

Novel Methodology for Selecting Representative Operating Points for the TNEP

Ricardo Alvarez, Albert Moser, *Member, IEEE*, and Claudia Andrea Rahmann, *Member, IEEE*

Abstract—As part of transmission network expansion planning (TNEP), a technical and economical assessment of several planning alternatives must be performed in order to ensure fulfillment of the network security criteria and to estimate the alternatives' expected operating costs. This task requires performing load flow calculations for different operating points (OPs) of the power system. Due to the high computational burden, considering all possible OPs is simply not possible. As a consequence, only a set of representative OPs is usually taken into account. Most works in the TNEP focus on issues related to optimization algorithms and modeling, neglecting the selection process of the representative OPs. Furthermore, most works only consider a few OPs, providing little or no insight about the criteria used in the selection process or about the error made when evaluating planning alternatives using a limited number of OPs. In this work, a novel methodology for selecting representative OPs to consider within the TNEP is presented. The proposal pays special attention to critical situations, where the network security may be endangered. Furthermore, the methodology allows quantifying the error made when evaluating network operation using a limited number of representative OPs.

Index Terms—Average-linkage, clustering, representative operating points, transmission network expansion planning.

I. INTRODUCTION

TRANSMISSION networks play an important role in power systems, enabling large-area energy transport from generation points to the final consumers. The planning of the transmission network is a critical task for every country and must be carefully thought out.

For the development of a transmission network expansion plan, possible alternatives must be technically and economically evaluated for different operating points (OPs) of the system. Considering all possible OPs within every year of the planning period is, given the high computational burden, not feasible. As a result, representative OPs are always selected. Many works in TNEP consider only one OP, which is normally the peak load

[1], [2]. Other publications such as [3]–[7] have considered more than one OP. Nevertheless, selection of the OPs is only done considering the load and generation situations, neglecting the network topology, allocation of the load and generating units. For instance, in [3], a total of 22 OPs were considered, which were selected based on the load curve. In [4], six OPs were taken into account, which were obtained by combining two load situations (peak and off-peak) and different wind productions. Weighting factors of each of the load flow patterns were obtained from historical data. In [5], typical days for each season were selected and the hourly load in these days was used for the TNEP. Although in this work a total of 96 OPs are considered for each year, the optimization algorithm is demonstrated in a six-busbars network (Garver system) and therefore its practical use in real transmission networks is not guaranteed. In [6], different OPs were considered through the development of load scenarios. However, no indication about the amount of OPs used or the way in which they were selected is given. Finally, in [7] different OPs were considered based on market-driven power-flow patterns. For the case study, five OPs were used, which were built for peak demand and five different generation dispatches.

A major drawback of only considering the load and generation patterns when determining representative OPs for the TNEP is that stressful situations in the transmission network do not necessarily occur in peak load conditions [8], or in situations with high amounts of variable generation technologies (VGT). Indeed, it also depends on the network topology and on the allocation of the load and generating units. Furthermore, for network security assessments, most works only focus on situations with high loading levels of the network elements, neglecting situations in which the network elements are lightly loaded. Especially in transmission systems with long lines, light load situations may result in significant amounts of reactive power flows, which can cause an increase of the voltage level outside the permissible voltage band, endangering the network security. Similar to peak load situations, cases where the network elements are lightly loaded do not necessarily correspond to situations with low load and/or low feed-in of VGT. In order to illustrate this fact, Fig. 1 shows the hourly based load and VGT feed-in for the Main Chilean Transmission System (CTS) in year 2013. In this figure, the OPs that lead to minimum and/or maximum power transfers in at least one of the corridors of the network are highlighted. It can be seen that maximum and minimum power transfer on the network corridors do not necessarily occur at high/low load situations or high/low VGT feed-in. As a consequence, ignoring the network topology may lead to an in-

Manuscript received February 17, 2016; revised July 12, 2016 and September 7, 2016; accepted September 12, 2016. Date of publication September 16, 2016; date of current version April 17, 2017. This work was supported in part by the Chilean Council of Scientific and Technological Research (CONICYT) under Grant Fondap/15110019, in part by the Complex Engineering Systems Institute under Grant ICM: P-05-004-F and Grant CONICYT: FB0816, and in part by the German Academic Exchange Service (DAAD). Paper no. TPWRS-00266-2016.

R. Alvarez is with the Energy Centre, Faculty of Mathematical and Physical Sciences, University of Chile, Santiago 52056, Chile (e-mail: rialvare@gmail.com).

A. Moser is with the Institute of Power Systems and Power Economics, RWTH Aachen University, Aachen 52062, Germany (e-mail: am@iaew.rwth-aachen.de).

C. Rahmann is with the Department of Electrical Engineering and the Energy Centre, Faculty of Mathematical and Physical Sciences, University of Chile, Santiago 52056, Chile (e-mail: crahmann@ing.uchile.cl).

Digital Object Identifier 10.1109/TPWRS.2016.2609538

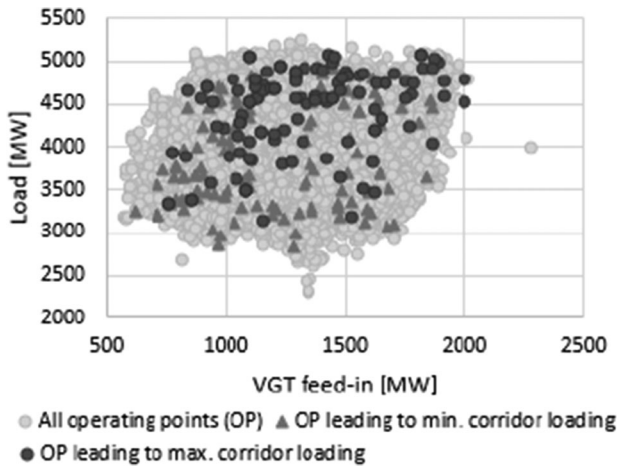


Fig. 1. Load and VGT feed-in in the CTS in the year 2013.

correct evaluation of the network security and thus to unfeasible or to suboptimal solutions within the TNEP.

