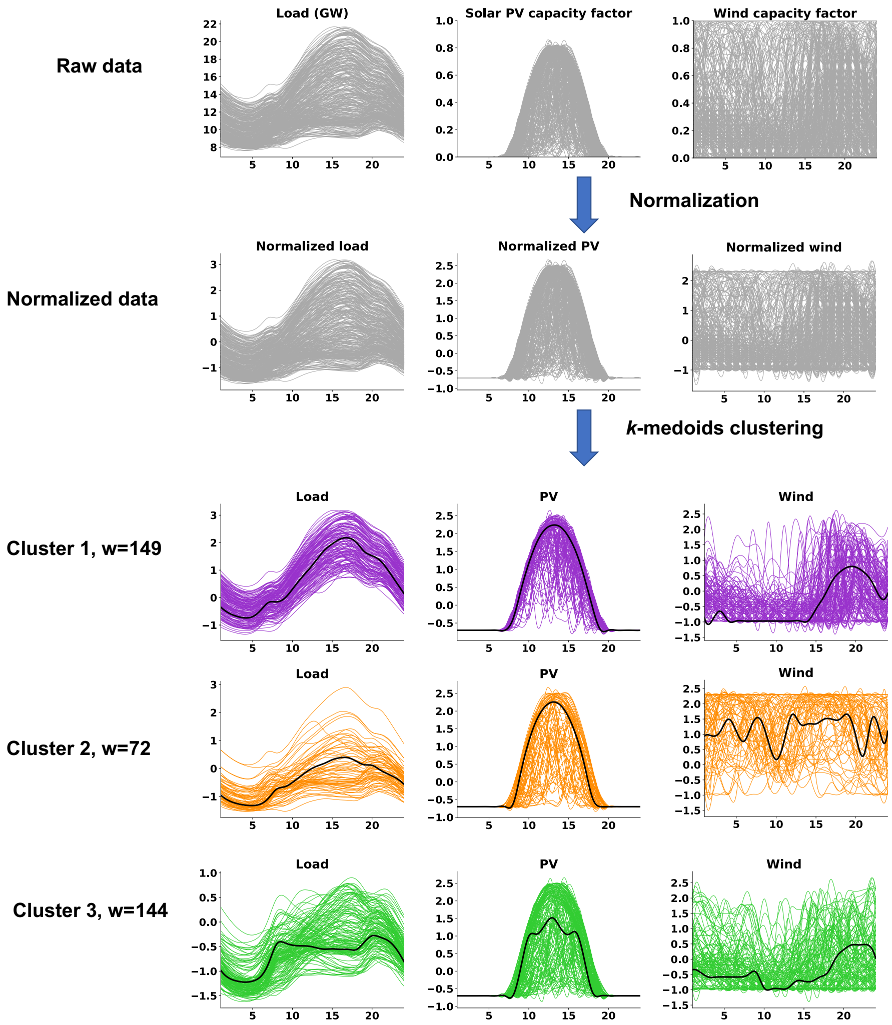

Figure 1: Illustration of the input-based approach

applying a clustering algorithm on $\{H_d, d \in \mathcal{D}\}$, the input data are usually normalized because they can be in very different magnitudes. For example, the capacity factors are from 0 to 1 while the load data are in GW. We normalize each type of data, e.g., the load data in a given node, using its mean and standard deviation over the whole dataset $\mathcal{D}$.

$$H_{d,p}^{\texttt{norm}} = \frac{H_{d,p} - \mu_p}{\sigma_p} \quad \forall p, d \tag{7}$$

where $H_{d,p}$ represents the entries in H_d corresponding to the data type p. For example, $H_{d,p}$ can correspond to the time series for the load in a given node in day d. μ_p and σ_p are scalars that represent the mean and standard deviation of $\{H_{d,p}, d \in \mathcal{D}\}$.

Other normalization schemes have been proposed in [18], where instead of normalizing by the mean and standard deviation of each data type p over the whole dataset, normalization can be performed based on the mean and variance of each day or each hour for a given data type p. In this paper, we only apply the whole dataset normalization approach.

After the normalization, the clustering algorithms can be applied to $\{H_d^{\texttt{norm}}, d \in \mathcal{D}\}$. Then, the whole dataset is partitioned into k clusters and the mean or the medoid of each cluster is selected as the input parameter corresponding to the representative day.

An illustrative example of the input-based approach is shown in Figure 1. The full dataset consists of 365 days of hourly load, hourly capacity factors of PV and wind corresponding to a single node, which is denoted as "raw data" in Figure 1. Each day is shown as a separate time series denoted by a gray line. After the normalization step, the time series become dimensionless and is shown as "normalized data". The k-medoids clustering is performed on the normalized data. For illustration, three clusters are shown in different colors with weights, 149, 72, and 144, respectively. The medoid of each cluster is shown as a bold black line. The medoids will be used as the input data corresponding to the representative days.

