



An objective-based scenario selection method for transmission network expansion planning with multivariate stochasticity in load and renewable energy sources



M. Sun, F. Teng^{*}, I. Konstantelos, G. Strbac

Electrical and Electronic Engineering, Imperial College London, London, UK

ARTICLE INFO

Article history:

Received 18 January 2017

Received in revised form

15 December 2017

Accepted 30 December 2017

Available online 5 January 2018

Keywords:

Clustering

Transmission network expansion planning

Resource variability

Wind power

ABSTRACT

Transmission Network Expansion Planning (TNEP) in modern electricity systems is carried out on a cost-benefit analysis basis; the planner identifies investments that maximize the social welfare. As the integration of Renewable Energy Sources (RES) increases, there is a real challenge to accurately capture the vast variability that characterizes system operation within a planning problem. Conventional approaches that rely on a large number of scenarios for representing the variability of operating points can quickly lead to computational issues. An alternative approach that is becoming increasingly necessary is to select representative scenarios from the original population via clustering techniques. However, direct clustering of operating points in the input domain may not capture characteristics which are important for investment decision-making. This paper presents a novel objective-based scenario selection framework for TNEP to obtain optimal investment decisions with a significantly reduced number of operating states. Different clustering frameworks, clustering variables and clustering techniques are compared to determine the most appropriate approach. The superior performance of the proposed framework is demonstrated through a case study on a modified IEEE 118-bus system.

© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

1.1. Motivation and literature review

The worldwide push for decarbonisation of electricity systems is expected to be largely achieved by the high penetration of renewable energy sources (RES). Given that the majority of RES is remotely located in most jurisdictions, large scale transmission network expansion is required to accommodate these RES. Historically, network design had been driven by the need to meet peak demand with sufficient reliability [1]; this peak-based approach has led to economically efficient solutions in systems dominated by thermal generators with high capacity value. However, under high penetration of intermittent RES, that have a much lower capacity value, accommodating peak flows during high demand seizures to be the primary investment driver. Consequently, transmission

investment is undertaken on a cost-benefit basis where the solution that minimizes total cost is pursued. Finding the right balance that minimizes the overall system cost including cost of transmission investment, cost of operation and cost of unserved load constitutes the transmission network expansion planning (TNEP) problem [2].

Under a cost-benefit planning framework, it is challenging to accurately capture the plethora of operating points that can occur. For this purpose, historical data of demand and wind across the different system nodes can be used. Nevertheless, the computational complexity of accommodating thousands of operating points in a large-scale Mixed Integer-Linear (MILP) planning model is highly problematic. Therefore, it is highly desirable to analyse the original dataset of historical operating points and select a small set of representative scenarios that can lead to efficient planning decisions. At the high level, there are three major approaches for tackling this task.

The simplest and most practical approach is heuristic selection, where a few scenarios are selected as representative snapshots by experts. They are chosen on the assumption that they describe critical states of the system according to the variations of the load

^{*} Corresponding author.

E-mail addresses: mingyang.sun11@imperial.ac.uk (M. Sun), fei.teng09@imperial.ac.uk (F. Teng), i.konstantelos@imperial.ac.uk (I. Konstantelos), g.strbac@imperial.ac.uk (G. Strbac).

Nomenclature			
<i>Sets and indices</i>		p_{γ}^{\max}	Maximum stable generation level of generator γ (MW)
Ω_T	Set of operating points, indexed t	$W_{t,w}$	Wind power injection from unit w at operating point t (MW)
Ω_B	Set of network buses, indexed b	b_l^E	Susceptance of existing line l (p.u.)
Ω_G	Set of all generators, indexed g	b_l^C	Susceptance of candidate line l (p.u.)
Ω_T	Set of thermal generators, indexed γ	F_l^{\max}	Power flow limit of transmission line l (MW)
Ω_W	Set of wind generators, indexed w	u_l	Originating bus of line l
Ω_{A1}^C	Set of existing transmission lines, indexed l	v_l	Terminating bus of line l
Ω_{A1}^C	Set of candidate transmission lines, indexed l	<i>Decision variables</i>	
Ω_A	Set of existing and candidate transmission lines, indexed l	n_l	Binary variable to build line l
<i>Input parameters</i>		$f_{t,l}^E$	Power flow on existing transmission line l (MW)
C_l	Cost of building line l (\$/year)	$f_{t,l}^C$	Power flow on candidate transmission line l (MW)
C_{γ}	Generation cost of thermal unit γ (\$/MWh)	$\theta_{t,n}$	Phase angle at bus n (rad)
V	Value of lost load (\$/MWh)	$p_{t,\gamma}$	Output of thermal generator γ (MW)
τ_t	Duration of operating point t (hour)	$p_{t,w}$	Output of wind generator w (MW)
π_t	Weighting of operating point t (scalar)	$u_{t,b}$	Load curtailment at bus n (MW)
$D_{t,n}$	Demand at bus n at operating point t (MW)		

and RES outputs [3]. Naturally, this approach lacks systematic selection criteria and presents inherent limitations regarding complexity.

A second approach is that the TNEP mathematical formulation can be modified to explicitly accommodate probabilistic input data. For example, in Ref. [4], point estimate methods are employed to describe the operation of a Micro-Grid; a full probabilistic description of state variables is obtained while numerically evaluating it at only at a few points. Fuzzy methods have also been applied in a similar vein (e.g. Ref. [5]). However, two major drawbacks of such point-estimate techniques are that (i) they are based on parametric marginal distribution functions which may not be a good fit for demand and RES injection variables and (ii) they do not capture dependence between variables which is a critical issue in a multivariate setting.

A final approach is the use of clustering techniques. In general, clustering is the task of grouping objects in such a way that objects in the same group are more similar to each other than to those in other groups. As such clustering can be a systematic way for 'information compression'. Clustering algorithms can be categorized as connectivity-based (e.g. hierarchical clustering), centroid-based (e.g. k-means clustering), model-based (e.g. Gaussian mixture model (GMM)), and density-based clustering (e.g. DBSCAN). Connectivity-based methods construct a hierarchy of groups based on distance metrics, whereas centroid-based clustering aims to identify a pre-determined number of central points that minimize a given centre-to-point distance metric. In contrast, model-based and density-based clustering techniques focus on the distribution and the density of the input data respectively. Comprehensive comparisons among these clustering techniques have been performed in terms of their statistical performance in Ref. [6]. Various applications of clustering methods to power systems have been proposed in the past, such as wake effect analysis in wind farm [7], characterizing electricity load profiles [8], reliability constrained congestion management [9], assessing the RES potential [10], grouping high-dimensional stochastic variables in power systems [11] and clustering of electricity consumption behavior dynamics [12].

Recently there has been some limited work on the topic of

scenario selection in planning problems. In Ref. [13] the authors employ k-means clustering to select representative operating points for investment on wind generation. In Ref. [3], the k-means algorithm is used to cluster system operation based on energy prices and non-controllable power injections. In Ref. [14], k-means is used to generate bounded intervals of demand and wind production levels for resilience-driven generation planning. The k-medoids method was employed in Ref. [15] for selecting operating points to carry out TNEP. A self-adaptive clustering technique is proposed in Ref. [16] to deal with the wind variability that characterizes the TNEP using historical wind data. Another application is to use of k-means to enhance the computational performance of Benders decomposition for multi-area TNEP [17].

All the above approaches select representative scenarios by directly clustering data in the input domain (e.g. energy price, demand), (referred to as 'input-based'), with the advantage of straightforward implementation. However, it is important to highlight that in the case of TNEP, the optimal investment decisions are not linearly related to the input variables. Therefore, direct clustering approaches may not be efficient since the effect of a chosen input scenario on the objective of the TNEP problem cannot be known a priori. To this end, authors in Ref. [18], in recognition of the fact that power flow patterns are key drivers for transmission investment, use a moment-matching algorithm to cluster operating points on the basis of their optimal power flow (OPF) patterns. According to numerical examples, the proposed OPF-based scenario selection algorithm indeed leads to a more effective reduction in the number of scenarios required to obtain the optimal investment decisions. In addition, an effective operational state aggregation technique has been proposed in Ref. [19] to select representative scenarios based on the line benefit. However, this algorithm needs to solve a relaxed TNEP problem with all the operating points, resulting in an increased computational cost before the clustering procedure. Recently, a novel method has been presented in Ref. [18] for the TNEP; important operating points are selected based on the expected power transfer of each corridor. Approaches such as this, where clustering takes place in terms of some output variable, are referred to as 'effect-based'. However, no proposal has been yet made to cluster operating points in terms of

Table 1
Overview of the scenarios selection methods for TNEP.

Method	Example References	Advantages and Disadvantages
Input-Based	[13] [15], [16] [17]	<p>Advantages:</p> <ul style="list-style-type: none"> ✓ Straightforward implementation. <p>Disadvantages:</p> <ul style="list-style-type: none"> – Not efficient since the effect of a particular operating point on the TNEP problem cannot be known a priori.
Effect-Based	[18] [19] [20]	<p>Advantages:</p> <ul style="list-style-type: none"> ✓ Generally more efficient than input-based methods. <p>Disadvantages:</p> <ul style="list-style-type: none"> – Further computational effort required to solve individual operation problems (albeit in isolation which typically translates to negligible complexity compared to the full problem at hand).
Objective-Based	Presented approach	<p>Advantages:</p> <ul style="list-style-type: none"> ✓ Highly efficient scenario selection method ✓ The proposed bi-level framework succeeds in capturing well a fully representative scenario set. <p>Disadvantages:</p> <ul style="list-style-type: none"> – Further computational effort required to solve individual operation problems (albeit in isolation which typically translates to negligible complexity compared to the full problem at hand).

the investment decisions themselves (referred to as ‘objective-based’ since identifying investment is the objective of the TNEP). The most important contribution of the present paper is that we demonstrate the superior performance of the proposed objective-based framework for solving the TNEP problem. An overview of the advantages and disadvantages of the existing and proposed methods is given in Table 1. Another topic of interest is clustering validation; analysing whether the chosen operating points effectively represent the full set of available data. Similar to the work carried in Ref. [21], which compares the impact of different scenario reduction techniques on the stochastic unit commitment problem, more work is required for identifying appropriate clustering technique for the TNEP problem.

1.2. Research questions and contributions

Three major research questions that pertain to the application of clustering-based approaches to TNEP can be summarized as:

- (1) Which variables should the clustering be based on? The choice can include combinations of variables in the input domain (e.g. demand and/or renewable injection), in the operational decision domain (e.g. power flows, bus angles) and in the investment decision domain (e.g. lines built or investment cost);
- (2) Which is the most appropriate clustering technique to be applied for a chosen set of variables (e.g. centroid methods, mixture models etc.);
- (3) After clustering the different scenarios, how to select the representative profile of each cluster (e.g. mean value or median point).