With this in mind, this paper presents a novel methodology for selecting representative OPs for the TNEP. The proposal is based on Average-Linkage Clustering, which represents an agglomerative strategy within hierarchical clustering [9]. The proposed algorithm also quantifies the error made when using only a limited number of OPs in the TNEP. The clustering is based on the expected power transfer of each corridor of the system, which considers not only the load and the generation situations, but also its allocation in the network.

This paper is organized as follows: Section II presents the framework of this study; Section III presents the cluster model implemented; Section IV presents the cluster methodology for selecting representative OPs; Section V summarizes the results obtained; and Section VI the conclusions.

II. FRAMEWORK FOR SELECTING RELEVANT OPs

In this study it is assumed that the TNEP is centralized in a deregulated environment. The optimization aims to minimize the investment and operating costs of the transmission network while maintaining a predefined level of network security.

As already mentioned in the introduction, in order to evaluate the network security and to quantify the yearly operating costs of planning candidates, several power flow calculations must be performed for a set of representative operating conditions. These cases must be able to cover both, critical situations within a year where the network security may be endangered, and normal operating conditions to estimate a representative yearly operating cost. To represent critical situations from a TNEP perspective, this work considers the thermal load capability of the network elements as security criterion. The fulfillment of the thermal limits is a usual criterion used in TNEP, where the traditional approach is to use a simplified DC model of the network.

The cluster methodology for selecting the representative OPs is based on the expected power transfer of the network corridors. Critical situations are represented by:

- OPs characterized by high power transfers in the network corridors, which may lead to: 1) unacceptable loading

levels from a thermal perspective and/or 2) unacceptable low voltage profiles. Under high loading conditions, the triggering factor of a critical situation (thermal limit or steady state stability) will depend, among others, on the system state and the length of the line [10]. It is worth to mention that the clustering process does not aim to indicate which factor will lead to a security problem in the future network, but rather to identify possible OPs that may endanger the network security. Indeed, a full security analysis is assumed to be performed later on within the TNEP itself.

- OPs that lead to low power transfers in the corridors, which may lead to unacceptable high voltages at some network busbars. These OPs are important in the sizing and allocation of network reactors within the TNEP.

The main challenge to select representative OPs for the TNEP considering the power transfer in the network is that this task must be performed before the TNEP itself, i.e., when the future network is still unknown. For this study, this situation is resolved by considering a simplified model of the transmission network and calculating the expected power transfer in the corridors, rather than in single circuits of the transmission lines. Indeed, while future circuits added to the network are not known at this stage (before the planning), future corridors can be estimated more accurately. New corridors are built at this stage assuming a fixed number of circuits and a predefined type of conductor. The expected power transfer in each of the corridors is calculated performing DC-power flows for each future load and generation situation expected during the planning period. Main advantage of using DC-power flow instead of an AC-power flow is that, as explained above, the use of an AC-power flow requires assuming a future network and therefore a network plan. This can biased the results of the TNEP or lead to a wrong identification of critical situations, since they will strongly depend on the network reinforcement assumed. For example, if too little network reinforcement in a specific area is assumed, an AC-power flow computation would either not converge or lead to unrealistic high network losses. Also, an important part of the loading of the transmission line will consist of reactive power flow, which makes it difficult or even impossible to estimate the active power that needs to be transferred in the future through the lines. The use of a DC-power flow enables to estimate in a simplified way the future power transfers that need to take place within the network, without having to make assumptions regarding network reinforcements. Also, traditional convergence problems of AC power flows do not appear.

Finally, it is worth to mention that no outages are considered when calculating the expected power transfer in each corridor. As usual in planning studies, the network is stressed in the TNEP itself, usually in the form of contingency analysis including generation and transmission outages. Only at this stage can critical situations –for each planning candidate– be truly identified. If outages were considered in the process of selecting representative OPs, the starting point for the network security analysis may already contain contingency situations, which could mislead the results of the TNEP.

III. MODELING APPROACH AND BASIC DEFINITIONS

A. General

The objective of cluster analysis is to break down a data set into groups, so that the elements within one group are similar to each other and different than the elements of the other groups [11]. Therefore, large intra-cluster homogeneity and a low inter-cluster homogeneity are the goals [9]. The data set consists of a collection of objects, where each object has a number of attributes with different values. In this study, objects represent hours, where each hour is related to a specific network OP. In a network with N corridors, the N attributes of each object (hour) are the expected loading levels of each network corridor for the year under consideration, which are expressed as a percentage of its capacity. Therefore, the objective of the clustering is to group hours based on the expected loading levels of the corridors in the future.

The starting point of the cluster analysis is a matrix $X(H \times N)$, where H represents the total number of OPs within a year (hours), and N the number of corridors. The elements of the matrix X are the loading levels of the corridors for each hour within the year under consideration. Thus the element $x_{i,k}$ is the expected loading level of corridor k at hour i . The set Ω_H contains all hours under consideration, and the set Ω_N all corridors.

B. Critical Operating Regions of Corridors

In order to identify possible critical situations, the first step of the cluster algorithm is to define critical operating regions of the corridors. As shown in Section II, critical situations are represented by those OPs that lead to high/low power transfers in the network corridors. Based on the cumulative probability density function of the corridor loading level, two critical operating regions for each corridor may be defined:

- An upper critical region, which contains situations of high corridor loading, and
- A lower critical region, which contain situations of low corridor loading.

Additionally, a normal region for each corridor is also defined, which aims to characterize normal operating conditions where no network security problems are expected.

