Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 2654 – 2671

doi:10.55730/1300-0632.3961

Application of hierarchical clustering on electricity demand of electric vehicles for GEP problems

Seyedkazem AFGHAH[1], **Hatice TEKİNER MOĞULKOÇ**[2], **Bijan BİBAK**[1,3,*]
[1]Department of Industrial and Systems Engineering, Institute of Pure and Applied Sciences, Marmara University, İstanbul, Turkey
[2]Department of Industrial Engineering, Faculty of Engineering, Marmara University, İstanbul, Turkey
[3]Department of Industrial Engineering, Faculty of Engineering, Koç University, İstanbul, Turkey

Received: 28.04.2022 • **Accepted/Published Online:** 25.10.2022 • **Final Version:** 28.11.2022

Abstract: Increasing fossil fuel consumption and consequently the effects of greenhouse gases (GHGs) on the environment and economy are a major concern for all nations and governments. Electric vehicles (EVs) with plug-in capabilities have the potential to ease such problems. However, the extracted power from the grid for charging the EVs' batteries will significantly impact daily power demand. To satisfy the increasing demand and ensure generation capacity adequacy, the generation expansion planning (GEP) problem is solved to determine the investment decisions for electricity generation sources. Even though there are no centralized utilities for generation planning in most markets, there is still a need to realistically solve the GEP problems and find the optimal investment decisions to tailor the incentives used by most governments to guide the market. There is also a need for a tool to analyze the effect of different charging power levels, charging policies, and penetration levels. The main goal of this paper is to provide a tool to determine realistic optimal investment plans and evaluate different cases. It is also very important to consider the stochastic nature of the electricity demand in GEP problems. We propose a scenario-based stochastic programming model to incorporate the variability in the electricity demand due to EV charging through a set of scenarios generated by Monte Carlo Simulation. The methodology starts with applying a simulation method to generate the electricity demand of EVs by considering all the possible factors affecting EVs' demand. Each iteration of this simulation represents a possible demand profile as a result of penetrating the EVs into the market. Using all these demand profiles in GEP is preferable, but it is not computationally efficient. Computational tractability is achieved by using the clustering technique to reduce the size of such scenarios. We propose clustering methods to select a representative set from the data sets generated by the simulation and integrate EVs into GEP problems by using the selected set. The GEP models are defined to represent EVs' demand explicitly and then solved to imply the benefit of the suggested methods. The results show that GEP models with a representative set produce more realistic solutions than the GEP models including only average EVs demand. To select representative sets, different clustering techniques and distance measurements are used and compared with respect to their performances. Two different methods are defined to choose the best number of clusters: the silhouette coefficient method and the elbow method. For each method, five different distance measurement techniques are used. In each of these techniques, three approaches are evaluated for the representative point: Min, Max, and Average. A key contribution of this article is to explore and evaluate the quality of GEP models for each case according to how close the total cost obtained from the GEP model by using clustered load curves to the total cost obtained by using the full data sets generated by simulation.

Key words: Electric vehicles, hierarchical clustering, generation expansion planning, Monte Carlo simulation

*Correspondence: bbibak20@ku.edu.tr

1. Introduction

Oil consumption, climate change, and greenhouse gases are serious concerns of today's world [1]. Transportation is considered one of the most impactful reasons for these matters since the fossil fuel consumption of this sector is very high [2]. Migrating from using internal combustion cars (IC cars) to the electrified transportation fleet is considered an alternative solution for environmental concerns. Even though EVs have so many environmental advantages, there are a lot of serious concerns about their negative impact on the electricity generation network. Increasing the number of EVs in the market equals increasing power demand that consequently directly impacts power generation, reliability, and emission of the network [3, 4]. Therefore, it is crucial to make a precise and efficient plan for the future development of the EVs fleet.

In most electricity markets, there are no central planning utilities to make generation investment decisions. However, in most of them, the government uses incentives, taxes, or subsidies to guide the market. For proper guidance strategies, an optimum investment plan needs to be found. Generation expansion planning (GEP) problem is a well-known problem for expanding the current electricity supply network according to any new situation, such as new demand or new resources, which makes it a good decision tool for changes due to the increasing number of EVs. To obtain realistic solutions for the GEP, the demand changes due to the EVs should be represented in GEPs explicitly. Another important issue towards a more electrified transportation system is the infrastructure to charge the EVs. There is a need for efficiently analyzing the different cases in terms of the charging power levels, charging policies, and penetration levels. This study aims to provide a tool to shape the future electricity generation network and charging infrastructure.