In this paper, we focus on exploring the first two issues of clustering variables and clustering techniques, and also analyse the impacts of selection methods. In summary, this work contributes to the existing literature on the following points:

1. A novel objective-based scenario selection framework is proposed by performing bi-level clustering on a combination of two variables; the incurred investment costs for candidate lines and the power flow patterns in existing lines.

2. A comprehensive analysis is performed, for the first time, to understand the suitability and efficiency of clustering variables and clustering techniques in the TNEP problem.

1.3. Analysis approach

Given a historical demand-wind dataset that consists of N observations of $m = |\Omega_N| + |\Omega_W|$ interdependent injections and loads $z_{t,n}$, where $t \in \Omega_T$ is the operating point and $n \in \{1, \dots, m\}$. We denote each operation point by the vector $\vec{z}_t \in \mathbb{R}^m$ and the set of all observations as $\mathcal{Z} = \{\vec{z}_t | t \in \Omega_T\}$. Our aim is to identify a set of operating points \mathcal{Z}^\dagger so that $C^*(\mathcal{Z}) = C^*(\mathcal{Z}^\dagger)$, where C^* denotes the optimal system cost, as shown later in Equation (1). To achieve this, we compare different clustering approaches; a comprehensive analysis framework has been developed that enables the evaluation of each approach's performance in terms of the TNEP solution accuracy.

The framework is shown in Fig. 1 and can be summarized in the following steps:

Step 1. We first solve the full TNEP problem considering all $|\Omega_T|$ available scenarios in \mathcal{Z} to establish $C^*(\mathcal{Z})$ which constitutes the ‘ground truth’ benchmark in terms of investment decisions and optimal cost.

Step 2. Given a number of clusters K , we perform one of four scenario selection approaches (denoted A1, A2, A3 and A4) with different combinations of clustering framework (i.e. single-level and bi-level) and clustering variables (i.e. demand-wind patterns, power flow patterns, and investment costs) to derive the set of representative scenarios \mathcal{Z}^\dagger . More information on this step is provided in Section 3.

Step 3. For each of the above methods, six different clustering techniques are applied to obtain \mathcal{Z}^\dagger , k-means, k-medoids, hierarchical average, hierarchical complete, hierarchical ward and GMM. More information on the different clustering methods is provided in Section 4.

Step 4. The optimal investment decisions n^\dagger are obtained by solving the TNEP with the reduced data set \mathcal{Z}^\dagger .

Step 5. The optimal operational cost $C_0^*(n^\dagger, \mathcal{Z})$ over the entire dataset \mathcal{Z} under the previously-identified investment scheme n^\dagger is determined.

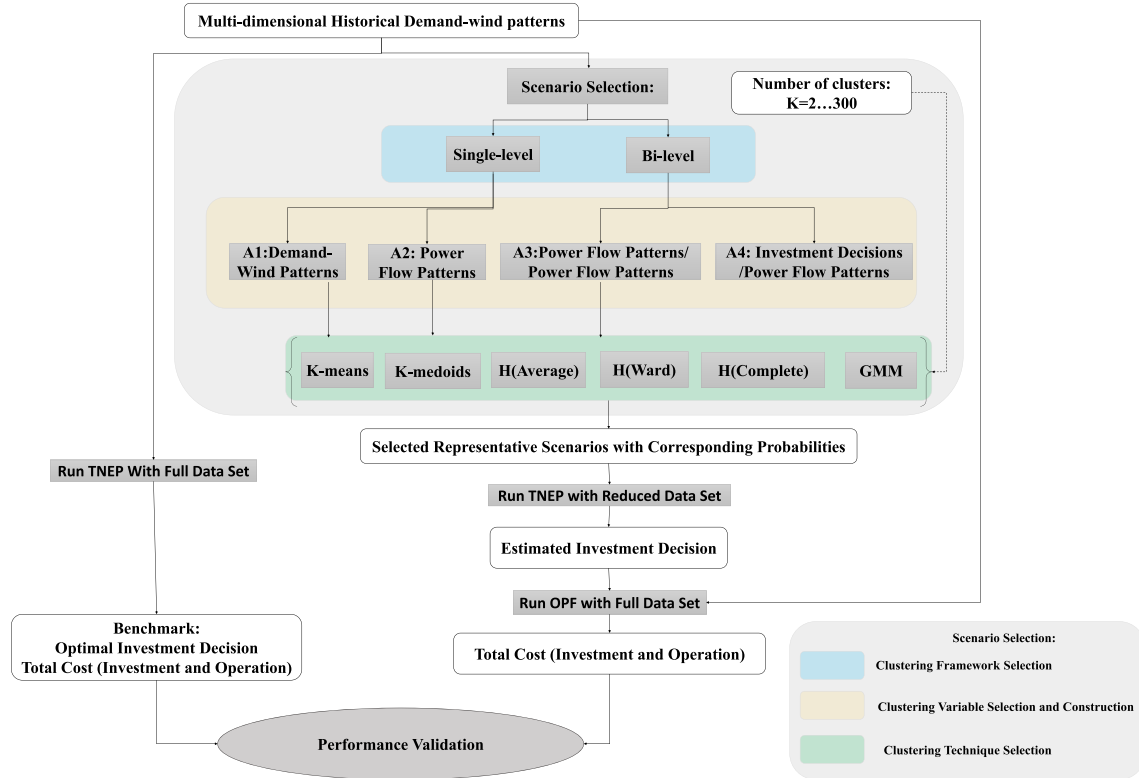


Fig. 1. Workflow of the tested scenario selection methods and the performance validation process.

Step 6. The optimal system cost (i.e. sum of investment and operation cost) is computed as $C^*(\mathcal{Z}^\dagger) = C_I(n^\dagger) + C_O(n^\dagger, \mathcal{Z})$

Step 7. The solution $C^*(\mathcal{Z}^\dagger)$ obtained in step 6 is compared to with the benchmark $C^*(\mathcal{Z})$ obtained in step 1.

The rest of paper is structured as follows: Section 2 presents TNEP model. Section 3 gives an introduction to various types of clustering variables and different clustering techniques for the TNEP problem. Section 4 explicitly introduces the structure of the proposed bi-level clustering based scenario selection framework. The case study results are presented in Section 5 to compare different clustering frameworks, clustering variables, and clustering techniques. Section 6 contains the concluding remarks.

2. Transmission network expansion planning model

2.1. Sources of stochasticity

With the introduction of shiftable load elements and the large-scale integration of intermittent energy sources, the vast number of stochastic variables beyond the transmission system operator (TSO)'s control result in a significant increase of uncertainty in the transmission operation and planning problem. In general, the stochastic variables beyond the operator's control include:

- Load levels (active and reactive);
- Uncontrollable renewable generation injections (e.g. wind, solar, etc.);
- External injections/withdrawals (e.g. cross-border imports/exports);

In this paper, we focus on the multivariate stochastic variables of demand and wind generation output as the sources of stochasticity

for the TNEP problem, by directly employing the historical measurements as the scenarios that represent the various operating conditions.

2.2. Mathematical formulation

In terms of the mathematical formulations, DC-power flow and AC-power flow are two types of transmission network modeling approaches for the TNEP problem [20]. For practical purposes, AC power flow model can be approximated via the linearized DC power flow model with the advantages of unique solution, reliability, and efficiency [21], especially in the context of the large-scale integration of intermittent generation sources. In addition, DC models, transportation models, hybrid models, and disjunctive models have been introduced and compared in Ref. [22]. Concretely, the transportation model is derived by relaxing the nonlinear constraint, an expression of Ohm's law, for the DC model, however, resulting in a possible higher investment cost than the optimal solution of DC model. On the other hand, the hybrid model integrates the characteristics of the transportation model and the DC model and only consider the Kirchhoff's voltage law (KVL) constraint for the existing circuits. In Ref. [23], the disjunctive model has been adopted to solve a stochastic TNEP problem with the consideration of the uncertainties of load and RES. Also, the DC model has been modified and used in Ref. [24] for solving a stochastic transmission expansion planning (STEP) problem. In this paper, the conventional DC model is used and adapted for solving the TNEP problem considering multivariate dependency in load and wind outputs. However note that the presented methodology can be readily applied to other variants of the TNEP formulation such as one employing AC power flow.

In this work, the objective of the TNEP problem is to minimize the total system cost which is the sum of generation cost, load

curtailment cost and line investment cost. When considering the variability in multivariate load and RES, the TNEP model can be formulated as a mixed integer-linear program as in Ref. [26]. It is imperative to highlight that, compared with the case in Ref. [26] which only captures the dependence between total load and total wind output, the proposed TNEP model takes into account the inter-spatial dependence of load and renewables output at various places which is an important factor for transmission network investment planning. Built on this foundation, our work also proposes a novel scenario selection framework for the TNEP problem to tackle the large number of scenarios that are introduced by considering the inter-spatial dependence.

$$C^*(\mathcal{Z}) = \min_{n,p,u} \{C_I(n) + C_O(n, \mathcal{Z})\} \quad (1)$$

where

$$C_I(n) = \sum_{l \in \Omega_A^C} n_l C_l \quad (2)$$

$$C_O(n, \mathcal{Z}) = \sum_{t \in \Omega_T} \tau_t \cdot \pi_t \left(\sum_{\gamma \in \Omega_\Gamma} p_{t,\gamma} C_\gamma + \sum_{b \in \Omega_B} u_{t,b} \text{VOLL} \right) \quad (3)$$

s.t.

$$n_l \in \{0, 1\}, \forall l \in \Omega_A^C \quad (4)$$

$$\sum_{g \in \Omega_G} p_{t,g} + \sum_{\{l \in \Omega_A^C: v_l=b\}} (f_{t,l}^E + f_{t,l}^C) - \sum_{\{l \in \Omega_A^C: u_l=b\}} (f_{t,l}^E + f_{t,l}^C) - u_{t,b} = 0, t \in \Omega_T, \forall b \in \Omega_B \quad (5)$$

$$0 \leq p_{t,w} \leq P_W^{\max} \cdot W_{t,w}, t \in \Omega_T, \forall w \in \Omega_W \quad (6)$$

$$0 \leq p_{t,\gamma} \leq P_\gamma^{\max}, t \in \Omega_T, \forall \gamma \in \Omega_\Gamma \quad (7)$$

$$-F_l^{\max} \leq f_{t,l}^E \leq F_l^{\max}, t \in \Omega_T, \forall l \in \Omega_A^E \quad (8)$$

$$f_{t,l}^E = b_l^E (\theta_t, u_l - \theta_{t,v_l}), t \in \Omega_T, \forall l \in \Omega_A^E \quad (9)$$

$$-n_l F_l^{\max} \leq f_{t,l}^C \leq n_l F_l^{\max}, t \in \Omega_T, \forall l \in \Omega_A^C \quad (10)$$

$$-M(1 - n_l) \leq f_{t,l}^C - b_l^C (\theta_t, u_l - \theta_{t,v_l}) \leq M(1 - n_l), t \in \Omega_T, \forall l \in \Omega_A^C \quad (11)$$