The critical regions are defined based on a minimum upper quantile \underline{Q}_U and a maximum lower quantile \bar{Q}_L of the corridor loading. An upper critical region of corridor k is assigned only if at least in one OP the power transfer through the corridor exceeds 50% of its capacity, which in this work is fixed for all corridors. The justification of the 50% for defining the upper critical operating regions is based on the fulfillment of the n-1 criterion. In this context, the worst case appears when considering a transmission line consisting only in two circuits. The n-1 criterion in this case requires that the failure of one of the circuits do not lead to a loading level above 100% in the remaining circuit (from a thermal perspective) in order to not compromise network security. Although in transmission lines consisting on more than two circuits this threshold should be higher, to apply 50% to all circuits represents a safe assumption

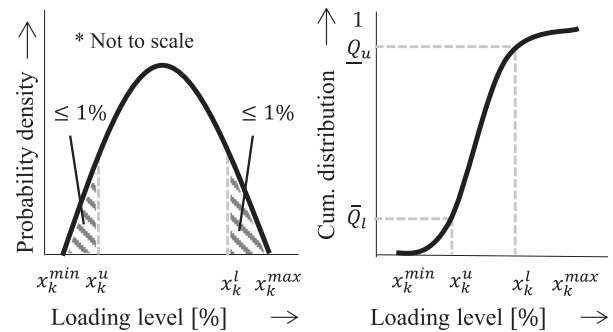


Fig. 2. Minimum and maximum critical operating regions for exemplary upper and lower quantiles of 99% and 1% respectively.

and a good *first* approximation for identifying potential critical situations. Similarly, a lower critical region is assigned to a corridor only if at least in one OP the power transfer is below 15% of its capacity, which is also fixed for all corridors. This lower limit is motivated by typical values of the surge impedance load of transmission lines, which is around 30% of its capacity. It is important to highlight that both limits (50 and 15%) do not affect the generality of the proposed methodology but only the definition of the critical regions. Indeed, different thresholds for each corridor may be also defined based on the operator experience without affecting the methodology.

Therefore, for each corridor k where a critical operating region is assigned, the limits of the lower and upper critical operating regions, x_k^u and x_k^l respectively, are calculated according to (1) and (2).

$$x_k^u = \max\{x(\underline{Q}_u), 50\%\} \quad (1)$$

$$x_k^l = \min\{x(\bar{Q}_l), 15\%\} \quad (2)$$

where the quantile function $x(Q)$ gives the corridor loading level for a given quantile Q . For example, if $\underline{Q}_u = 99\%$, the upper critical operating region contains 1% of the maximum corridor loading level, unless the expected corridor loading level for that quantile is below 50%. In that case, the upper critical region contains less than 1% of the maximum corridor loading levels. Similarly, if $\bar{Q}_l = 1\%$, then the lower critical operating region contains 1% of the lowest corridor loading levels, unless the corridor loading levels for that quantile are above 15%. Fig. 2 shows the limits of the lower and upper critical operating regions of corridor k considering 99% and 1% for the upper and lower quantiles.

C. Proximity Measures Between Two Objects

The proximity measure is a quantification of the distance or similarity between two objects (in this case, between two hours representing different OPs). Objects with a smaller distance are considered more similar than objects with a higher distance and are therefore more likely to be grouped. The proximity measure used in this work is the maximum weighted distance $d_{i,j}^{w,\max}$ between the vectors containing the corridors loading levels in

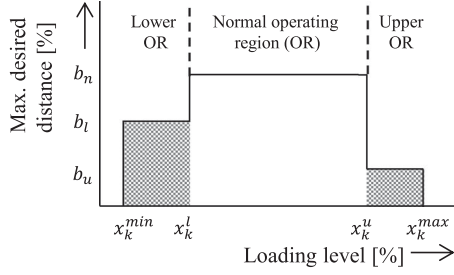


Fig. 3. Maximum desired distances for the lower, normal, and upper critical operating regions.

hours i and j , as presented in (3).

$$d_{i,j}^{w,\max} = \max_{k \in \Omega_N} \{ |x_{i,k} - x_{j,k}| \cdot \Delta(x_{i,k}, x_{j,k}, x_k^u, x_k^l) \} \quad (3)$$

where $x_{i,k} \in X$ represent the loading level of corridor k in hour i , and x_k^u and x_k^l represent the upper and lower limits of the critical operating region of corridor k , respectively and Ω_N is the set containing all corridors. The weighting function $\Delta(x_{i,k}, x_{j,k}, x_k^u, x_k^l)$ aims to penalize the distance between the loading levels of the corridors in two different hours if this exceeds a maximum desired distance b . The goal of this function is to achieve a better representation of critical situations by defining different values of b for each of the operating regions. This is shown graphically in Fig. 3, where the maximum desired distance for the critical regions is lower than for the normal operating region. In this figure b_u , b_l , and b_n represent the maximum desired distance for the upper, lower and normal operating regions, respectively.

The maximum desired distance used when calculating the proximity measures of a corridor in two hours depends on the corridor's loading levels (in both hours) and on the pertinent operating regions. In this way, the loading level of corridor k in hour i could lie in the normal operating region, while for hour j in the upper critical operating region. Since only one maximum desired distance of corridor k between hours i and j can be used ($b_k^{i,j}$), the minimum of the desired distances is selected according to:

$$b_k^{i,j} = \min \{ b_k^i, b_k^j \} \quad (4)$$

The weighting function Δ for calculating the proximity measures of the loading level of corridor k between hours i and j is shown in (5).

$$\Delta(x_{i,k}, x_{j,k}, x_k^u, x_k^l) = \begin{cases} 1/b_k^{i,j} & \text{if } |x_{i,k} - x_{j,k}| \leq b_k^{i,j} \\ \frac{(a-2) \cdot |x_{i,k} - x_{j,k}|}{2 \cdot (b_k^{i,j})^2} - \frac{a-4}{2 \cdot b_k^{i,j}} & \text{if } |x_{i,k} - x_{j,k}| > b_k^{i,j} \end{cases} \quad (5)$$

The parameter a defines the penalization function. It can be chosen arbitrarily restricted to $a \geq 2$. Fig. 4 left shows the value of the weighting function Δ depending on the distance between the loading level of corridor k in the hours i and j . Fig. 4 right shows the value of the weighted distance $w_k^{i,j} = |x_{i,k} - x_{j,k}| \cdot \Delta(x_{i,k}, x_{j,k}, x_k^u, x_k^l)$

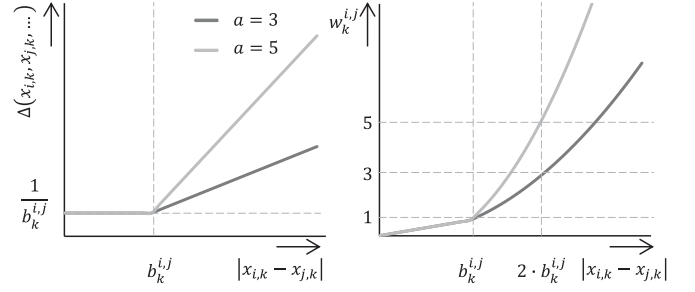


Fig. 4. Value of the weighting function (left) and of the weighted distance (right).