We propose a methodology to generate load curves by considering all the impact factors via simulation and selecting a representative set via clustering methods to integrate them into the GEP problem to maintain computational efficiency. We also provide an analysis to demonstrate the usage of the proposed approach to analyze the different cases.

Clustering is one of the fundamental concepts in data mining, and it is considered unsupervised learning [5]. The final aim of clustering is descriptive. Partitioning clustering is a clustering method that is used to classify different observation points, in a data set, into different groups based on similarity. In most cases, the number of clusters should be specified. The partitioning clustering uses an iterative relocation technique that tries to improve the partitioning by changing object clusters from one group to another [6]. The hierarchical clustering method creates the clusters by consecutively partitioning the data in either a top-down or bottom-up method [7]. Density-based clustering method is constructed by identifying groups of points that form high-density regions in the data space [8]. In this study, we will use the hierarchical clustering method.

There are many studies concerning electric vehicles. Bibak and Tekiner [9] presented a comprehensive analysis and precise categorization of the recent studies around electric vehicles and vehicle to grid system (V2G) based on their proposed methodology. Hannan et al. [10] provided a review study on hybrid electric vehicles and their challenges. They pointed out that the increasing amount of fossil fuel used in cars is a real concern because of its negative impacts on the environment. They introduced alternative resources such as fuel cells (FC) and supercapacitors. Thounthong et al. [11] explored different techniques and aspects of hybrid electrical vehicles (HEVs) like propulsion systems and energy management systems (EMS). Hadley et al. [12] evaluated the effects of HEVs on a power generation network and applied a simulation method to analyze the change in peak demand, increasing the prices and reducing the margins after penetration of the HEVs. Momtazpour et al. [13] evaluated the importance of EVs and their role in the future of power networks. They utilized a developed clustering technique to model the electric network for urban areas to figure out the best placement of charging

stations. Coordinated clustering is used to get the best placement of charging stations.

Shukla et al. [14] evaluated the increasing penetration level of EVs and their need for charging stations. They applied K-mean and fuzzy C-mean clustering to find the best locations of the fast-charging stations (FCSs). The author in [15] utilized the generation expansion planning (GEP) problem to find the optimal plan for electricity network, technologies, location, size, and time of building new power plants by minimizing the total cost in the long term. Tekiner et al. [16] investigated the GEP problem using two-stage programming to incorporate reliability analysis into the dispatching problem and expansion planning. They implemented a multiobjective optimization to evaluate the trade-offs between expense and ecological effects for wind turbines. Ramirez et al. [17] proposed a planning model to optimize operation and investment costs considering EVs' demand flexibility. Hu et al. [18] proposed a new robust optimization-based GEP model by considering the virtual inertia that is supported by wind farms. This study aims to minimize the expansion cost by considering all the uncertainties and maximize the probability of increasing the level of renewable energy sources. Abdin et al. [19] introduced a multistage adaptive robust GEP model specifically for the short-term unit, which simultaneously covers multiperiod and multiregional planning. This study analyzed the stochasticity of electricity demand and generated power by renewable energy sources by presenting bounded intervals.

Yu [20] proposed a model for a national energy system to predict the need for new power generation plants by assuming that the penetration of Plug-in Hybrid Electric Vehicles (PHEVs) will increase in coming years. For daily charging strategies, four different types are investigated: uniform, off-peak, home-based, and vehicle-to-grid (V2G). The impact of these charging profile on GEP is studied by a scenario-based analysis. Foley et al. [21] utilized a common analysis tool called WASP-IV which is a traditional generation expansion planning tool for analyzing the charging of the EVs in the all island grid (AIG) by long-term GEP. They investigated three different scenarios as base-case, peak, and off-peak charging. The simulation of the impacts of EVs is done up to the year 2025 in AIGs [22]. Tekiner [23] solved the GEP problem where demand due to EVs is presented explicitly by using an optimized-based reduction procedure. Guo et al. [24] considered revenue adequacy constraints in a two-stage GEP model. The results indicated that investing in renewable generation capacity causes a small increase in the total costs.