In the objective function (1), transmission investment cost and operation cost are included as the first and second term respectively. The investment cost C_I is the annualized capital cost of building candidate lines; it is a function of the binary variables n_l as described in constraints (4). The expected operational cost of each scenario (t) shown in (3) is equal to the products of the weighting (π_t) of scenario t , operating hours (τ_t) and the sum of the dispatch cost of conventional generators and the cost of curtailed demand penalized at the Value of Lost Load (VOLL). Constraint (5) represents the power balance equation for each system bus; the locally-produced power as well as the incoming/outgoing power flows must equal the demand minus the load curtailment. Maximum capacity limits for wind generators and thermal plants are

introduced in constraints (6) and (7) respectively. Finally, constraints (8)–(11) denote thermal capacity limits as well as DC power flow constraints for existing and candidate transmission lines. In particular, the standard DC power flow constraints (9) are enforced in the case of existing lines. For candidates lines, a big-M formulation is employed to express the disjunctive constraint (11) which becomes active only the line n_l is built and inactive in the case that $n_l = 0$. It should be mentioned that security problem is not considered in this research. Benders' decomposition [27] is one of the most popular techniques for solving Mixed Integer Linear problems. In this work, the proposed two-stage TNEP model can be decomposed into a master problem and sub-problems via Benders' decomposition. Note that a multi-cut formulation has been adopted as in Ref. [28].

3. Clustering variables and clustering techniques

3.1. Clustering variables

The first step of the scenario selection problem is to choose the variables upon which clustering will take place. For the TNEP problem, scenario selection is typically done by clustering directly the operating points \vec{z}_t of \mathcal{Z} . For example, demand-generation patterns multiplied by the nodal price were used as clustering variables to select snapshots from historical data in Ref. [3]. In Ref. [14], a polyhedral set was built by clustering demand and wind power generation data to characterize variability for TNEP. The wind generation was also selected as a clustering variable in Ref. [16] to generate scenarios based on a proposed self-adaptive clustering technique. However, clustering based on the information in the input domain may not lead to an efficient scenario reduction in the TNEP problem as some significantly different scenarios can lead to identical investment decisions. To this end, instead of considering the demand-generation patterns, the clustering process proposed in Ref. [16] was based on the power flow patterns that measures the benefits of the investments in the candidate lines, calculated by solving the TNEP problem for each scenario based on a Network Capacity Unconstrained Economic Dispatch (NCUED) model. Specifically, overloading scenarios relate to the network congestion cases that lead to a potential increase in the operation costs, whereas non-overloading scenarios should also be considered when the investments can decrease network losses. In Ref. [17], the benefits of potential network reinforcements are used as clustering variables; these were quantified and calculated by solving a relaxed version of the TNEP problem. Recently, a novel selection method has been proposed in Ref. [18] that performs the clustering based on the expected power transfer of each corridor of the system. However, the clustering variables proposed in the existing literature are still only marginally more relevant. Therefore, it may be more reasonable to perform the clustering procedure directly on the decision variables that arise if the TNEP problem was solved for each historical operating scenario. This is due to the fact that the solutions of the TNEP are not linearly related to the input demand-generation patterns due to the complexity of the network structure and the non-linear nature of the optimization problem. In this context, two new clustering variables are considered in this paper:

- 1) Active power flows $\mathcal{F} = \{\vec{f}_t, \forall t \in \Omega_T\}$, the power flow patterns over the transmission network are highly related to transmission congestions and losses, conveying information on potential investments; Note that the active power flows are obtained by solving the original TNEP problem for each scenario,

proposed in Section 2, rather than a NCUED model [18] or a relaxed model [19].

- 2) Investment costs $\mathcal{J} = \{C_1 n_l | \forall l \in \Omega_L^C\}$, the capital cost associated with each of the new-built lines are most relevant metrics for TNEP problem.

To perform clustering based on these two new clustering variables, a TNEP problem is firstly solved for each individual scenario n . Given the temporal mismatch between investment and operation (TNEP objective function is expressed in terms of annual costs, whereas each scenario has duration of 1 h) the duration of each scenario is made equivalent to the whole investment horizon. Subsequently, the outputs (power flows over transmission lines or investment costs of each new-built transmission lines) are used to group the operating points into clusters.

3.2. Clustering techniques

When solving the TNEP problem, the input scenarios are not linearly related to the optimization outputs of operational cost and investment decisions. Therefore, it is important to assess the performance of different clustering techniques for this specific problem; a clustering technique with good statistical performance may not guarantee an accurate solution. A large number of clustering techniques have been proposed and investigated in the past. However for practical purposes in this paper, we focus on some particularly popular and widely used clustering techniques in power systems according to the power systems literature. The clustering techniques considered here can be classified into three categories: centroid-based clustering (k-means, k-medoids), connectivity-based clustering (hierarchical clustering), and distribution-based clustering (GMM).

As one of the most popular unsupervised clustering algorithms, the k-means clustering technique [29] is used to classify data into K clusters by means of an iterative procedure, where K is an a priori defined integer value. Although k-means clustering is a technique that can be rapidly deployed and has high computational efficiency, the quality of clustering is highly sensitive to the randomly initialized centroids and the number of clusters K . Based on the k-means clustering and the medoid shift algorithm, k-medoids clustering method aims to minimize the sum of dissimilarities between the data points assigned in a cluster and its corresponding central point [30]. The partitioning around medoids (PAM) algorithm is one of the most widely used k-medoid clustering. The difference between k-means and k-medoids methods is that the mean value in each cluster is replaced with the median, which may be more appropriate for the TNEP problem as no new samples are generated [15]. Nevertheless, this method also suffers from high sensitivity to the initial conditions and the choice of K .

Hierarchical clustering constructs a hierarchy of clusters by employing a measure of similarity between groups of data measurements rather than relying on a pre-defined number of clusters [31]. There are two approaches to hierarchical clustering; divisive clustering (top-down approach) and agglomerative clustering algorithm (bottom-up approach); the latter is considered in this research. In terms of the intergroup similarity, different linkage criteria have been proposed with varieties of properties [32]. In this paper, we implement average linkage, complete linkage, and ward linkage in order to evaluate and compare their performance in the proposed scenario selection framework. From the statistical point of view, hierarchical clustering has the deficiency that the result cannot be re-evaluated and further adjusted due to its deterministic nature.

Beyond centroid and connectivity models, an alternative

approach is finite mixture model technique, where the whole dataset is described as a mixture of parametric distributions. Mixture models tackle the issue of determining the optimal number of clusters by employing information criteria [33]. Mixture modeling is a ‘soft’ clustering approach that ascribes a probability measure of classification to each data point. The Gaussian distribution is employed for fitting continuous data, which is suitable to the case of continuous demand and power injection data. The optimal parameters that maximize the likelihood of a Gaussian mixture model are usually estimated by the Expectation-Maximization (EM) algorithm [34].

4. Frameworks for different clustering approaches

This section presents in detail the frameworks that are used by the four clustering approaches outlined in Section 1. In particular, two frameworks combined with different clustering variables are proposed. In the case of approach A1, the scenario selection process is performed based on the input variables (i.e. demand-generation patterns); this approach has been widely used in the literature (e.g. Ref. [3]). For the cases A2 to A4, the proposed clustering frameworks (i.e. single-level and bi-level) as well as the clustering variables (i.e. the power flow patterns and the investment decisions obtained via solving the original TNEP problem for each scenario) are all the original contributions of this work.

4.1. Single-level framework

Conventionally, Single-level (SL) framework in Fig. 2 has been used for clustering-based scenario selection. Based on this framework, the clustering approaches A1 and A2 are explicitly described as follows:

Approach 1 (A1): For the TNEP problem, scenario selection is typically done by clustering directly the operating points \vec{z}_t of \mathcal{Z} . This is the most straightforward approach for scenario reduction where no further consideration is made regarding how the state variables of interest depend on the input operating points.

Approach 2 (A2): Indirect clustering approach based on clustering power flows across all the lines is applied in A2, which can be described as follows:

Step 1. Given the multi-dimensional historical demand-wind input data \mathcal{Z} , the TNEP problem is first solved for each individual scenario (total N problems) by assuming the scenario repeats across the whole investment horizon.

Step 2. From the solution, we form the dataset $\mathcal{F} = \{\vec{f}_t, \forall t \in \Omega_T\}$. The set \mathcal{F} consists of $|\Omega_T|$ vectors, where

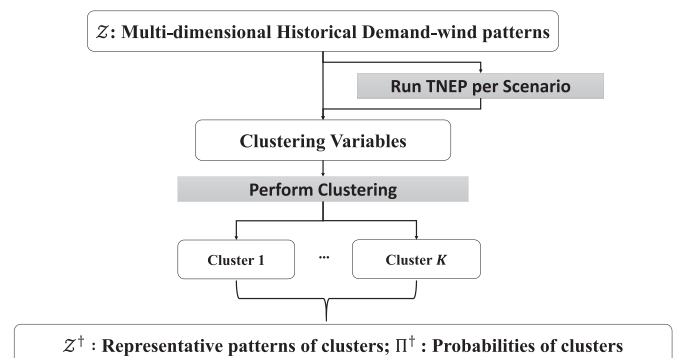


Fig. 2. The structure of conventional scenario selection framework.

each vector $\vec{f}_t \in \mathbb{R}^{|\Omega_\Lambda|}$ describes the power flows across all existing and candidate lines i.e. $|\Omega_\Lambda| = |\Omega_\Lambda^E| + |\Omega_\Lambda^C|$.

Step 3. A clustering technique is applied to partition \mathcal{S} in K clusters. We define $\Omega_{T,k}$ as the set of all operating points that belong to cluster k .

Step 4. We proceed by ‘mapping’ the clusters obtained across the dataset \mathcal{S} to the domain of the input demand-wind data \mathcal{Z} . For each cluster $k \in 1, \dots, K$ we create the set $\mathcal{Z}_k = \{\vec{z}_t, \forall t \in \Omega_{T,k}\}$ and we define the vector $\vec{z}_k^\dagger \in \mathbb{R}^m$, as the mean of the set \mathcal{Z}_k . Note that instead of the mean, the median can also be used.