D. Selection of the Cluster Centrum

For this study, the cluster centrum corresponds to the hour lying closer to the center of gravity of the cluster. For this, first, the average loading level of each corridor considering all hours belonging to the cluster is calculated as stated in (6). Then, the hour with minimum distance to the center of gravity is chosen as the cluster centrum, as stated in (7).

$$x_{C_z, k} = \frac{1}{|C_z|} \sum_{i \in C_z} x_{i,k} \quad (6)$$

$$r_{C_z} = \min_{i \in C_z} \{ d_{i, x_{C_z, k}}^{w,\max} \} \quad (7)$$

where $x_{C_z, k}$ represents the average loading level of corridor k within the cluster C_z , and r_{C_z} represents the cluster centrum (representative hour) of cluster C_z . Since for each cluster its centrum (and therefore the representative hour) corresponds to one of the hours within the cluster itself, it already contains the corresponding load and generation profiles, which can be then directly used in the TNEP.

E. Definition of the Distance Between Clusters

The distance between two clusters C_y and C_z ($D_{C_y, C_z}^{\text{aver}}$), is defined as the average weighted maximum distance between the members of each cluster according to (8).

$$D_{C_y, C_z}^{\text{aver}} = \frac{1}{|C_y| \cdot |C_z|} \sum_{i \in C_y} \sum_{j \in C_z} d_{i,j}^{w,\max} \quad (8)$$

where $|C_a|$ indicates the number of hours in cluster a .

IV. CLUSTERING METHODOLOGY

An overview of the clustering methodology is presented in Fig. 5. Input data for the clustering process are 8760 operating points representing one year. For the particular case of a dynamic TNEP, this process should be performed several times, depending on the planning period, stages and scenarios.

A. Determination of Loading Level and Operating Regions of the Corridors

The first step of the methodology is to determine the loading level of each corridor for all OPs under consideration. To do this, the DC power flows throughout the network are calculated for

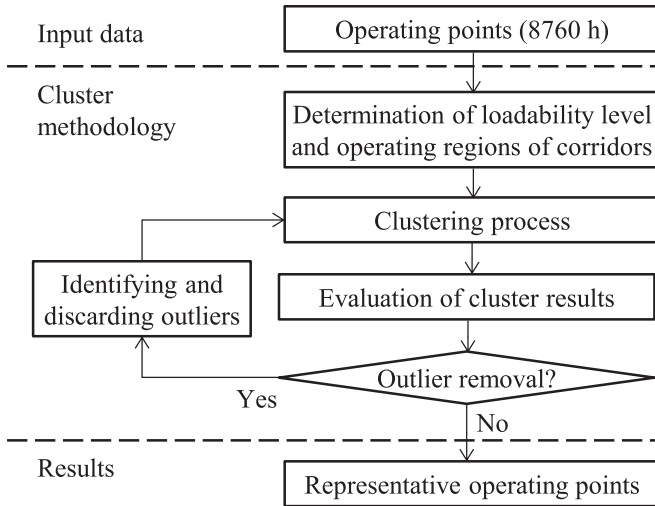


Fig. 5. Overview of the clustering methodology.

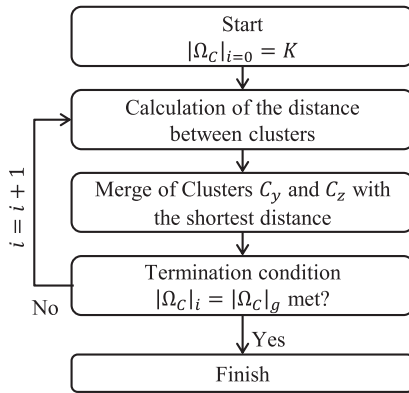


Fig. 6. Overview of the clustering process based on average-linkage.

each OP. In order to reduce the computational efforts, redundant information is discarded at this stage based on two criteria [9]:

- A corridor loaded equally over the whole year will not be further considered since it gives no relevant information for the clustering process.
- If the loading level of two corridors has a correlation above 99%, then only one of them will be considered since both give redundant information.

After the loading level of the corridors is obtained, the next step is to determine the operational regions for each corridor according to Section III-B.

B. Clustering Process

The clustering process is based on an adapted version of the average-linkage clustering [9]. The main adaptation is the proximity measure used, which considers penalization factors (see Section III-C). The clustering process using the average-linkage algorithm is shown in Fig. 6. Before the process starts, the desired final number of clusters (target), $|\Omega_C|_g$, must be defined.

As shown in Fig. 6, the clustering process is done iteratively. In the first iteration ($i = 0$), each hour represents a cluster, i.e.,

$|\Omega_C|_{i=0} = H$. For the particular case of one year, $H = 8760$. At each iteration, the process calculates the distances between clusters according to (8) and the two clusters with the shortest distance are merged. This is performed until the final number of clusters reaches the target number, i.e., $|\Omega_C|_i = |\Omega_C|_g$.

C. Evaluation of Cluster Results

In order to evaluate the results obtained in the clustering process, two different performance indexes are used. The first index evaluates the obtained clusters by quantifying how accurately the hours within a cluster are characterized by the corresponding representative hour (representativeness of the cluster centrums).

In the second index, the clusters are evaluated by quantifying how accurately the loading levels of the corridors within a year are characterized by the values corresponding to the representative hour (representativeness of the loading level of the corridors). Based on these indicators, it is possible to determine if the obtained clusters are acceptable for their use in the TNEP. If the results do not meet the expectations, a further step consisting of removing possible outliers is performed in order to improve the quality of the results (see next section).