3.2. Cost-based approach

Besides clustering the days based on input data, another approach is based on solving a CEP problem with operating decisions for only a single day d in

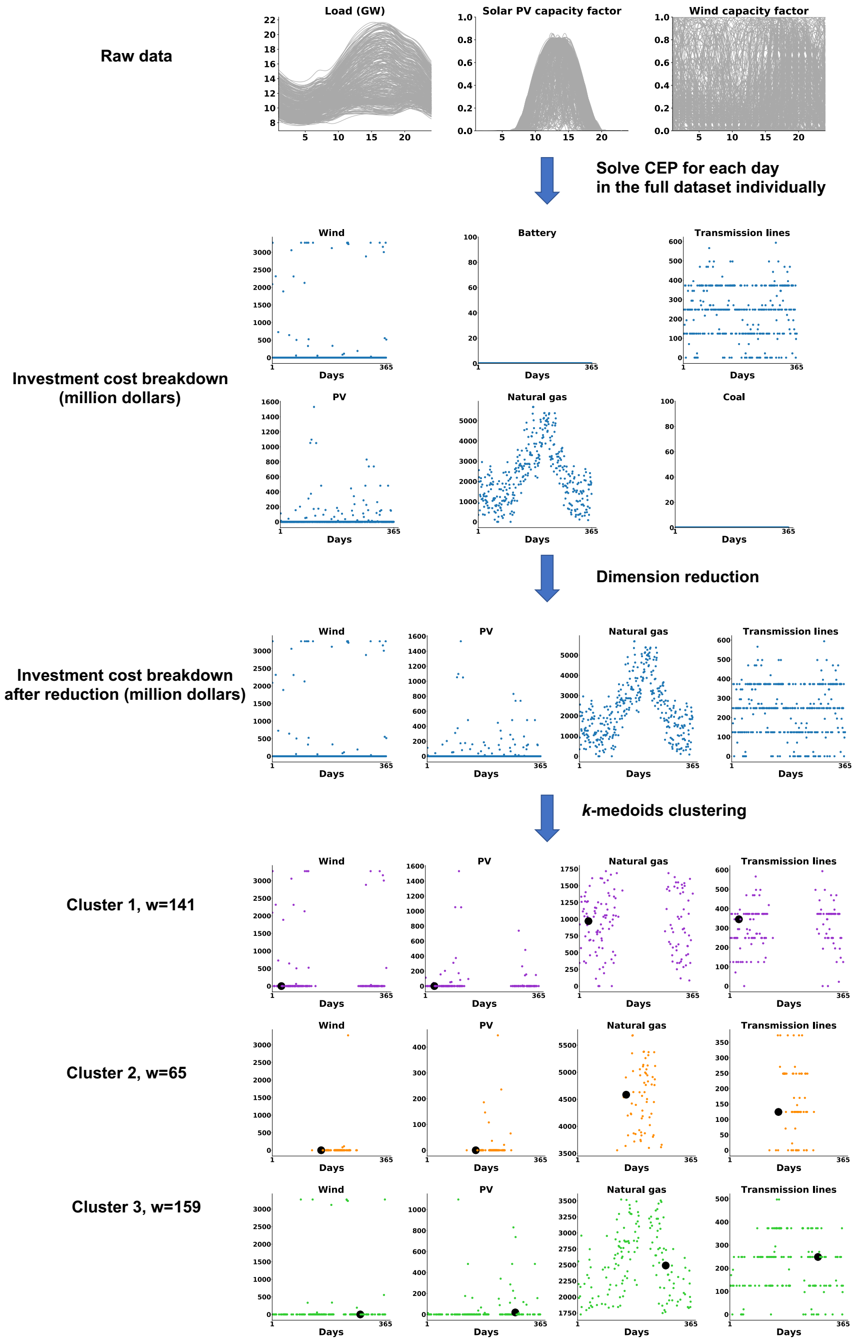

Figure 2: Illustration of the cost-based approach

$\mathcal{D}$ at a time. The CEP problem with one single day is small and can be solved relatively fast. After solving the single-day CEP problem for each day $d \in \mathcal{D}$, the optimal investment decisions are obtained for all $d \in \mathcal{D}$. The hypothesis is that the days with similar optimal investment decisions, i.e., the days that need similar generators, transmission lines, and storage units, are similar and should be assigned to the same cluster. One option is to perform clustering based on the optimal solutions themselves. However, normalization is needed because the number of generators, transmission lines, and storage units are expressed in different magnitudes. One natural way to perform the normalization is to transform the number of units to the cost of different types of units as proposed in [17] following the work reported in [22].

Our main contributions compared with [17] in the cost-based approach are outlined below, which are discussed after an illustrative example is presented.

(i) We provide an approach that accommodate planning problems with a time horizon longer than one year.

(ii) Both k-means and k-medoids clustering are adapted to the cost-based approach, while [17] only considers a k-medoids approach.

(iii) A simple dimension reduction method is proposed, and the effects of such method are analyzed combined with the two clustering algorithms.

To provide a conceptual overview of the proposed approach, an illustrative example is shown in Figure 2. The available raw data are hourly load, and hourly capacity factors of PV and wind for 365 days. A CEP problem that has a single day is solved individually for each of the 365 days. The optimal investment costs in the wind, PV, natural gas, and coal generating units, batteries, and transmission lines are shown in the "investment cost breakdown" charts for each of the 365 days. Note that in practice, investment costs can be associated with each node of the CEP problem, i.e., each node has its associated costs in installing generating units and batteries. The transmission line costs are associated with any two nodes. For illustration purpose, we only show the investment costs of generating units and batteries corresponding to one selected node, and the transmission line costs connecting two selected nodes. Dimensional reduction is performed on the cost data by removing the units that are never installed in any of the 365 days. Specifically, the battery and the coal generators are never installed and therefore removed because they do not contribute to the clustering error. k-medoids clustering is performed on the investment costs breakdown after dimension reduction. Three clusters are obtained shown in different colors. The medoid of each cluster is highlighted using large black dots. The days corresponding to the medoids are selected as the representative days.