Step 5. We define the reduced scenario set $\mathcal{Z}^\dagger = \{\vec{z}_k^\dagger, k = 1, \dots, K\}$. The probability set of \mathcal{Z}^\dagger is defined as $\Pi^\dagger = \{\pi_k, k = 1, \dots, K\}$ that includes the weighting of each representative scenario \vec{z}_k^\dagger, π_k , calculated by $\pi_k = |\Omega_{T,k}|/|\Omega_T|$.

4.2. Bi-level framework

The proposed bi-level framework is motivated based on the fact that usually only a limited number of scenarios will explicitly drive investment. However, the large number of remaining scenarios should not be neglected (since they play a major role in the cost-benefit determination) and should be properly represented within the problem, in an aggregated form. To this end, we present a novel bi-level framework to select scenarios that accurately represent both sets; the set of operating points that directly drive investment and the set of investment that do not result in new investment. In the proposed TNEP model, transmission lines can be categorized into existing lines Ω_A^E and candidate lines Ω_A^C for investment. In the operational level, it is effective to select the representative scenarios by clustering the operating scenarios based on the power flows over all transmission lines as the operational cost is highly related to the power flow patterns. However, the problem at hand is driven by the power flows over the candidate lines. Clustering based on the power flows over all transmission lines would under-weight the impact of this need. Consequently, it may be more efficient to select and represent scenarios based solely on power flows over candidate lines \mathcal{S}^C . Specifically, it is important to focus on the non-zero power flow patterns of candidate lines because they can basically reflect the need of transmission expansion. However, in practice, scenarios without investment usually occupy a large proportion of all the scenarios, which renders it unreasonable to group them into a single cluster. In fact, these scenarios should be further classified according to their power flow patterns of existing lines. To this end, we first propose a bi-level (BI) clustering scenario selection framework based on \mathcal{S}^E and \mathcal{S}^C (A3) that is illustrated in Fig. 3 and explained in detail below:

Approach 3 (A3):

Step 1. The TNEP problem is first solved for each individual scenario; a total of N TNEP problems are solved.

Step 2. From the solution, we form the datasets $\mathcal{S}^E = \{\vec{f}_t^E, \forall t \in \Omega_T\}$ and $\mathcal{S}^C = \{\vec{f}_t^C, \forall t \in \Omega_T\}$. These sets consist of $|\Omega_T|$ vectors. Each vector $\vec{f}_t^E \in \mathbb{R}^{|\Omega_A^E|}$ describes the power flows across all existing lines and each vector

$\vec{f}_t^C \in \mathbb{R}^{|\Omega_A^C|}$ describes the power flows across all candidate lines. Furthermore, we define $\mathcal{S}_0^C = \{\vec{f}_t^C : f_{t,l}^C = 0, \forall l \in \Omega_A^C\}$

i.e. the set of vectors \vec{f}_t^C where all power flows on candidate lines are zero. The operating points where this occurs are denoted Ω_T^{C0} . In a similar vein we introduce their complementary sets \mathcal{S}_{inv}^C and Ω_T^{Cinv} . Note that $\Omega_T = \{\Omega_T^{Cinv} \cup \Omega_T^{C0}\}$.

Step 3. A clustering technique is applied to partition \mathcal{S}_{inv}^C in K_1 clusters and $\mathcal{S}_0^E = \{\vec{f}_t^E : f_{t,l}^E = 0, \forall l \in \Omega_A^E\}$ in K_2 clusters (note that in the latter case, clustering takes place on the \mathcal{S}^E dataset); we define $\Omega_{T,k}^{Cinv}$ and $\Omega_{T,k}^{C0}$ as the set of operating points that belong to cluster $k_1 = 1, \dots, K_1$ and $k_2 = 1, \dots, K_2$ respectively.

Step 4. Based on Ω_T^{Cinv} and Ω_T^{C0} obtained in step 3, we proceed by ‘mapping’ the clusters obtained to the domain of the input demand-wind data \mathcal{Z} . For each cluster $k = 1, \dots, K_1$ we define the vector $\vec{z}_k^{iC} \in \mathbb{R}^m$, as the mean of the set $\mathcal{Z}_k^{iC} = \{\vec{z}_t : \forall t \in \Omega_{T,k}^{Cinv}\}$. Similarly, for each cluster $k \in 1, \dots, K_2$ we define the vector $\vec{z}_k^{iE} \in \mathbb{R}^m$, as the mean of the set $\mathcal{Z}_k^0 = \{\vec{z}_t : \forall t \in \Omega_{T,k}^{C0}\}$. Note that, as already mentioned, the median can also be used instead of the mean.

Step 5. Having performed the above mapping for all K clusters, we define the reduced scenario set $\mathcal{Z}^\dagger = \{\vec{z}_k^{iC}, k = 1, \dots, K_1\} \cup \{\vec{z}_k^{iE}, k = 1, \dots, K_2\}$. The probability set $\Pi^\dagger = \{\pi_k^{iC}, k = 1, \dots, K_1\} \cup \{\pi_k^{iE}, k = 1, \dots, K_2\}$ includes the weighting of each representative scenario is defined as the ratio of $|\mathcal{S}_k^{iC}|$ or $|\mathcal{S}_k^{iE}|$ over the total number of operating points $|\Omega_T|$ in the original input data set \mathcal{Z} .

Approach 4 (A4):

Although \mathcal{S}^C can indirectly signify the investment solutions, different power flows' values on candidate lines may also cause inaccurate solution as the scenarios with different investment decisions can be grouped together due to the fact that most of clustering methods are based on the Euclidean distance. Alternatively we can define the set of investment cost results $\mathcal{I} = \{\vec{i}_t, \forall t \in \Omega_T\}$, where $\vec{i}_t = \{c_t n_{t,l}, l \in \Omega_A^C\}$. The capital cost associated with each of the new-built lines is the most relevant metric for the TNEP problem. To this end, the proposed objective-based scenario selection framework, termed A4, is constructed based on the investment costs of candidate lines \mathcal{I} instead of \mathcal{S}^C in the first level clustering stage.

As shown in Fig. 4, the set of operating points is split between Ω_T^{C0} and Ω_T^{Cinv} , where the former is the set of operating points that do not give rise to any investment whereas the latter result in non-zero investment cost. Same as A3, the scenarios without investment are clustered based on the power flow patterns of existing lines (level 2). It is important to highlight that another advantage of A4 is that the upper bound to K_1 can be defined as the combination of unique investment vectors \vec{i}_t . For example, given the number of candidate lines $|\Omega_A^C| = 5$, there are $2^5 = 32$ possible combinations of investment decisions/investment costs. In practice, solving the individual this number will lead to a number significantly smaller than $2^{|\Omega_A^C|}$ because some investment combinations will never happen.

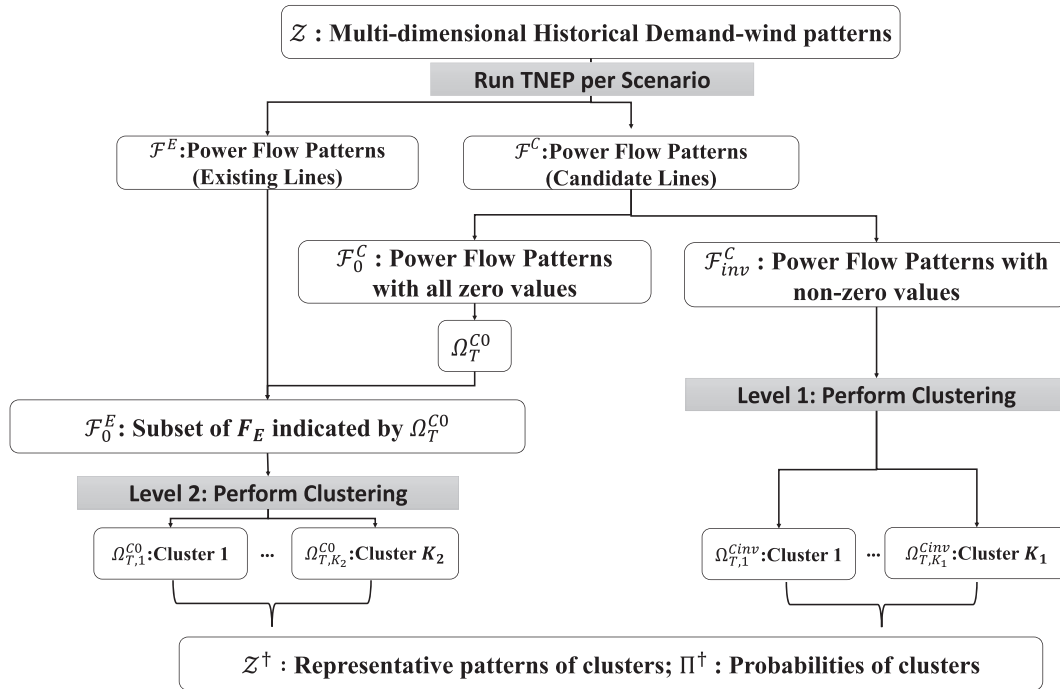


Fig. 3. The structure of the proposed power flows-based scenario selection framework.

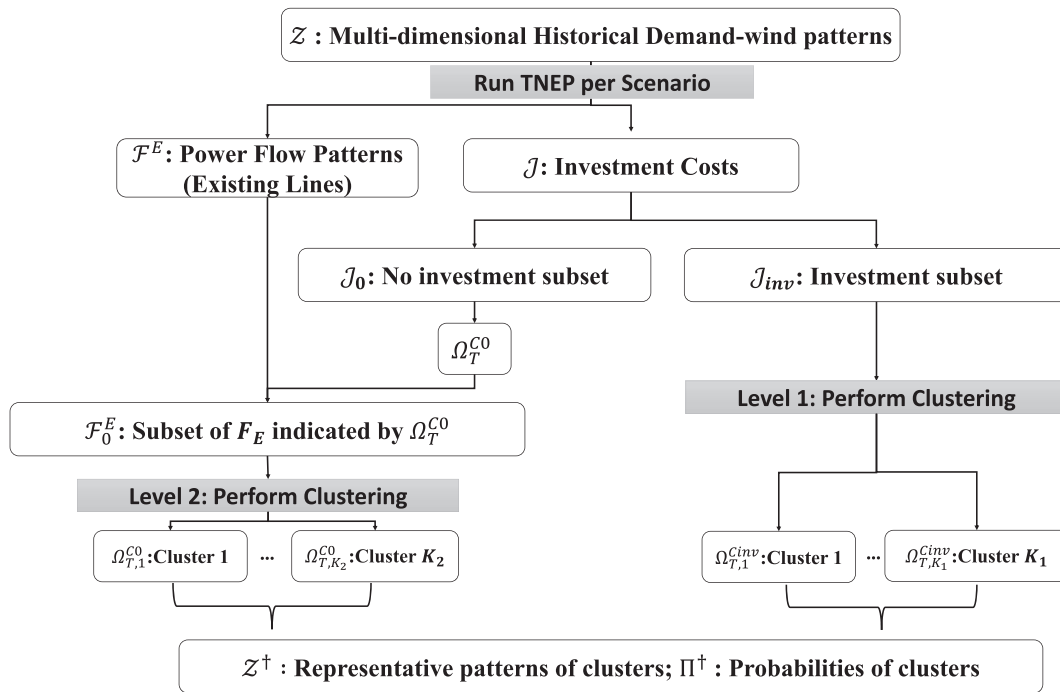


Fig. 4. The structure of the proposed objective-based scenario selection framework.