1) *Representativeness of the Cluster Centrums*: The objective here is to determine how well the hours within a cluster are characterized by its representative hour, i.e., by the center of the corresponding cluster. To quantify this, first the maximum distance between each hour and the corresponding cluster centrum is calculated for all operating regions. However, a non-weighted distance is used in this case, since the weighted distance used for the proximity measure has no physical meaning.

The mathematical formulation to determine the non-weighted maximum distance of hour $i \in C_z$ to the cluster centrum r_{C_z} of cluster C_z is presented in (9)–(11). In these equations, the upper indexes u , l and n represent the upper and lower critical operating regions and the normal operating region respectively.

$$d_{i,r_{C_z}}^{u,\max} = \max_{k \in \Omega_N} \{ |x_{i,k} - x_{r_{C_z},k}| \mid \forall x_{i,k} > x_k^u \} \quad (9)$$

$$d_{i,r_{C_z}}^{n,\max} = \max_{k \in \Omega_N} \{ |x_{i,k} - x_{r_{C_z},k}| \mid \forall x_k^l \leq x_{i,k} \leq x_k^u \} \quad (10)$$

$$d_{i,r_{C_z}}^{l,\max} = \max_{k \in \Omega_N} \{ |x_{i,k} - x_{r_{C_z},k}| \mid \forall x_{i,k} < x_k^l \} \quad (11)$$

where $x_{i,k} \in X$ represents the loading level of corridor k in hour i , x_k^u and x_k^l represents the upper and lower limits of the critical operating region of corridor k , respectively and Ω_N represents the set containing all corridors.

In order to have a unique indicator for the representativeness of the cluster centrums, the maximum value among the maximum distances between each hour and the corresponding cluster centrum is determined. This is presented in (12)–(14) for the upper, normal and lower operating regions respectively.

$$D^{u,\max} = \max_{i \in \Omega_H} \{ d_{i,r_{C_z}}^{u,\max} \} \quad (12)$$

$$D^{n,\max} = \max_{i \in \Omega_H} \{ d_{i,r_{C_z}}^{n,\max} \} \quad (13)$$

$$D^{l,\max} = \max_{i \in \Omega_H} \{ d_{i,r_{C_z}}^{l,\max} \} \quad (14)$$

Using this definition, it can be said that an hour is being “well represented” by the corresponding cluster if all three distances lie within a given maximum desired value defined a priori for the evaluation purposes. It should be noted that this maximum desired value must not be the same as the maximum desired value defined for the weighting function.

2) *Representativeness of the Corridors Loading Levels*: The second indicator used for evaluating the quality of the clusters is to determine whether or not the loading levels of the corridors for all hours are being well represented by the centers of the corresponding clusters. The total loading error θ_k for corridor k is obtained based on the average Euclidean distance considering all hours as stated in (15).

$$\theta_k = \frac{1}{|\Omega_H|} \sqrt{\sum_{i \in \Omega_H} (x_{i,k} - x_{r_{C_z,k}})^2}, \quad i \in C_z \quad (15)$$

where $x_{i,k}$ is the original loading level of corridor k in hour i , $x_{r_{C_z,k}}$ is its value in the corresponding cluster centrum and $|\Omega_H|$ represents the total number of hours.

D. Identifying and Discarding Outliers

Outliers have a negative impact in the performance of the clustering process. They can appear for different reasons, for example, as a result of error within the data acquisition (e.g., wrong input data of VGT feed-in) or as a result of extreme (abnormal) weather situations, which lead to unusual operating conditions. If the obtained clusters do not meet predefined performance criteria, possible outliers are identified and then discarded in order to improve the clustering results (see Fig. 5).

The identification of outliers is done iteratively for each cluster. In each iteration i , all cluster centums are newly defined based on a density criteria and a given distance d_{out}^i for iteration i . For each cluster, the new cluster centrum is selected as the one that maximizes the number of surrounding hours within the distance d_{out}^i . All hours whose distance to the newly defined cluster centrum is greater than d_{out}^i are regarded as potential outliers for the corresponding distance d_{out}^i . For the first iteration, d_{out}^0 corresponds to the minimum of the maximum desired distances between all operating regions (see Section III-C). Within the iterative process, the distance d_{out}^i is continually increased and the outliers defined according to this distance are recorded. This iterative process is performed as long as no more outliers are obtained for all clusters. This is graphically displayed in Fig. 7. In this figure, an exemplary cluster with its original members is depicted, as well as the new defined cluster centums for each iteration i . A total of four iterations is shown.

As a result of the outlier identification process, each potential outlier o can be assigned to a distance d_o^{out} . Outliers (hours) with a larger distance d_o^{out} are more likely to have a greater negative impact on the evaluation of the corresponding cluster and are therefore prone to be discarded.

Finally, potential outliers are discarded one by one. For each outlier discarded, the cluster centrum is again calculated according to (6) and (7) and the obtained maximum distance $D^{r,max}$ for each operating region is again calculated according to (12)–(14).

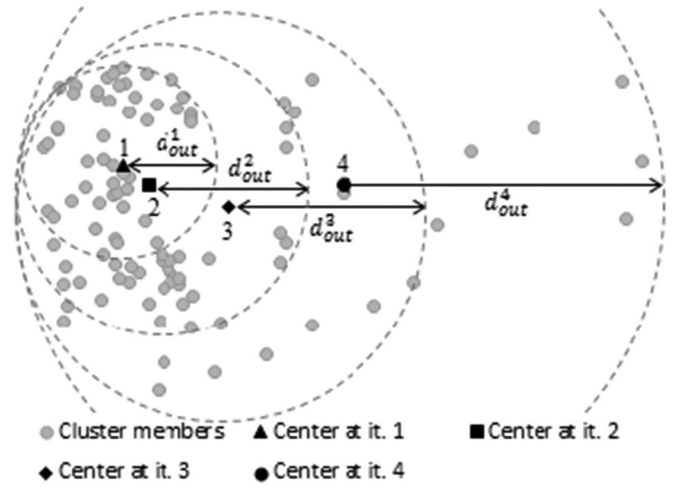


Fig. 7. Graphical display of selecting new cluster centums for discarding outliers.