Next, we present our contributions regarding (i)-(iii). For (i), the proposed approach can be applied to solve a CEP model with T years by considering the total discounted cost of each type of investment over the T years. For (ii), it is not straightforward to apply k-means clustering because the mean of the investment costs does not correspond to any days. However, it is possible to transform the data back to the input domain and take the mean of the input data for each cluster. For (iii), the dimension reduction technique that removes the all-zero entries can reduce the sizes of the MINLP problem for k-means clustering. On the other hand, the MILP formulation of the k-medoids clustering is independent of the dimension of the problem. Heuristic algorithms for clustering are generally computationally effective and do not suffer from the "curse of dimensionality". Since the MINLP formulation is too expensive to solve and will not be used in CEP problems, we conclude that the dimension reduction techniques have marginal effects on the clustering algorithms that are used.

The detailed steps of the cost-based algorithm are described in Algorithm 1.

3.3. Extreme days selection

As discussed in the introduction section, including only $|\mathcal{K}|$ representative days to solve the CEP model can lead to expansion decisions that are infeasible under some extreme days. To examine whether the optimal investment decisions

Algorithm 1

```
Initialization: A dataset $\mathcal{D}$ includes the load, and wind and solar capacity factor data. The data corresponding to the $d$th day in $\mathcal{D}$ is denoted as $H_d$.
for $d = 1$ to $|\mathcal{D}|$ do
    Solve a $T$ year CEP problem using $H_d$ as the single representative day data. Denote the optimal investment costs for each type of the generating and storage units in each node, and the transmission lines that connect any two nodes as a concatenated vector $c_d$. Denote the total cost of the CEP model as $tc_d$
end for
Considering matrix $C = [c_1, c_2, \dots, c_{|\mathcal{D}|}]$, delete the all-zero rows in $C$ to derive matrix $\tilde{C}$.
Perform a $k$-clustering algorithm on the columns of matrix $\tilde{C}$. Select the representative days.
```

found by the reduced problem (RD) is feasible for the fullspace problem (FD), the operating problem corresponding to each day $d \in \mathcal{D}$ is solved with the optimal investment decisions fixed evaluate the objective $OBJ_{FD}(\mathbf{x}^{RD})$. If there are no infeasible days in $\mathcal{D}$, i.e., $OBJ_{FD}(\mathbf{x}^{RD}) < +\infty$, the investment decisions found by solving (FD) are sufficient to satisfy the "extreme" scenarios. If not, suppose the set of infeasible days are represented by $\mathcal{I}$. In this case, we select the "extreme day" or "most infeasible day" in set $\mathcal{I}$ and add this extreme day to set $\mathcal{K}$ as a new cluster with only 1 day. Then, the weights of the representative days are adjusted accordingly. With the added extreme day, the reduced problem (RD) is solved again to find the new investment decisions. We repeat this procedure until all the days in the set $\mathcal{D}$ are feasible for our optimal investment decisions.

We next focus on how to choose the "extreme day" from the set of infeasible days $\mathcal{I}$. We propose two approaches to select the extreme days.

3.3.1. Load shedding cost

The operating problems for the infeasible days are solved again with load shedding variables added to energy balance constraints for each node at each hour. The load shedding variables are used to balance the load when the power generation is not enough to satisfy the load. The objective function is changed to minimizing the total load shedding cost over the planning horizon. The day with the highest total load shedding cost is chosen as the extreme day.

This approach is similar to the one proposed in [19]. The difference is that [19] proposes to choose the infeasible day with the highest load shedding cost for a single hour.

3.3.2. Highest cost

Using the cost-based algorithm, the optimal total cost tc_d (investment cost + operating cost) of each day in the dataset $\mathcal{D}$ is obtained by solving the optimization problem described in Algorithm 1. We choose the infeasible day with the highest total cost obtained in Algorithm 1 as the extreme day.

$$d^{\texttt{select}} = \underset{d, d \in \mathcal{I}}{\arg\max}\, tc_d \tag{8}$$

The rationale for this selection is that the day with the highest cost is likely to include the extreme events that trigger installing additional units and result in a high operating cost. The advantage of the highest cost approach is that it exploits information calculated already in Algorithm 1 and therefore there is no need to solve the infeasible CEP model again to find the extreme day.

3.4. Properties of the proposed algorithms

In this subsection, we analyze the proposed algorithms by characterizing their properties.

3.4.1. Lower bound

An upper bound of the optimal objective value of the fullspace problem (FD) is available by evaluating $OBJ_{FD}(\mathbf{x}^{RD})$ as shown in Equation (6). Additionally, it is desirable to provide a lower bound of (FD) by solving (RD) to obtain an estimate of the suboptimality of the investment decisions obtained from (RD). The following theorem provides such a lower bound.