5. Simulation study and results analysis

5.1. Historical data and test system

To demonstrate the performance of the proposed framework, the original IEEE 118-bus system consisting of 54 generators and 186 existing transmission lines has been modified by including 10 wind farms each of size 100 MW as well as 20 more candidate

transmission lines each with capacity 1000 MW. For the historical database of variables, a large library of historical measurements of load and wind power injections was provided by RTE, the French Transmission System Operator (TSO). The library consists of 14,251 measurements at 15-min time intervals and spans over 7000 load buses and 200 wind plants [35]. In this work, we focus on a subset; a 128-dimensional dataset consisting of 118 demand buses and 10 wind generators chosen randomly. To map the selected variables

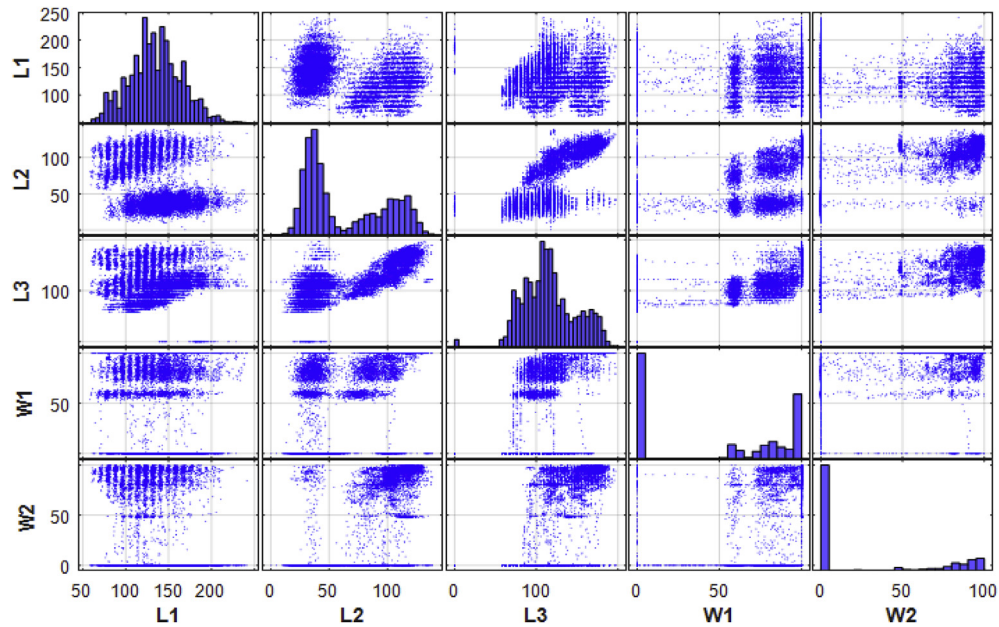


Fig. 5. Histograms of marginals distributions and bivariate scatter plots of active loads (first three variables; MW) and wind power output (two last variables; MW) for randomly selected buses in France.

Table 2
Candidate lines for the IEEE 118-bus system.

Line ID	Lines	Cost (\$ million/year)	Line ID	Lines	Cost (\$ million/year)
1	(8–9)	5.49	11	(89–90)	17.95
2	(9–10)	5.80	12	(83–84)	19.80
3	(26–30)	15.48	13	(69–75)	18.30
4	(64–65)	5.44	14	(17–113)	5.42
5	(68–116)	0.91	15	(77–78)	2.79
6	(49–51)	20.55	16	(76–118)	9.79
7	(70–74)	19.85	17	(105–106)	9.85
8	(44–45)	16.22	18	(103–110)	27.19
9	(80–96)	27.30	19	(114–115)	2.34
10	(92–100)	44.25	20	(12–117)	21.00

onto the test system, the historical dataset is scaled by calculating the ratio between the maximum coincident peak demand of the original data and the sum of active power demand across all buses defined in the system. By this way, the temporal correlations and inter-spatial dependencies between the 3 variables can be retained. An example in Fig. 5 shows the histograms and bivariate scatter plots of five variables taken from the French transmission system; variables L1, L2 and L3 are loads at different locations while W1 and W2 are wind injections from two wind farms. It can be observed that the variables have highly non-normal marginal distributions (diagonal histograms) and non-linear dependencies (scatter diagrams). The dependence between the different loads, between the outputs of the two wind farms as well as the dependence between individual loads and wind power injections are highly complex and non-standard. As such, for TNEP problem, it is of great potential benefit to move beyond the assumption of perfect correlation between loads at different locations and investigate the dependence

structure at the level of disaggregated variables in more detail. To this end, the proposed clustering framework is designed to select representative scenarios in a multi-dimensional space.

Table 2 shows the 20 candidate lines and their annualized investment costs that are estimated by the product of cost multipliers, cost per mile and line length [36]. Regarding the parameters of the proposed TNEP model, value of lost load V is set to \$40,000/MWh. In addition, we use $\tau_t = 8,760$ hours and a penalty factor $M = 20,000,000$.

5.2. Performance evaluation

In order to illustrate the comparative advantage of the proposed framework, the frameworks (A1, A2, A3 and A4) designed in Section 4 are compared. Furthermore, the clustering methods performed here are k-means, k-medoids clustering, hierarchical clustering with average linkage (H(Average)), with Ward linkage (H(Ward)), with complete linkage (H(Complete)) and GMM. The performances of all 24 methods are evaluated according to their estimated investment costs, operation costs and total costs. To assess solution quality of each method, the benchmark, given in Table 3, can be obtained by solving the TNEP problem with all the historical dataset of 14,251 operating points.

Note that, given investment decisions, operation costs are

Table 3
TNEP solution benchmark with all the scenarios.

	Operational Cost (\$million/year)	Investment cost (\$ million/year)	Total cost (\$ million/year)
All scenarios solution	2667.14	69.98	2737.12

computed by solving the economic dispatch problem considering all the scenarios. In this section, we firstly compare the investment decisions for each clustering method. Then sub-section 5.3 performs the overall comparison considering the total cost of each scenario selection framework under the associated clustering technique with best performance.

5.2.1. TNEP results of A1

In this case, the representative scenarios are selected by directly clustering the historical variables according to their statistical similarities via the above-mentioned clustering techniques. Note that, for this method, it is not required to solve the TNEP problem per operating point before carrying out the clustering process. From the view of statistical performance, clustering validity indicators have been proposed and broadly used to quantitatively analyse the quality of clustering schemes based on their similarities [6]. Consider that K partitioned scenario clusters with computed centroids $C = \{c_1, \dots, c_K\}$, are obtained by a clustering technique. Let X_k denote the scenarios that belongs to cluster k , for $k = 1, \dots, K$, the two indicators considered in this paper are computed as follows.

- 1) One of the commonly used indicators applied in this paper, the mean index adequacy (MIA) [35], indicates the average of the distances between the centroid of a cluster and each scenario in this cluster:

$$MIA(X, K) = K^{-1} \sum_{k=1}^K d^2(X_k, c_k) \tag{12}$$

- 2) The clustering dispersion indicator (CDI), defined as the ratio of the mean intraset distance between the scenarios in the same cluster and the intraset distance between the centroids of all the K clusters [36]:

$$CDI(X, K) = \hat{d}(C)^{-1} \sqrt{K^{-1} \sum_{k=1}^K \hat{d}^2(X_k)} \tag{13}$$

where $d(\cdot, \cdot)$, $\hat{d}(\cdot)$ represent cluster-to-cluster distance and intra-set distance (Euclidean distance), respectively, as defined in Ref. [36]. Note that both indicators have the same characteristic that a lower criterion value indicates better clustering performance. Before solving the TNEP problem, the original data set of demand-wind patterns are clustered via k-means, k-medoids clustering, hierarchical clustering with average linkage (H(Average)), Ward linkage (H(Ward)), complete linkage (H(Complete)) and GMM. Fig. 6 shows the results of adequacy comparisons among the tested clustering methods by using the CDI and MIA validity indicators for clusters counts K ranging from 2 to 300. Note that 300 was chosen as a sufficiently large number of clusters (which was validated by our analysis).

The results illustrated in Fig. 6 shows that H(Average) is the best-performing method among these tested on the historical load and wind dataset, indicated by its lowest indicator values for both of the indicators. In contrast, the limitations of GMM is indicated by the large indicator values across all the number of clusters. In addition, it is constructive to note that K-means, K-medoids, H(Complete) and H(Ward) have comparable performances in terms of the statistical similarities.

The next step is to assess the TNEP solution quality of the above clustering methods when using demand-wind patterns as clustering variables. To obtain the benchmark, the TNEP problem based on the modified IEEE-118 system is firstly solved with all the historical dataset of 14,251 operating points. The overall total cost consisting of optimal operational cost and investment cost are given in Table 3.

Based on the constructed clusters for each method, most of the representative scenarios are selected by using the mean value of each group with corresponding probabilities except the K-medoids method whose final scenarios are presented by the actual median. Fig. 7 show the TNEP solutions of investment cost obtained by performing the tested clustering methods on the historical demand-wind patterns X (14251 observations of 128 variables), for the number of clusters K from 2 to 300.