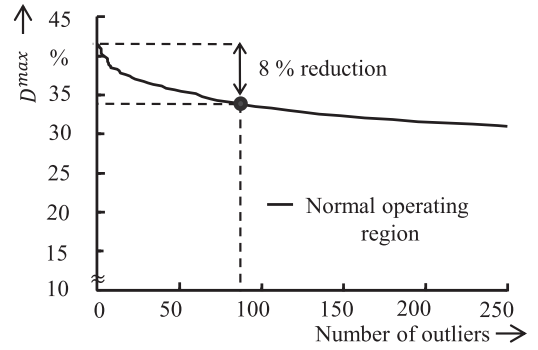


Fig. 8. Illustrative example of the identification of outlier.

Fig. 8 graphs exemplary results of $D^{n,max}$ (normal operating region), when discarding outliers. The point highlighted in the figure indicates that if 87 outliers are discarded, then an 8% reduction of $D^{n,max}$ can be achieved.

It should be noted that the discarding of outliers is restricted to those hours regarded as non-relevant. For this study, relevant hours, which cannot be discarded in this step are defined as those where

- the loading level of any corridor reaches either its maximum or minimum value and
- at least 75% of the corridor loading levels are located in each one of the critical regions.

This ensures that no relevant information is lost when discarding outliers. After discarding outliers, the clustering process must be performed again, as shown in Fig. 5.

V. CASE STUDIES

The clustering algorithm presented in Section IV was first tested in a model of the main Chilean Transmission System (CTS). The network model consisted of 110 buses and 285 lines, covering four different voltage levels between 110 kV and 500 kV. Network data and technical information of the generating units, water reservoirs and load profile were taken

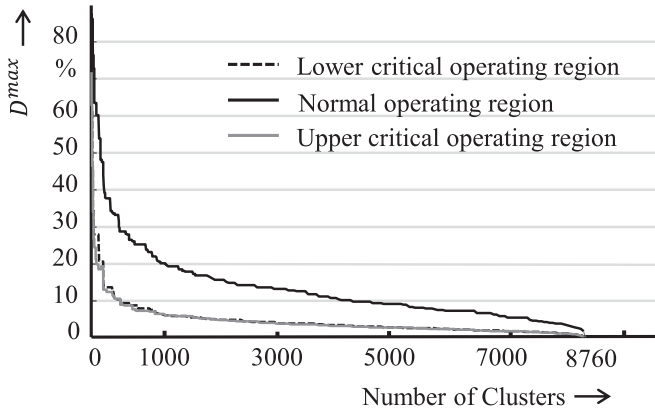


Fig. 9. Distance D^{\max} obtained for different number of clusters.

from [12]. Feed-in profiles of VGT, as well as the water inflow in the reservoirs were developed for an average hydrological year, using information from [12] and [13]. The data required to replicate the results obtained in this case study is on-line available in [14].

In total, 8760 operating points, consisting of load and generation feed-in, were considered as input data for the clustering for both years. The dispatch of the generating units was calculated using a MILP approach calculated with CPLEX according to [15].

The critical operating regions were defined using upper and lower quantiles of $Q_u = 99\%$ and $Q_l = 1\%$ respectively. A maximum desired distance between corridors loading levels for the weighting function were set to $b_n = 40\%$ for the normal operating condition and $b_u = b_l = 15\%$ for the upper and lower critical operating regions.

A. Results Obtained for Different Numbers of Clusters

Fig. 9 shows the results of the clustering process when considering different cluster targets. In this figure, the distance D^{\max} for each one of the operating regions is depicted (see equations (12)–(14)). As can be seen, at first D^{\max} decreases rapidly in all operating regions with increasing number of clusters and then slower. The break-even point, i.e., where the curve starts decreasing slower, is around 250 clusters. For this number of clusters, the distance D^{\max} for the normal operating region is around 40% and for both critical operating regions around 15%.

As can be seen in Fig. 9, the definition of a more stringent maximum desired distance for the critical operating regions ($b_n = 40\%$, $b_u = b_l = 15\%$) proves to be effective. Indeed, significantly smaller errors (which are related to D^{\max}) are obtained in the upper and lower critical operating regions as compared to the normal operating region. This means that critical operating regions are characterized with more detail by the obtained clusters, enabling a special focus to critical situations in the TNEP.

B. Representativeness for a Fixed Number of Clusters

In this section the results considering 250 clusters are analyzed. Fig. 10 shows, for a specific corridor, the cumulative

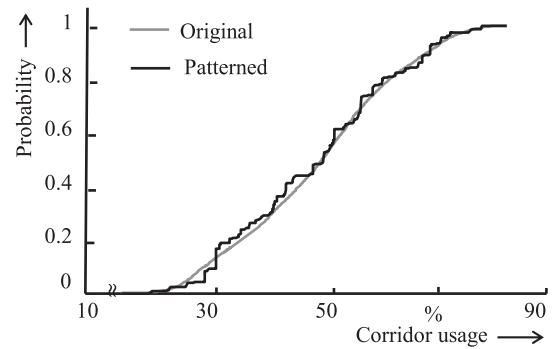


Fig. 10. Original and patterned CDF for a given corridor.

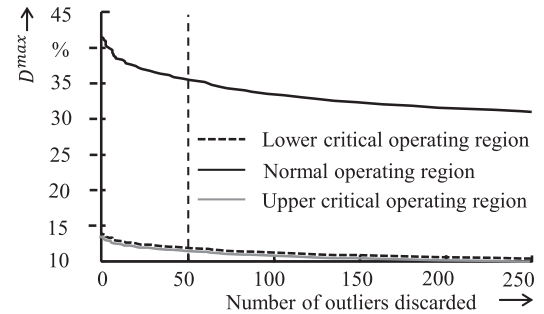


Fig. 11. Distance D^{\max} versus the number of outliers discarded.

distribution function (CDF) of the loading level considering both the original data and the data related to the corresponding representative hour. As can be seen in Fig. 10, corridor loading levels are very well patterned when using 250 representative operating points (hours).