Theorem 1. *For both cost-based and input-based approaches, if k-means clustering in* (1) *is used, (RD) provides a lower bound for the optimal objective* *value of (FD), i.e., $OBJ_{RD} \leq OBJ_{FD}$. This lower bound holds before and after adding extreme days.*

Proof. The proof is provided in Appendix A.1. □

Since both the upper bound and the lower bound of problem (FD) are available, an optimality gap of the solution $\mathbf{x}^{RD}$ can be estimated using

$$\texttt{Gap} = \frac{OBJ_{FD}(\mathbf{x}^{RD}) - OBJ_{RD}}{OBJ_{FD}(\mathbf{x}^{RD})} \times 100\% \tag{9}$$

3.4.2. Relationship between the clustering error and the optimality gap

The clustering error refers to the sum of deviations from the mean or the medoids in the k-means or k-medoids clustering, respectively. The mathematical expressions for the clustering errors are given in Equations (1) and (3a) for k-means and k-medoids, respectively. As the number of clusters increases, the clustering error is guaranteed to decrease. Intuitively, one would expect the optimality gap defined in Equation (9) to decrease as the clustering error diminishes, or at least, the upper bound, $OBJ_{FD}(\mathbf{x}^{RD})$, to improve as k increases. Our computational results in section 4 show that this actually happens in most cases. However, no formal characterization is available in the literature. Here, we provide several mathematical properties.

A relatively simple question is whether the optimality gap is zero if the clustering error is zero. For the input-based approach, the following theorem holds.

Theorem 2. *For the input-based approach, if the clustering error is zero and the integrality constraints on variables $y_{t,d}$ are not relaxed in (RD), then the optimality gap defined in* (9) *is zero, i,e., $\boldsymbol{x}^{RD}$ is optimal for (FD).*

Proof. The proof is provided in Appendix A.2. □

For the cost-based approach, we can prove the following theorem.

Theorem 3. *For the cost-based approach, if the k-medoids is used and the clustering error is zero, $|\mathcal{K}| = 1$, $|\mathcal{T}| = 1$, and the integrality constraints on* *variables $y_{t,k}$ are not relaxed in (RD), then the optimality gap defined in* (9) *is zero, i,e., $\boldsymbol{x}^{RD}$ is optimal for (FD).*

Proof. The proof is provided in Appendix A.3. □

It is easy to observe that additional assumptions are needed in Theorem 3 than in Theorem 2, such as the conditions that the number of clusters and the number of years are both one. This is due to the fact that transforming the data to the cost domain entails loss of information. In other words, any two days d and d' with the same input data have the same optimal costs but not vice versa. When the assumptions in Theorem 3 are not satisfied, the cost-based approach is not guaranteed to find the optimal investment decisions and is, therefore, a rather empirical approach compared with the input-based approach. However, computational results in section 4 indicate the cost-based approach is empirically favorable under certain criteria.

An additional relevant question is whether the lower bound and the upper bound improve as k increases. The following Theorem provides a sufficient condition to improve the lower bound. A similar Theorem can be found in [18] for LP problems.

Theorem 4. *Suppose that (RD) is solved with cluster number $|\mathcal{K}_1|$ and $|\mathcal{K}_2|$, $|\mathcal{K}_1| > |\mathcal{K}_2|$ that come from a k-means clustering algorithm. Denote the objective value of (RD) as $OBJ_{RD}^{\mathcal{K}_1}$, $OBJ_{RD}^{\mathcal{K}_2}$ for $|\mathcal{K}_1|$ and $|\mathcal{K}_2|$, respectively. If each cluster in the $|\mathcal{K}_1|$ clusters is contained in one of the clusters of the $|\mathcal{K}_2|$ clusters, then $OBJ_{RD}^{\mathcal{K}_1} \geq OBJ_{RD}^{\mathcal{K}_2}$.*

Proof. The proof is provided in Appendix A.4. □

This theorem provides a condition for lower-bound improvement in k-means clustering. Following Theorem 4, it is easy to see that the lower bound provided by (RD) improves after the extreme days are added when k-means clustering are used because the added extreme days are contained in some of the original clusters.

4. Computational results

The case study of the Electric Reliability Council of Texas (ERCOT) reported in [28] is used to test the representative day selection methods analyzed in this paper. The problem is a generation transmission expansion planning problem with a planning horizon of 5 years. The ERCOT region is modeled using 5 nodes, South, Coast, Northeast, West, and Panhandle. The investment decisions include the number of coal, natural gas, nuclear, solar, and wind generating units, the number of storage units that are installed in each node, and the number of transmission lines that are installed connecting any two nodes each year. The operating decisions involve unit commitment, and must comply with the DC power flow equations. The historical dataset $\mathcal{D}$ consists of 365 days.

The data used for clustering can be the input data or cost data. Both k-means and k-medoids clustering algorithms are used. For the input-based approach, the extreme days are selected using the load shedding cost approach. For the cost-based approach, both the load shedding cost and the highest cost approach are used.

All the models and algorithms are implemented using Pyomo/Python [29]. k-means clustering is solved using the heuristic provided by scikit-learn [26]. k-medoids clustering is solved with the MILP formulation shown in (3) using the solver CPLEX [30].

4.1. Comparison of different algorithms

To test the proposed representative day selection methods, the following six algorithm options (shown in Table 1) are tested.

The computational results of the six algorithm options are shown in Table 2 where k represents the initial number of representative days for the problem (RD). "#infeasible day" represents the number of infeasible days resulting from

[*] The values of OBJ_{RD} for algorithm option 2,3,4 are not valid lower bounds and are shown in italics. Therefore, the gaps for these options are omitted.

Table 1: The input data, clustering algorithm, and extreme day method used by the six cases.