Regarding the optimal investment costs, H(Ward) and H(Complete) exhibit superior performance compared to the other techniques indicated by their ability to approach the benchmark solution at $K = 300$, whereas the other clustering methods cannot

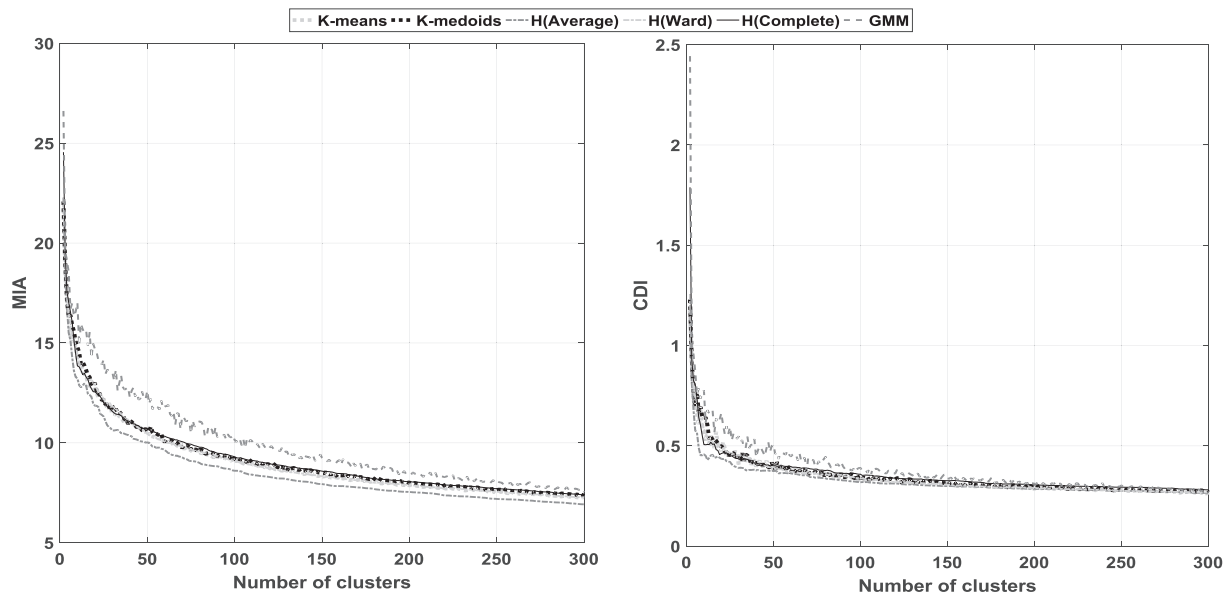


Fig. 6. Comparisons among the clustering techniques by using (left) the MIA indicator and (right) the CDI indicator for $K = 2$ to 300.

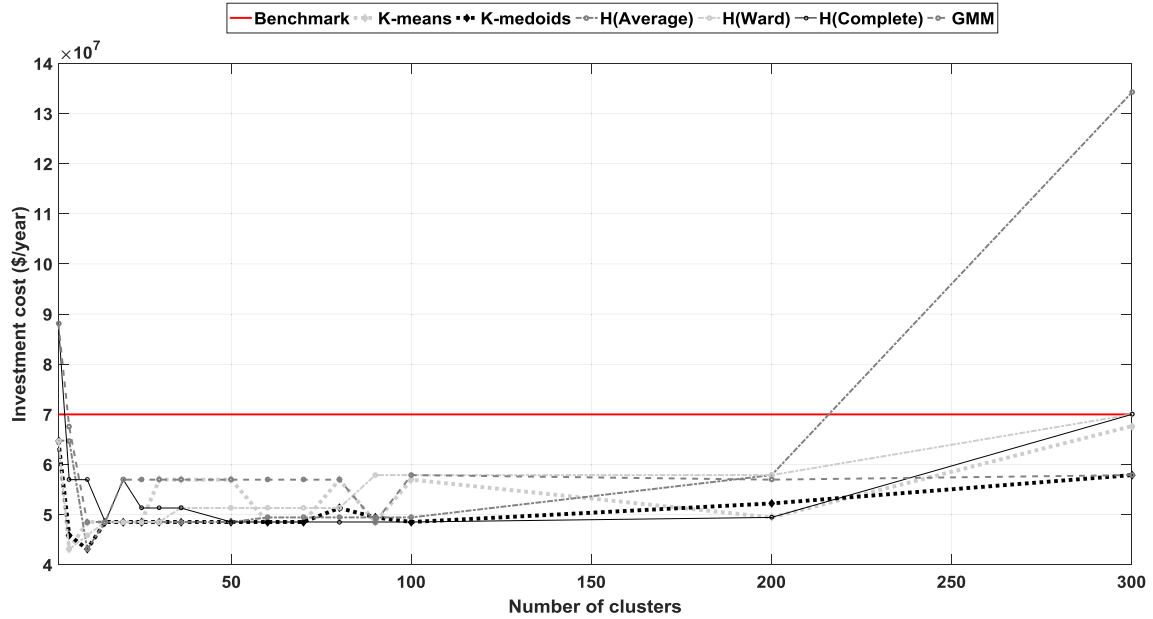


Fig. 7. Comparisons among clustering methods based on the TNEP solutions of investment costs. The input scenarios are selected via A1.

achieve the optimal solution in the tested range of K . Specifically, most of clustering methods result in an underinvestment on the transmission lines, whereas H(Average) method makes overinvestment decisions as more outliers (e.g. worst-case patterns) are detected by this method with higher weights. In addition, model-based clustering method, GMM, has relatively stable performance after about $K = 20$. For the centroid-based clustering methods, solutions of K-means are slightly closer to the target value than K-medoids which indicates that taking the mean value of each cluster has better performance than using the actual median in this case. When comparing Fig. 7 with Fig. 6, it is imperative to note that the validity indicators are not good predictors of accuracy regarding the eventual TNEP solution i.e. the insights gained from a validity indicator analysis do not match the TNEP results. This demonstrates the importance of investigating the performance of a clustering

method beyond their inherent statistical properties (i.e. intra-cluster dispersion) in order to determine their suitability to such a complex problem.

5.2.2. TNEP results of A2

After solving the TNEP problem for each scenario, clustering variables \mathcal{F} are formed that consists of \mathcal{F}^E (14,251 observations of 186 variables) and \mathcal{F}^C (14,251 observations of 20 variables). Applying different clustering methods on \mathcal{F} , the results of indices and corresponding probabilities are assigned to the historical stochastic variables X for the number of clusters K from 2 to 300. According to the constructed groups for each K , representative scenarios are used as the final inputs for the proposed TNEP problem with a certain probability for each operating point. The

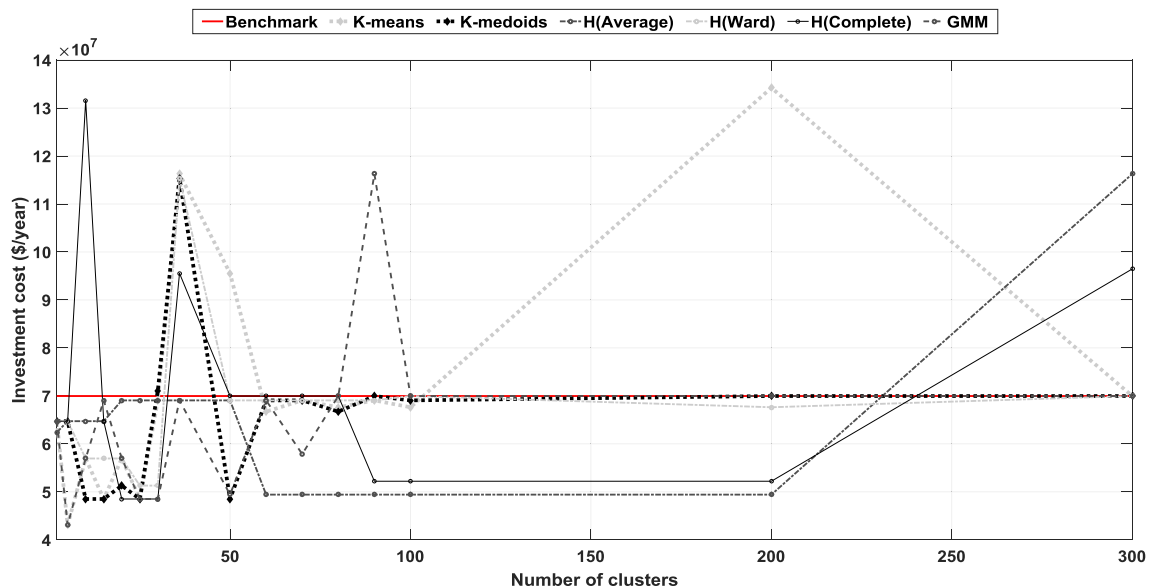


Fig. 8. Comparisons among clustering methods based on the TNEP solutions of investment costs (A2).

TNEP solutions of investment cost for A2 are shown in Fig. 8. In this case, K-medoids performs the best in this case that the solutions achieve a relatively accurate level from $K = 80$ and converge to the exact benchmark investment cost at $K = 200$. Then GMM and H(Ward) methods also exhibit high efficient to approach approximate but not exact optimal solutions that may also result in huge different in terms of the total cost. On the contrary, H(Average), H(Complete) and K-means methods cannot achieve the benchmark value in the tested range of number of clusters, and even increase to an overestimated value from about $K = 100$, $K = 230$, and $K = 240$, respectively.

5.2.3. TNEP results of A3

For A3, the constructed \mathcal{F}^E and \mathcal{F}^C are further classified into a series of useful dataset for clustering according to the power flow patterns of candidate lines. In particular, 3594 scenarios are extracted from \mathcal{F}^C to form the first stage clustering variables, \mathcal{F}_{inv}^C , which only containing the non-zero power flow patterns. Afterward, the second level clustering variables \mathcal{F}_0^E of size $10,657 \times 10$ are indicated and constructed based on \mathcal{F}_0 that result in all zeros power flow values (no investments) in \mathcal{F}^C . Note that although nearly 75% percent of scenarios are included in \mathcal{F}_0^E , the most influential scenarios that lead to investment are identified by \mathcal{F}_{inv}^C . To this end, we determine the number of clusters $K_1 = 2, \dots, 300$ for the first stage and then set $K_2 = 5$ for the second level clustering for convenience.

Using the selected scenarios obtained from A3, the results of investment for the tested clustering methods are shown in Fig. 9. All the methods achieve the benchmark results at the end of tested range of $K = K_1 + K_2$. In details, K-means presents the most inefficient performance that obtains the steady optimal solution from roughly $K = 90$. Afterward, GMM and K-medoids converge to the target investment cost at $K = 70$ and $K = 80$, respectively. Comparing to the centroid-based and the distribution-based clustering techniques, the tested connectivity-based methods exhibit better solution qualities with the increasing number of K , especially for H(Ward), the fastest approach to reach the exact optimal investment decisions from $K = 36$.

5.2.4. TNEP results of A4

Finally, the proposed objective-based scenario selection framework is implemented based on the constructed power flow patterns of the existing lines \mathcal{F}^E and the occurred investment costs of the candidate lines \mathcal{F} . In the first clustering stage, subset \mathcal{F}_{inv} are constructed with 3594 vectors of non-zero investment costs from \mathcal{F} . Then the subset \mathcal{F}_0^E for the second-level clustering is formed by choosing the corresponding vectors from \mathcal{F}^E indicated by Ω_T^{C0} . It is imperative to note that one of the important benefits of performing the first stage clustering on \mathcal{F} is the determined maximum number of clusters that requires for this problem. Mathematically, there should be $2^{20} = 1,048,576$ possible combinations of investment cost as it is a constant value for each candidate line. However, only 56 sets of expansion plan actually occur in the investment solutions when solving each scenario independently. Therefore, $\bar{K} = 56$ is defined as the maximum essential number of clusters because it already leads to exactly identical investment decision in each cluster, and hence it is meaningless to further increase the number clusters. Nevertheless, for the purposed of comparing with other methods, the number of clusters are also defined as in Section 5.2.3 (i.e. $K_1 = 2, \dots, 300, K_2 = 5$).