C. Influence of Removing Outliers

In order to evaluate the influence of removing outliers, Fig. 11 shows the distance D^{\max} for each operating region for different numbers of outliers discarded. As can be seen, removing outliers is a beneficial strategy to improve the quality of the obtained clusters. For instance, discarding 50 outliers (from a total of 8760 hours, i.e. 0.57%) reduces the distance D^{\max} in the critical operating regions around 3% and around 10% in the normal operating region.

D. Effects of Different Number of Representative OPs in the TNEP

This section illustrates the importance of the number of OPs considered in the TNEP. To accomplish this, a TNEP for the year 2030 was performed considering different cases:

- In the first case the TNEP was carried out considering only one OP, corresponding to peak demand
- In the second case the TNEP was carried out considering two OPs: The peak demand and an OP representing the yearly average load.
- In the last six cases the TNEP was performed considering the OPs obtained when using the proposed methodology with 10, 20, 50, 150, 250 and 350 clusters.

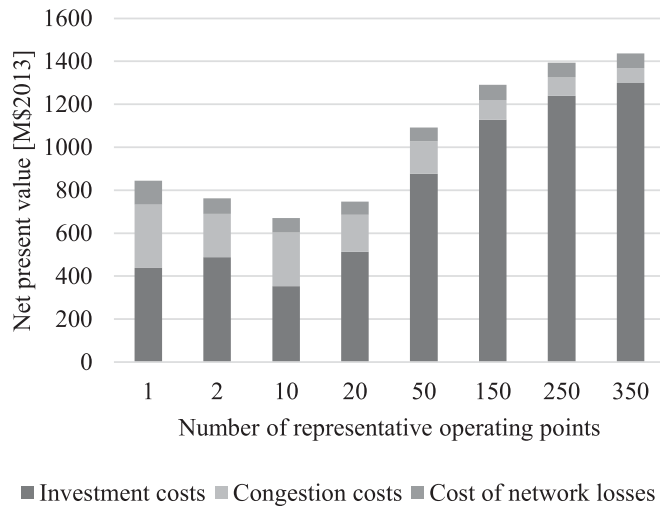


Fig. 12. NPV of the optimal expansion plans for different numbers of representative OPs.

The TNEP was solved using an adapted ant colony optimization strategy. A detail explanation of the optimization model and methodology used can be found in [15]. For this case study, a single scenario containing the generation capacity expansion, the load forecast and the price for primary energy resources was developed for the year 2030 based on [16]. Possible planning projects were built considering 12 different types of conductors and transformers, and 2 types of condenser banks and reactors. In total, around 4.500 planning projects were considered for generating planning candidates.

The Net Present Value (NPV) of the optimal expansion plans for each case is shown in Fig. 12.

As seen in Fig. 12, the total NPV of the optimal expansion plans first decreases when using less than 10 representative OPs and then starts to increase as the number of OPs grows beyond 10. Saturation of the NPV is achieved for more than 250 Ops. This indicates that only minor improvements in the results are obtained in the TNEP when considering more than 250 OPs at the expense of significantly higher computational efforts. Comparing the results obtained for the first two cases, it can be seen that including an OP representing the yearly average load leads to a higher NPV of the investment costs but to lower congestion costs and costs of network losses. This is explained because adding the OP that reflects the average load situation introduces an additional restriction for the TNEP, which leads to higher investment costs. Since the congestion and network losses for the average load situation is expected to be much smaller than those for the peak load situation, lower congestion costs and network losses are obtained.

When considering 10 OPs obtained through the proposed clustering methodology, neither the peak load situation nor the average load situation were selected. From Fig. 12 it can be seen that considering 10 OPs leads to a lower NPV of the investment costs when compared to the first two cases. This is due to the fact that the peak load situation poses a very stringent restriction for the TNEP, which cannot be covered using only 10 representative OPs. As the number of OPs increase over 10, the results of

TABLE I
RESULTS OF THE EVALUATION OF THE NETWORK OPERATION
FOR THE DIFFERENT CASES CONSIDERING ALL OPS

N° OPs	% unfeasible OP	% error congestion costs	% error network losses
1	49.8	64.7	41.0
2	50.0	4.6	9.6
10	50.3	14.5	14.7
20	17.1	9.8	17.6
50	8.3	12.0	12.7
150	3.6	0.2	0.7
250	0.7	2.5	3.7
350	0.5	0.9	0.1

the TNEP show an increase in the NPV of the investment costs and a decrease in the NPV of the congestion costs. The network losses remain more or less constant. These results can be explained because increasing the number of OPs in the TNEP introduces additional restrictions leading, therefore, to higher investment costs. As a consequence, lower congestion costs are also achieved.

To evaluate the effects of only considering a limited number of OPs, the network operation for each of the final networks was evaluated considering all operating points within a year (8760). Table I shows the percentage of OPs where the resulting network was not feasible (where the technical restrictions were not fulfilled), as well as the error obtained in the yearly congestion costs and the costs of the network losses.

As seen in Table I, the TNEP when using too few OPs results in a large number of unfeasible situations, i.e. where the technical requirements are not fulfilled. Indeed, using 50 OPs leads to 8.3 % of unfeasible situations for the whole year. The table also shows that the number of unfeasible situations obtained in the TNEP when considering 10 representative OPs is similar to the case where only the peak load was considered, although in this latter case less network reinforcement takes place. This shows that, although the peak load situation poses a very stringent restriction for the TNEP (thus leading to higher investment costs), it fails to cover a large number of critical situations.

The error made when evaluating the operating costs (congestion and network losses) is also significant when the TNEP is performed using less than 250 operating points. This shows that using only a limited number of OPs for the TNEP may lead to a significant underestimation of the transmission network expansion capacity needs. As a consequence, the obtained network plan must be reinforced in later stages.