Algorithm option	Data	Clustering algorithm	Extreme day method
1	Input	k-means	load shedding cost
2	Input	k-medoids	load shedding cost
3	Cost	k-medoids	highest cost
4	Cost	k-medoids	load shedding cost
5	Cost	k-means	highest cost
6	Cost	k-means	load shedding cost

Table 2: Computational results of the six algorithm options

Option	k	#infeasible day	X	$OBJ_{FD}(\mathbf{x}^{RD})$	$OBJ_{RD}{}^*$	Gap*	Extreme days selected
1	5	70	3	79.16	76.09	4.0%	212,213,214
	10	63	2	79.04	76.29	3.6%	212,213
	15	42	2	78.81	76.58	2.9%	212,213
2	5	35	3	78.92	*76.98*	-	212,213,177
	10	21	2	78.72	*73.10*	-	212,213
	15	40	2	78.74	*76.18*	-	212,213
3	5	98	5	78.83	*72.74*	-	221,249,212,213,177
	10	13	3	78.67	*77.39*	-	221,212,213
	15	12	3	78.81	*76.05*	-	221,212,213
4	5	98	3	78.93	*72.51*	-	212,213,177
	10	13	2	78.79	*77.32*	-	212,213
	15	12	1	78.75	*75.94*	-	212
5	5	34	4	78.98	76.16	4.2%	221,249,212,213
	10	30	6	79.09	76.64	3.7%	221,249,212,213,177,214
	15	29	4	78.98	76.74	3.4%	221,249,212,213
6	5	34	3	79.12	76.15	3.9%	212,213,177
	10	30	4	78.93	76.63	3.0%	212,214,213,177
	15	29	3	78.81	76.73	2.7%	212,177,213

the investment decisions of the k representative days problem. X is the number of extreme days added in addition to the k representative days to make $\mathbf{x}^{RD}$ feasible for all $d \in \mathcal{D}$, i.e., to attain $OBJ_{FD}(\mathbf{x}^{RD}) < +\infty$. The objective value of the (RD), OBJ_{RD}, after the X extreme days are added, are reported. Note that OBJ_{RD} is only a valid lower bound when the k-means clustering is used. The values of OBJ_{RD} for algorithm options 2,3,4 that use the k-medoids clustering are shown in italics in Table 2 since they do not provide valid lower bounds. $OBJ_{FD}(\mathbf{x}^{RD})$ is the "actual cost" when evaluating the investment decision of (RD) on the full dataset and therefore provides an upper bound. The "Gap" is calculated using Equation (9) for algorithm option 1,5,6, where k-means clustering is used. The days in the historical dataset $\mathcal{D}$ are numbered from 1 to 365. The "extreme days selected" column shows the indices of the extreme days.

In most of the algorithm options, $OBJ_{FD}(\mathbf{x}^{RD})$ (upper bound) and OBJ_{RD} (provided that it's a valid lower bound) improve as k increases. If k-means clustering is used, the optimality gap improves as k increases. However, the upper bound is usually better when k-medoids clustering is used. We conclude that there is no clear winner between the k-medoids and the k-means clustering. When k-means clustering is used, the lowest optimality gap is achieved in algorithm option 6 with $k = 15$. There is a trend of decrease in the optimality gap as k increases in option 1,5, and 6 if the k-means is used. If one wishes to achieve an even lower optimality gap, k has to be increased. For our analysis, we consider that the optimality gap of 2.7% is sufficient to show the capability of the algorithms. It should be noted that the suboptimality comes from both the use of representative days instead of the full dataset and the relaxation of the integer variables in the operating subproblems. The relaxation of the integer variables result in an gap of around 1% in our computational experiments [28].

The number of infeasible days if using only the initial representative days is a good indicator of the effectiveness of the algorithm. One would expect the algorithms with better performance to have fewer infeasible days. Overall, as expected, the number of infeasible days decreases as k increases. The cost-based approach with k-medoids clustering and 15 representative days has the smallest number of initial infeasible days.

We note that X, the number of extreme days added to make $\mathbf{x}^{RD}$ feasible for the whole dataset $\mathcal{D}$, is an indicator of whether the extreme day selection approach is effective. The load shedding cost approach needs a smaller number of extreme days to achieve feasibility than the highest cost approach. Among all the algorithm options, there is an overlap in the extreme days selected. Days number 212,214,221, 177, are most frequently selected. The two extreme days selection approaches sometimes select the same days as extreme days. It should be noted that day 177 happens to be the day with the peak load. The highest ramp occurs in day 12, which is not selected by our proposed extreme day selection algorithms. The results show that our methods select sometimes different extreme days than just focusing on the input data, such as peak load and peak ramp. Selecting extreme days solely based on peak load and peak ramp does not guarantee feasibility.

Next, we analyze the computational time of different algorithm options. All the problems are solved using CPLEX version 12.9.0.0 [30] using one processor of an Intel Xeon (2.67GHz) machine with 64 GB RAM. For the cost-based approach, a CEP problem is solved for each day in the dataset $\mathcal{D}$ to obtain the optimal investment costs. To save computational time, the LP relaxations of these CEP models are solved. As a matter of fact, we also tested solving these CEP models with the integrality constraints and find that the optimal investment costs vary very little with and without the integrality constraints. The total computational time to solve the LP relaxations of these 365 CEP models is 2,091 seconds. Additionally, the solution time of the Benders decomposition algorithm for different algorithm options with and without extreme days are shown in Table 3. All these CEP models can be solved within 10 hours. Although the sizes of the CEP models does not depend on the algorithm option if they have the same number of representative days, their solution time can vary significantly.