The TNEP solutions of the proposed objective-based framework (A4) are shown in Fig. 10. Among the tested clustering techniques, GMM requires the most number of scenarios to obtain the expected result ($K = 60$). In addition, it is evident that the scenarios selected via A4 exhibit superior performance when using the centroid-based and connectivity based clustering techniques. When applying K-means, the most efficient method for A4, only 15 selected scenarios are required to approaches the target investment cost. Meanwhile, both of H(Ward) and H(Complete) achieve the accurate investment cost from $K = 20$. Then K-medoids and H(Average) converge to the benchmark solution at $K = 25$ and $K = 30$, respectively.

5.3. Overall comparison

In this part, a comprehensive comparison is given among the overall performance of different scenario selection methods to highlight the contributions of the proposed framework.

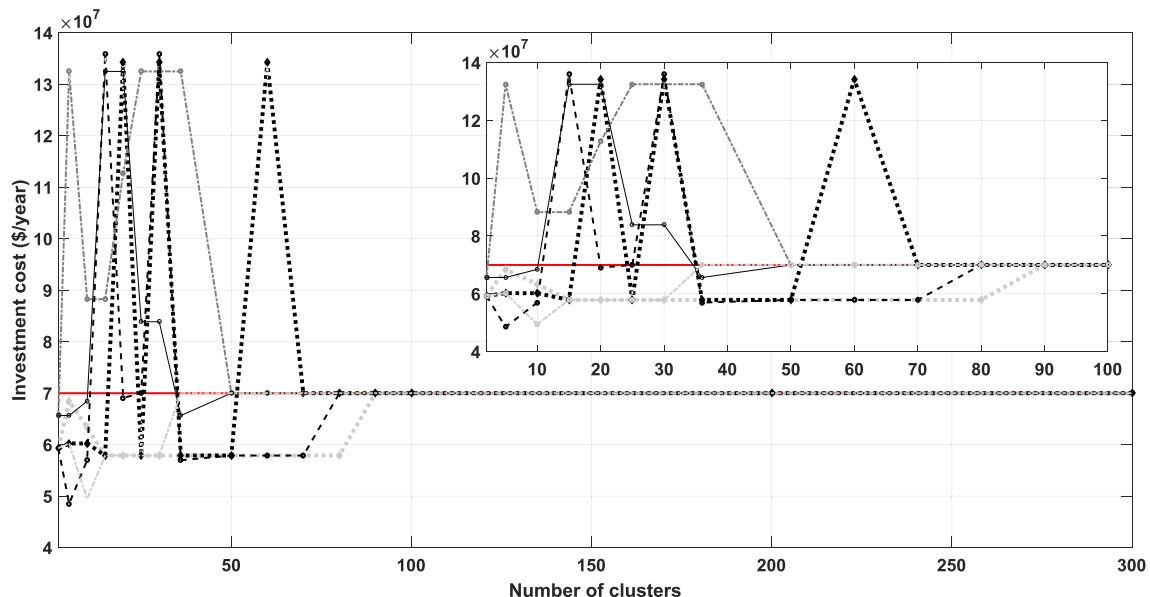


Fig. 9. Comparisons among clustering methods based on the TNEP solutions of investment costs (A3).

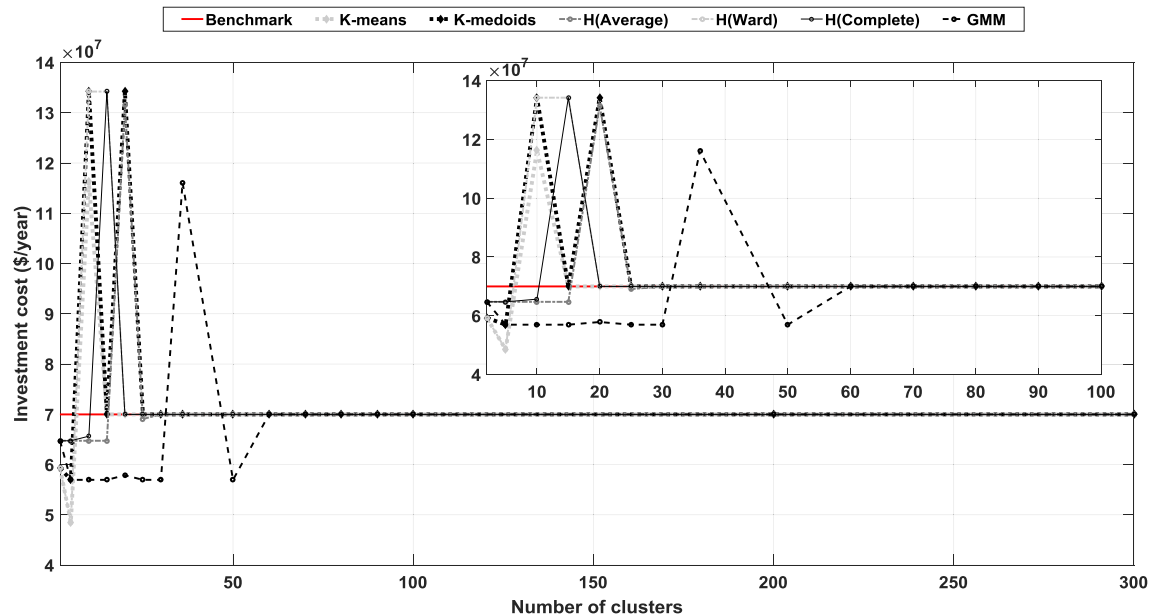


Fig. 10. Comparisons among clustering methods based on the TNEP solutions of investment costs (A4).

Firstly, scenarios selection based on conventional clustering variable, demand-wind patterns (A1), have the benefits of simple implementation and readily comprehensible clustering procedure. However, it suffers from the issues of inaccurate and underestimated investment costs with the increasing number of clusters even for the best-performed method, H(Ward), whose estimated solution approaches the optimal value at $K = 300$.

To this end, the optimal power flow patterns of all transmission lines \mathcal{S} are proposed to be used as clustering variables in the conventional framework (A2) to decrease the required number of scenarios for the optimal TNEP solution. Comparing Fig. 8 with Fig. 7, it is constructive to note that the estimated investment costs of all the clustering technique is improved by using the power flow patterns as clustering variables, especially for the K-medoids method that achieves the benchmark objective investment cost steadily after $K = 200$ clusters.

Nevertheless, as shown in Fig. 9, an obvious improvement has been observed when considering the power flow patterns of existing lines and candidate lines as well as investment and no-investment scenarios separately (A3). Regarding the network expansion plan, the proposed OPF based bi-level clustering framework with H(Ward) exhibits the best estimation that achieves the optimal decisions from $K = 36$, resulting in approximately 80% scenarios reduction. This enhancement is contributed to the fact that the scenarios with investment in the full TNEP problem should also lead to investment when solving the problem per operating points. However, it is important to note that the contrary of the above statement is not always true because it is a probabilistic problem when solving the TNEP problem considering all the scenarios together. For example, the 9th candidate line is planned to be built in some scenarios when solving the problem individually, however, it is not worth to invest this line eventually according to the optimal investment decision obtained by solving the full problem.

Although such great achievements have been attained by A3, the main contribution of this paper is to propose the idea that it is the most efficient way to select the representative scenarios via clustering directly on the objective of problem, which is the expansion plan for TNEP problem. Consequently, the objective-

based scenario selection framework (A4) presents its superior performance than all the other methods that it can reach the optimal solution at $K = 15$, which is about 50% further reduction of A3, when using the most efficient clustering technique, K-means, in this case. Overall, the best-performed clustering technique for each method can be concluded as follows: H(Ward) is the most appropriate clustering technique for A1 and A3; Centroid-based clustering techniques, K-medoids and K-means, can be regarded as the suitable clustering techniques for A2 and A4, respectively.

Fig. 11 shows the optimal total costs associated with the transmission expansion plan obtained by solving the TNEP problem using the tested methods A1, A2, A3, and A4 with their most appropriate clustering techniques. It is evident that the performance regarding the total costs is highly consistent with the results concluded from the view of investment costs.

Note that the tested scenario selection methods and the TNEP optimization problem were implemented in MATLAB and FICO Xpress, respectively, and run on an Intel Xeon E5-2690 PC with 8 cores. For comparison, time consumptions of solving the full TNEP problem as well as the TNEP problem with the minimum required number of scenarios for each type of selection method with its best-performed clustering technique are given in Table 4. Comparing to the computation times of the full problem, it is constructive to highlight that the computational cost has been successfully reduced from hours to seconds when employing the clustering-based scenario selection methods. In details, clustering based on OPF patterns (A2) can reduce about 47% of the computational cost comparing to A1, while the power flow patterns based bi-level clustering method (A3) further improve A2 by saving roughly 70% computing times. Finally, another 73% reduction in the computational time of A2 is accomplished via the proposed objective-based framework (A4).

6. Conclusions

This paper proposes a novel objective-based scenario selection framework to choose the representative operating points for the TNEP problem with high penetration of RES. Beyond the conventional clustering variable, demand-generation patterns, we propose

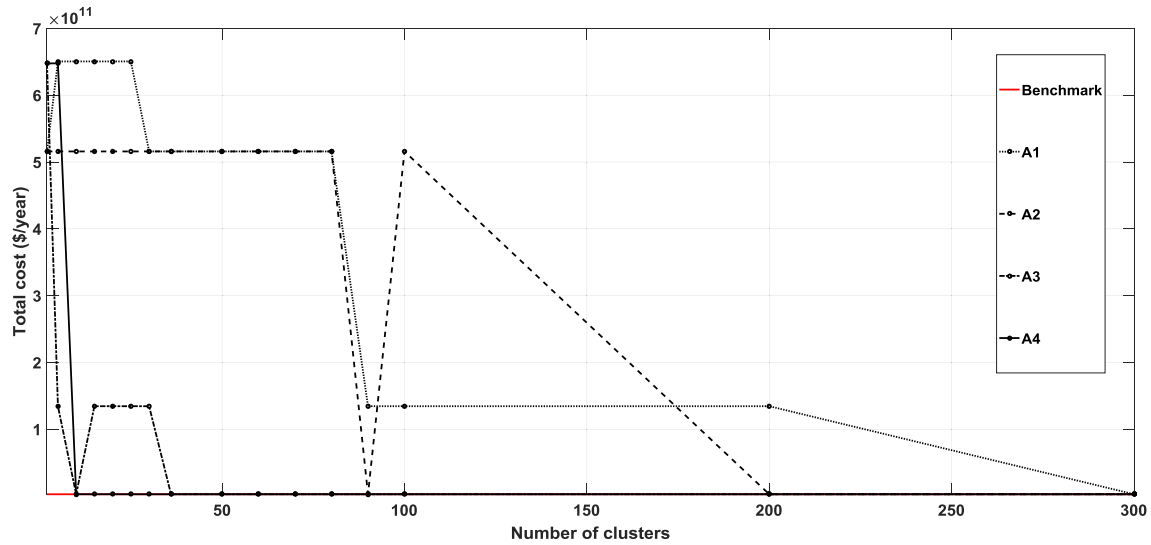


Fig. 11. Comparisons among the TNEP solutions of different selection algorithms according to total cost, based on the best-performed clustering method for each algorithm (i.e. A1:H(Ward), A2: K-medoids, A3: H(Ward), A4(K-means)).