Comparing the results of the network evaluation from Table I with the results obtained for the clustering process for different number of clusters (see Fig. 9 in Section V-A), it can be seen that the acceptable number of OPs to be used for the TNEP can be derived from the clustering results. Indeed, as shown in Section V-A, the results of the clustering process indicates that the reduction of the error (obtained from D^{\max}) starts decreasing slower after 250 clusters. This shows the advantage of being able to select the appropriate number of OPs before performing the TNEP, which is always subject to large computational efforts. This is consistent with the results obtained for the planning process, where the results show that an acceptable solution can

be achieved using 250 representative OPs. These results show the importance of choosing an appropriate number of representative OPs for the TNEP. Using too little representative OPs not only leads to a large number of unfeasible situations, but also to a large error when evaluating the network operation, which may have a negative impact when economically evaluating different planning alternatives. It must be noted that the optimal number of OPs to be used in the TNEP depends on the characteristics of the power system under study and therefore independent analysis must always be carried out.

VI. CONCLUSION

In this study, a novel cluster methodology for selecting representative OPs for the TNEP was presented. The cluster methodology selects the representative OPs for a predefined number of clusters based on the expected loading levels of the corridors, which are calculated using a simplified model of the transmission network and the yearly generation and load. Also, the methodology pays special attention to critical situations, where the network security may be compromised, by achieving a much better representation of these critical situations for a given number of clusters as compared to normal operating conditions.

The results showed that the number of representative OPs to be used within the TNEP has a major impact on the optimal expansion plans. Using too few OPs may lead to an underestimation of the transmission network expansion capacity needs and also to large errors in the estimation of the yearly network operating costs. As a consequence, the results obtained in the TNEP may lead to an unacceptable number of unfeasible situations where the security criteria is not met and to unfeasible or suboptimal expansion plans, which must be altered later for practical use. For the specific case of the CTS, the results showed that in order to achieve a valid solution, at least 250 representative OPs should be considered when performing a TNEP. Using a large number of OPs within the TNEP leads to much more accurate solutions, but at the expense of large computational effort. The further development of optimization strategies that can cope with this restriction is therefore critical, if its results are meant to be used for the TNEP of actual networks.

REFERENCES

- [1] A. Escobar, R. Gallego, and R. Romero, "Multistage and coordinated planning of the expansion of transmission systems," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 735–744, May 2004.
- [2] J. Choi, A. El-Keib, and T. Tran, "A fuzzy branch and bound-based transmission system expansion planning for the highest satisfaction level of the decision maker," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 476–484, Feb. 2005.
- [3] P. Maghouli, S. H. Hosseini, M. Oloomi Buygi, and M. Shahidehpour, "A scenario-based multi-objective model for multi-stage transmission expansion planning," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 470–478, Feb. 2011.
- [4] E. Sauma, C. Muñoz, J. Aguado, J. Contreras, and S. De La Torre, "Impact of high wind power penetration on transmission network expansion planning," *IET Gener. Transm. Distrib.*, vol. 6, no. 12, pp. 1281–1291, 2012.

- [5] M. Xie, J. Zhong, and F. F. Wu, "Multiyear transmission expansion planning using ordinal optimization," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1420–1428, Nov. 2007.
- [6] A. Braga and J. T. Saraiva, "A multiyear dynamic approach for transmission expansion planning and long-term marginal costs computation," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1631–1639, Aug. 2005.
- [7] R. Fang and D. J. Hill, "A new strategy for transmission expansion in competitive electricity markets," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 374–380, Feb. 2003.
- [8] A. K. Kazerooni and J. Mutale, "Transmission network planning under security and environmental constraints," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 1169–1178, May 2010.
- [9] K. Backhaus, *Multivariate Analysemethoden*. Berlin, Germany: Springer, 2000.
- [10] P. Kundur, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill, 1994.
- [11] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001, pp. xxxii, 546.
- [12] "Centro de Despacho Económico de Carga del Sistema Interconectado Central." (2016). [Online]. Available: <http://www.cdecsc.cl/>
- [13] Ministerio de Energía, "Campaña de medición del recurso Eólico y Solar," (2016). [Online]. Available: <http://ernc.dgf.uchile.cl/mediciones/>
- [14] "Centro de Energía, Universidad de Chile." (2016). [Online]. Available: <http://centroenergia.cl/desarrollos/publicaciones/papers-isi/>
- [15] R. Alvarez, "Ausbauplanung des chilenischen Übertragungsnetzes unter Berücksichtigung von Unsicherheiten," Ph.D. dissertation, Inst. Power Syst. Power Econ., RWTH Aachen Univ., Aachen, Germany, 2016.
- [16] CDEC-SIC, "Revisión 2014 Estudio de Transmisión Troncal Cuatrienio 2011–2014," 2014. [Online]. Available: <http://www.cdec-sic.cl>

Ricardo Alvarez was born in Salvador, Chile, on February 25, 1980. He received the degree in electrical engineering from the University of Chile, Santiago, Chile, in 2005 and the Ph.D. degree in electrical engineering from the RWTH Aachen University, Aachen, Germany, in 2016. He is currently a Researcher at the Energy Centre, Faculty of Mathematical and Physical Sciences, University of Chile. His main interests are power system planning, operation, and optimization.

Albert Moser was born in Linz, Germany, in 1965. He received the Diploma and the Ph.D. degrees in electrical engineering from the RWTH Aachen University, Aachen, Germany, in 1991 and 1995, respectively. Since March 2009, he has been the Head of the Institute of Power Systems and Power Economics. He was in charge of the business development at European Energy Exchange (EEX) in Leipzig, Germany, and the Head of the clearing and settlement department of EEXs clearing house. From 1997 to 2000, he was a Product Developer for TSO applications with Siemens AG in Nuremberg, Germany, and Minneapolis, MN, USA.

Claudia Rahmann was born in Santiago, Chile, on September 15, 1979. She received the degree in electrical engineering from the University of Chile, Santiago, Chile, in 2005 and the Ph.D. degree in electrical engineering from the RWTH Aachen University, Aachen, Germany, in 2010. She is currently a Professor in the Electrical Engineering Department, University of Chile, and an Associate Researcher in the Solar Energy Research Center of Chile. Her main interests are dynamic modeling of electrical power systems, power systems control and stability, energy storage systems, and control strategies for integration of wind and PV power plants into power systems.