Table 3: Solution time of the Benders decomposition algorithm for different algorithm options with and without adding X extreme days (secs)

Option	1	2	3	4	5	6
k=5	938	1,460	2,078	2,392	1,004	1,436
k=5+X	3,079	6,769	16,961	15,240	3,567	2,705
k=10	1,520	9,175	4,404	4,557	2,897	3,367
k=10+X	3,190	12,722	6,223	17,896	6,905	6,994
k=15	4,647	17,914	26,514	27,019	6,235	7,433
k=15+X	5,095	9,126	25,320	31,899	10,568	13,395

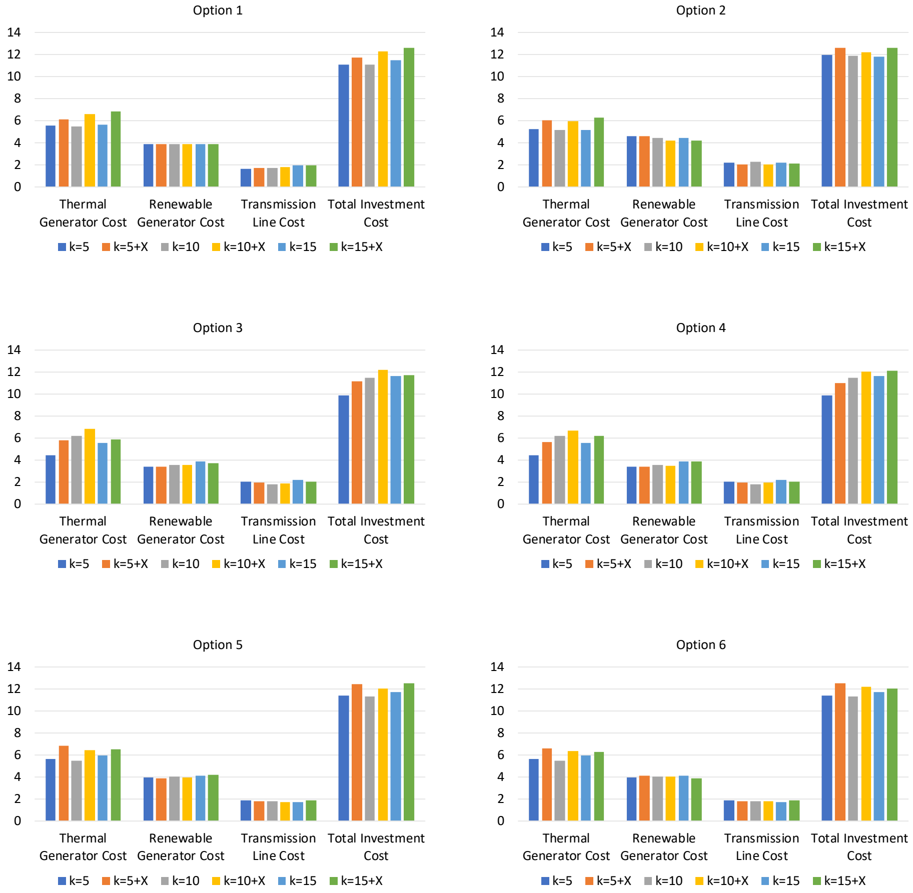

Figure 3: Thermal generating unit cost, renewable generating unit cost, transmission line cost and total investment cost change with and without the extreme days

4.2. Comparison of the capacity expansion results with and without adding the extreme days

Once the extreme days are added, the investment decisions $\mathbf{x}^{RD}$ change compared with the investment decisions that are optimal for the initial k representative days. We compare in Figure 3 the thermal generator cost, the renewable generator cost, the transmission line cost, and the total investment cost in the 5 years of the planning horizon with and without adding the extreme days. In this Figure, $k = 5$ denotes the costs associated with the 5 representative day problem. $k = 5 + X$ denotes the problems with the initial 5 days plus the X extreme days. In all the algorithm options, the total investment cost increases after adding the X extreme days. The majority of this increase comes from the increase in the thermal generator costs while there is no consistent trend in the change in the costs of the renewable generators and the transmission lines. This can be explained by the fact that more dispatchable generating units are needed after the extreme days are included.

To provide a more quantitative view of the change in the costs, we select option 6, where we use cost-based k-means clustering, start with 15 representative days, and add extreme days according to the load shedding cost. After adding 3 extreme days to make all the days feasible, the total investment cost increases from 11.74 trillion dollars to 12.06 trillion dollars (2.7% increase). The number of transmission lines increases from 15 to 16. There is a small increase as well in the storage investment cost (0.2 million dollars). Additionally, there is an increase in total thermal generator cost (346 million dollars) and a decrease in renewable generator cost (-212 million dollars). In this case, the portfolio of the dispatchable and nondispatchable generating units is adjusted after the extreme days are included in order to make the planning decisions suitable for the extreme events.

5. Conclusion

In the context of power system expansion planning, we have presented an input-based approach and an cost-based approach for the selection of representative days. Two novel extreme day selection algorithms, known as load shedding cost and highest cost, are proposed. The properties of the proposed algorithms are theoretically analyzed. In particular, an upper bound and an lower bound of the optimal objective value of the fullspace problem (FD) are obtained. We also identify the conditions under which zero clustering error implies zero optimality gap for both input-based and cost-based approaches. It turns out that the conditions for the cost-based approach are more restrictive than the input-based approach because there is an information loss in projecting the data from the input space to the cost space.