Table 4
TNEP solutions and computation Times(Seconds).

	Optimal clustering technique	Minimum required K	CPU Times(s)
Benchmark	—	14,251	1.47×10^5
A1	H(Ward)	300	198.28
A2	K-medoids	200	103.78
A3	H(Ward)	36	30.13
A4	K-means	15	8.07

three new clustering variables \mathcal{F}^E , \mathcal{F}^C , and \mathcal{F} , allowing the historical scenarios to be classified based on their effects. In addition, conventional clustering based framework and the proposed bi-level clustering framework has been introduced and integrated with the abovementioned clustering variables to construct four types of tested scenario selection methods: clustering based on demand-wind patterns (A1), clustering based on all power flow patterns (A2), bi-level clustering based on power flow patterns (A3), objective-based bi-level clustering (A4). Among varieties of clustering techniques, we analyse the unsupervised clustering techniques: centroid-based clustering (K-means, K-medoids), connectivity-based clustering H(Average), H(Complete), and H(Ward)), and distribution-based clustering (GMM). To evaluate their performance, the selected scenarios are used to solve the TNEP problem based on a modified IEEE-118 system with real historical data in France.

One major conclusion stemming from the analysis is that the proposed objective-based scenario selection framework (A4) with K-means exhibits superior performance when compared to the other methods, indicated by its fastest convergence speed to approach the optimal TNEP solutions. In addition, another key advantage of the proposed method is that we can define the maximum number of clusters that required for the TNEP problem by counting the actual number of existing combinations of expansion plans which is far less than the theoretical number $2^{|\Omega_s^C|}$ in practice. Furthermore, this work identifies three major directions in terms of the clustering based scenario selection method: (1) selection of clustering variable; (2) selection of clustering technique; (3) selection of representative scenario in each constructed cluster.

According to the results, we conclude that clustering based on the impacts of the input data indeed improves the performance of the selected scenarios obtained by clustering the input dataset itself. Regarding the power flow patterns, constructed by pre-solving the TNEP problem per scenarios, it is of interest to consider the existing lines and candidate lines as well as scenarios with and without investment, separately, realized by the proposed bi-level clustering framework. Although a great achievement has been obtained via implementing the power flow based bi-level clustering framework, directly based on the objective, investment costs, exhibit the highest efficiency to reach the optimal solution. For the second direction, various clustering techniques have been demonstrated to have significantly different performances for each scenario selection method. Finally, a primary conclusion of choosing the scenario in each cluster is that taking the actual median in each cluster is more suitable for A2 and A3, whereas using the mean value can result in a good performance for A4.

Future research should be mainly devoted to further improve the efficiency of the proposed framework, and implementing it into a larger multi-energy system with more complicated network structures. The increasing complexity of the TNEP problem, introducing by the multi-energy resources, leads to the changelings of high variability and high-dimensionality. Therefore, feature extraction and dimensionality reduction techniques can be integrated into the framework. In addition, contingency constraints could be appended to the TNEP model. Lastly, the idea of objective-based clustering can also be extended to the other applications such as generation planning and system scheduling.

Acknowledgements

This research was partially supported by the EPSRC Grant EP/K002252/1.

References

- [1] Garver LL. Transmission network estimation using linear programming. *IEEE Trans Power App Syst Sep./Oct. 1970*;PAS-89(7):1688–97.
- [2] Hemmati R, Hooshmand RA, Khodabakhshian A. Comprehensive review of generation and transmission expansion planning. *IET Gener Transm Distrib* 2013;7(9):955–64. Sept.
- [3] Agapoff S, Pache C, Panciatici P, Warland L, Lumbieras S. Snapshot selection based on statistical clustering for Transmission Expansion Planning.

- Eindhoven: PowerTech, 2015 IEEE Eindhoven; 2015. p. 1–6.
- [4] Mohammadi Sirus, Mozafari Babak, Solimani Soodabeh, Niknam Taher. An adaptive modified firefly optimisation algorithm based on Hong's point estimate method to optimal operation management in a microgrid with consideration of uncertainties. *Energy* March 2013;51(1):339–48.
- [5] Yu L, Li YP, Huang GH. A fuzzy-stochastic simulation-optimization model for planning electric power systems with considering peak-electricity demand: a case study of Qingdao, China. *Energy* March 2016;98(1):190–203.
- [6] Chicco G. Overview and performance assessment of clustering methods for electrical load pattern grouping. *Energy* Jun. 2012;42(1):68–80.
- [7] Al-Shammari, Tamah Eiman, Shamsirband Shahaboddin, Petković Dalibor, Zalnezhad Erfan, Yee Lip, et al. Comparative study of clustering methods for wake effect analysis in wind farm. *Energy* 2016;95:573–9.
- [8] Viegas Joaquim L, Vieira Susana M, Melício R, Mendes VMF, Sousa João MC. Classification of new electricity customers based on surveys and smart metering data. *Energy* July 2016;107(15):804–17.
- [9] Tabandeh Abbas, Abdollahi Amir, Rashidinejad Masoud. Reliability constrained congestion management with uncertain megawatt demand response firms considering repairable advanced metering infrastructures. *Energy* June 2016;104(1):213–28.
- [10] Grigoras Gheorghe, Scarlatache Florina. An assessment of the renewable energy potential using a clustering based data mining method. Case study in Romania. *Energy* March 2015;81(1):416–29.
- [11] Wang Y, Zhang N, Kang C, Miao M, Shi R, Xia Q. An efficient approach to power system uncertainty analysis with high-dimensional dependencies. *IEEE Trans Power Syst* 2017;99. 1–1.
- [12] Wang Y, Chen Q, Kang C, Xia Q. Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE Trans Smart Grid* Sept. 2016;7(5):2437–47.
- [13] Dominguez R, Conejo AJ, Carrion M. Toward fully renewable electric energy systems. *IEEE Trans Power Syst* Jan. 2015;30(1):316–26.
- [14] Dehghan S, Amjadi N, Conejo AJ. Reliability-constrained robust power system expansion planning. *IEEE Trans Power Syst* May 2016;31(3):2383–92.
- [15] Torbaghan SS, Gibescu M, Rawn BG, Meijden M v d. A market-based transmission planning for HVDC grid—case study of the North Sea. *IEEE Trans Power Syst* March 2015;30(2):784–94.
- [16] Li Yunhao, Wang Jianxue. Flexible transmission network expansion planning under uncertainty based on a self-adaptive clustering technique. In: *International conference on renewable power generation (RPG 2015)*. IET; 2015. p. 1–6.
- [17] Munoz F, Hobbs B, Watson J. New bounding and decomposition approaches for MILP investment problems: multi-area transmission and generation planning under policy constraints. *Eur J Oper Res* 2016;248(3):888–98.
- [18] Fitiwi Desta Z, de Cuadra F, Olmos L, Rivier M. 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* 2015;90:1360–76.
- [19] Ploussard Q, Olmos L, Ramos A. An operational state aggregation technique for transmission expansion planning based on line benefits. *IEEE Trans Power Syst* July 2017;32(4):2744–55.
- [20] Alvarez R, Moser A, Rahmann CA. Novel methodology for selecting representative operating points for the TNEP. *IEEE Trans Power Syst* May 2017;32(3):2234–42.
- [21] Du Ershun, Zhang Ning, Kang Chongqing, Bai Jianhua, Cheng Lu, Ding Yi. Impact of wind power scenario reduction techniques on stochastic unit commitment. In: *2016 second international symposium on stochastic models in reliability engineering, life science and operations management (SMRLO)*. IEEE; 2016. p. 202–10.
- [22] Hemmati R, Hooshmand R-A, Khodabakhshian A. State-of-the-art of transmission expansion planning: comprehensive review. *Renew Sustain Energy Rev* Jul. 2013;23:312–9.
- [23] Stott B, Jardim J, Alsac O. DC power flow revisited. *IEEE Trans Power Syst* Aug. 2009;24(3):1290–300.
- [24] Romero R, Monticelli A, Garcia A, Haffner S. Test systems and mathematical models for transmission network expansion planning. *IEE Proc Generat Transm Distrib* 2002;149(1):27–36.
- [26] Park H, Baldick R, Morton DP. A stochastic transmission planning model with dependent load and wind forecasts. *IEEE Trans Power Syst* Nov. 2015;30(6):3003–11.
- [27] MacRae CAG, Ernst AT, Ozlen M. A Benders decomposition approach to transmission expansion planning considering energy storage. *Energy* October 2016;112(1):795–803.
- [28] Konstantelos I, Strbac G. Valuation of flexible transmission investment options under uncertainty. *IEEE Trans Power Syst* March 2015;30(2):1047–55.
- [29] MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Berkeley, CA: Univ. California Press; 1967. p. 281–97.
- [30] Mirkes EM. K-means and K-medoids applet. University of Leicester; 2011.
- [31] Johnson SC. Hierarchical clustering schemes. *Psychometrika* Sep. 1967;32(3):241–54.
- [32] Gordon AD. A review of hierarchical classification. *J Roy Stat Soc A* Sta 1987;150:119–37.
- [33] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002;97(458):611–31.
- [34] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc* 1977;39(Series B):1–38.
- [35] Sun M, Konstantelos I, Tindemans S, Strbac G. Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems. In: *PSCC '16*; 2016. p. 1–8. Genoa.
- [36] Zhang H, Heydt GT, Vittal V, Quintero J. An improved network model for transmission expansion planning considering reactive power and network losses. *IEEE Trans Power Syst* Aug. 2013;28(3):3471–9.