A case study of the ERCOT region is used to compare the different proposed algorithms. There is no clear winner between the k-means clustering and the k-medoids clustering algorithms in terms of the upper bound. However, the k-means clustering is recommended since it provides a valid lower bound. The cost-based approach and the input-based approach have varying performances in combination with different clustering and extreme day selection methods. The load shedding cost approach outperforms the highest cost approach in extreme day selection in the sense that it needs fewer extreme days to achieve feasibility. We also compare the change in the investment decisions with and without adding the extreme days. The major change is that additional dispatchable generating units are added to make the planning decisions feasible under extreme events.

Appendix A Proofs of the Theorems

A.1 Proof of Theorem 1

Proof. It suffices to prove that for any feasible solution to (FD) it is always possible to construct some feasible solution to (RD) that yields the same objective value. Suppose (x_t, $y_{t,d}$, $\forall t \in \mathcal{T}, d \in \mathcal{D}$) is a feasible solution of (FD). Suppose that by applying the k-means clustering algorithm, set $\mathcal{D}$ is partitioned into $|\mathcal{K}|$ clusters, i.e., $\mathcal{D} = \cup_{k \in \mathcal{K}} \mathcal{D}_k$ where set $\mathcal{D}_k$ represents the days in the kth cluster. Constraint (4b) can be rewritten as

$$A_{t,d} x_t + B_t y_{t,d} \leq b_{t,d} \quad \forall d \in \mathcal{D}_k \tag{A.1}$$

for all $k \in \mathcal{K}, t \in \mathcal{T}$. We aggregate all the constraints in (A.1) for a given cluster k in year t and obtain

$$(\sum_{d \in \mathcal{D}_k} A_{t,d}) x_t + B_t (\sum_{d \in \mathcal{D}_k} y_{t,d}) \leq \sum_{d \in \mathcal{D}_k} b_{t,d} \tag{A.2}$$

Divide (A.3) by $|\mathcal{D}_k|$ on both sides we have

$$A_{t,k} x_t + B_t (\frac{\sum_{d \in \mathcal{D}_k} y_{t,d}}{|\mathcal{D}_k|}) \leq b_{t,k} \tag{A.3}$$

because by definition of the k-means clustering algorithm, $A_{t,k}$, $b_{t,k}$ are the mean values of $\{A_{t,d}, \forall d \in \mathcal{D}_k\}$ and $\{b_{t,d}, \forall d \in \mathcal{D}_k\}$, respectively.

It is easy to see that if we set $y_{t,k} = \frac{\sum_{d \in \mathcal{D}_k} y_{t,d}}{|\mathcal{D}_k|}$ for all $k \in \mathcal{K}, t \in \mathcal{T}$. ($x_t, y_{t,k}, k \in \mathcal{K}, t \in \mathcal{T}$) is a feasible solution of (RD) that yields the same objective value as (x_t, $y_{t,d}$, $\forall t \in \mathcal{T}, d \in \mathcal{D}$) for problem (FD). This completes the proof. □

A.2 Proof of Theorem 2

Proof. Since the clustering error is zero, the input parameters in each cluster must be the same. Therefore, in the fullspace problem (RD), the optimal operating decisions for the days in each cluster $k \in \mathcal{K}$ are the same, i.e., $y_{t,d} = y_{t,d'}$

for any d, d' within the same cluster. The decisions within each cluster can be aggregated without any sacrifice of optimality. The aggregated problem is exactly the problem (RD) without relaxing the integrality constraints on $y_{t,k}$. $\square$

A.3 Proof of Theorem 3

Proof. Since there is only one cluster and one year in the planning horizon, the clustering error of the cost-based approach being zero implies that the optimal investment decisions of all the days are the same if the CEP problem is solved for each day in $\mathcal{D}$ individually. Denote this investment decision as $\mathbf{x}^{common}$. Clearly, $\mathbf{x}^{common}$ is optimal for (FD). Furthermore, (RD) is solved using the medoid of the single cluster that corresponds to one of the days in $\mathcal{D}$. By definition, $\mathbf{x}^{common}$ is also optimal for (RD). $\square$

A.4 Proof of Theorem 4

Proof. Using the same proof reasoning as in Theorem 1, the variables and constraints corresponding to $\mathcal{K}_1$ can be aggregated to produce the variables and constraints corresponding to $\mathcal{K}_2$ and the inequality follows. $\square$

Acknowledgements

This research was conducted as part of the Institute for the Design of Advanced Energy Systems (IDAES) with funding from the Simulation-Based Engineering, Crosscutting Research Program of the U.S. Department of Energy's Office of Fossil Energy. Can Li would like to thank Dr. Dimitri Papageorgiou from Exxon Mobil for some fruitful discussions.

References

[1] A. J. Conejo, L. Baringo, S. J. Kazempour, A. S. Siddiqui, Investment in electricity generation and transmission, Cham Zug, Switzerland: Springer International Publishing 119.

[2] N. E. Koltsaklis, A. S. Dagoumas, State-of-the-art generation expansion planning: A review, Applied Energy 230 (2018) 563–589.

[3] R. Hemmati, R.-A. Hooshmand, A. Khodabakhshian, State-of-the-art of transmission expansion planning: Comprehensive review, Renewable and Sustainable Energy Reviews 23 (2013) 312–319.

[4] C. L. Lara, D. S. Mallapragada, D. J. Papageorgiou, A. Venkatesh, I. E. Grossmann, Deterministic electric power infrastructure planning: Mixed-integer programming model and nested decomposition algorithm, European Journal of Operational Research 271 (3) (2018) 1037–1054.

[5] B. Palmintier, M. Webster, Impact of unit commitment constraints on generation expansion planning with renewables, in: 2011 IEEE power and energy society general meeting, IEEE, 2011, pp. 1–7.

[6] D. S. Mallapragada, D. J. Papageorgiou, A. Venkatesh, C. L. Lara, I. E. Grossmann, Impact of model resolution on scenario outcomes for electricity sector system expansion, Energy 163 (2018) 1231–1244.

[7] A. Almaimouni, A. Ademola-Idowu, J. N. Kutz, A. Negash, D. Kirschen, Selecting and evaluating representative days for generation expansion planning, in: 2018 Power Systems Computation Conference (PSCC), IEEE, 2018, pp. 1–7.

[8] L. Kotzur, P. Markewitz, M. Robinius, D. Stolten, Impact of different time series aggregation methods on optimal energy system design, Renewable Energy 117 (2018) 474–487.

[9] B. Bahl, T. Söhler, M. Hennen, A. Bardow, Typical periods for two-stage synthesis by time-series aggregation with bounded error in objective function, Frontiers in Energy Research 5 (2018) 35.

[10] Á. García-Cerezo, L. Baringo, R. García-Bertrand, Representative days for expansion decisions in power systems, Energies 13 (2) (2020) 335.

[11] Y. Liu, R. Sioshansi, A. J. Conejo, Hierarchical clustering to find representative operating periods for capacity-expansion modeling, IEEE Transactions on Power Systems 33 (3) (2017) 3029–3039.

[12] P. Nahmmacher, E. Schmid, L. Hirth, B. Knopf, Carpe diem: A novel approach to select representative days for long-term power system modeling, Energy 112 (2016) 430–442.

[13] S. Pfenninger, Dealing with multiple decades of hourly wind and pv time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability, Applied Energy 197 (2017) 1–13.

[14] I. J. Scott, P. M. Carvalho, A. Botterud, C. A. Silva, Clustering representative days for power systems generation expansion planning: Capturing the effects of variable renewables and energy storage, Applied Energy 253 (2019) 113603.

[15] P. Seljom, A. Tomasgard, Sample average approximation and stability tests applied to energy system design, Energy Systems (2019) 1–25.

[16] F. J. De Sisternes, M. D. Webster, Optimal selection of sample weeks for approximating the net load in generation planning problems, esd. mit. edu.

[17] M. Sun, F. Teng, X. Zhang, G. Strbac, D. Pudjianto, Data-driven representative day selection for investment decisions: A cost-oriented approach, IEEE Transactions on Power Systems 34 (4) (2019) 2925–2936.

[18] H. Teichgraeber, A. R. Brandt, Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison, Applied Energy 239 (2019) 1283–1293.

[19] H. Teichgraeber, C. P. Lindenmeyer, N. Baumgärtner, L. Kotzur, D. Stolten, M. Robinius, A. Bardow, A. R. Brandt, Extreme events in time series aggregation: A case study for optimal residential energy supply systems, arXiv preprint arXiv:2002.03059.

[20] A. Yeganefar, M. R. Amin-Naseri, M. K. Sheikh-El-Eslami, Improvement of representative days selection in power system planning by incorporating the extreme days of the net load to take account of the variability and intermittency of renewable resources, Applied Energy 272 (2020) 115224.

[21] W. W. Tso, C. D. Demirhan, C. F. Heuberger, J. B. Powell, E. N. Pistikopoulos, A hierarchical clustering decomposition algorithm for optimizing renewable power systems with storage, Applied Energy 270 (2020) 115190.

[22] S. Pineda, A. Conejo, Scenario reduction for risk-averse electricity trading, IET Generation, Transmission & Distribution 4 (6) (2010) 694–705.

[23] K. Poncelet, H. Höschle, E. Delarue, A. Virag, W. D'haeseleer, Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems, IEEE Transactions on Power Systems 32 (3) (2016) 1936–1948.

[24] A. Shapiro, D. Dentcheva, A. Ruszczyński, Lectures on stochastic programming: modeling and theory, MOS-SIAM Series on Optimization, 2014.

[25] E. W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics 21 (1965) 768–769.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[27] H.-S. Park, C.-H. Jun, A simple and fast algorithm for k-medoids clustering, Expert systems with applications 36 (2) (2009) 3336–3341.

[28] C. Li, A. J. Conejo, P. Liu, B. Omell, J. D. Siirola, I. E. Grossmann, Mixed-integer linear programming models and algorithms for generation

and transmission expansion planning of power systems, Optimization Online.

[29] W. E. Hart, J.-P. Watson, D. L. Woodruff, Pyomo: modeling and solving mathematical programs in python, Mathematical Programming Computation 3 (3) (2011) 219.

[30] IBM, IBM ILOG CPLEX Optimization Studio CPLEX User's Manual, Tech. rep., IBM Corp. (2015).