Castro et al. [25] analyzed a short-term GEP with capacity for multiterminal transmission systems. The presented formulation is based on mixed-integer linear programming (MILP), in which transmission losses are considered by piecewise linearization. Singh et al. [26] discussed some optimization strategies for distributed generation, EVs, and distributed generations by EVs with different load models. Borozan et al. [27] introduced investment and operation models of EVs, specifically in V2G and V2B systems, for the large-scale and long-term network expansion planning problem. In this study, the presented multi-stage stochastic planning can find the optimal investment by minimizing the cost and risk of investment.

The most common way in cases of integrating EVs demand for GEP problems is to take the average of the possible load curves. However, there are some significant disadvantages to this. Firstly, by taking the average, many actual and valuable data are being ignored. Most of the extreme points, which are usually the case in peak hours and are the most important points, are being neglected by taking the average. To guarantee the reliability of the generation system, all the possible load curves should be explicitly represented in GEP models. However, using all the possible load curves in GEP is not computationally efficient. In this paper, we suggest using hierarchical clustering instead of taking the average to reduce the problem size while maintaining a more realistic presentation of the demand. The optimal clustering method is first determined in terms of the distance

measurement method and centroid point. The selected clustering method is used to create the representative subsets of the possible load profiles for the GEP model. The proposed cluster-based methodology creates more realistic GEP models, resulting in more realistic solutions that allow it to be used to analyze many different cases. The main contribution of this paper is to generate realistic load curves due to the penetration of EVs and efficiently represent them in the GEP problem.

The first part of the analysis consists of determining the optimal clustering method. Then, we perform an analysis to demonstrate the usage of our proposed method to evaluate the different power charging levels, charging policies, and penetration levels. In this analysis, we determine a small set of EV demand profiles to represent all profiles correctly and then compare the results of using full data, clustered data, and the average, and show that using the clustering gives a relatively close answer to full data.

This paper is organized as follows: the suggested methodology and GEP models are discussed in Section 2. Section 3 covers the simulation procedure and then shows the influential factor for defining different scenarios and explains how demand profiles are generated. In Section 4, GEP problems are explained, and the mathematical model is given. Different cases are defined in Section 5. In Section 6, two types of hierarchical clustering utilized in this paper are explained. The outputs of simulation and numerical analysis are given in Section 7 and Section 8. Finally, conclusions are drawn in Section 9.

2. Methodology

The main idea of the proposed methodology is to integrate EVs into GEP models starting with generating load curves by using the modified Monte Carlo simulation presented in [28]. Next, the GEP model is proposed, which explicitly shows the EVs' demand. In the first part of the analysis, we define different clustering methods with respect to the distance measurement methods and the centroid points used. Each of these methods is then used to select a representative set for the defined base case. The GEP models using the clustered data for each method are solved. The results are compared with the result obtained by the GEP model using the full data (including all the load curves generated via Monte Carlo Simulation). The clustering method with the smaller deviation from the result obtained by the GEP model with full data is selected as the optimum clustering method. In the second part of the analysis, different cases are defined in terms of the penetration level of EVs, the preference distribution of the charging places and policies for charging times, and GEP models using the representative sets generated by the optimal clustering method are solved. The solutions of the GEP model with full data, representative load curves (clustered data), and the average of load curves are compared to justify that the proposed model produces realistic solutions and can be used to analyze different charging power levels distributions and policies. The flowchart of selecting the optimal clustering method is given in Figure 1a. The flowchart for the GEP problems for analyzing different cases is shown in Figure 1b.

3. Simulation procedure

To generate EVs' demand profiles, we modify the Monte Carlo simulation presented in [28]. This simulation consists of different users' profiles with respect to daily driven distances, arrival and departure time of charging points, battery sizes, and charging infrastructure types. The Monte Carlo simulation, coded in MATLAB, generates many random drivers for a given penetration level. For each driver, based on the driver's charging and driving characteristics, the simulation randomly determines the time of charging and the amount of electricity extracted from the power grid. At the end of each iteration, electricity withdrawn from the grid by each hour is obtained by summing the electricity demand of each driver for that hour. Each iteration corresponds to a

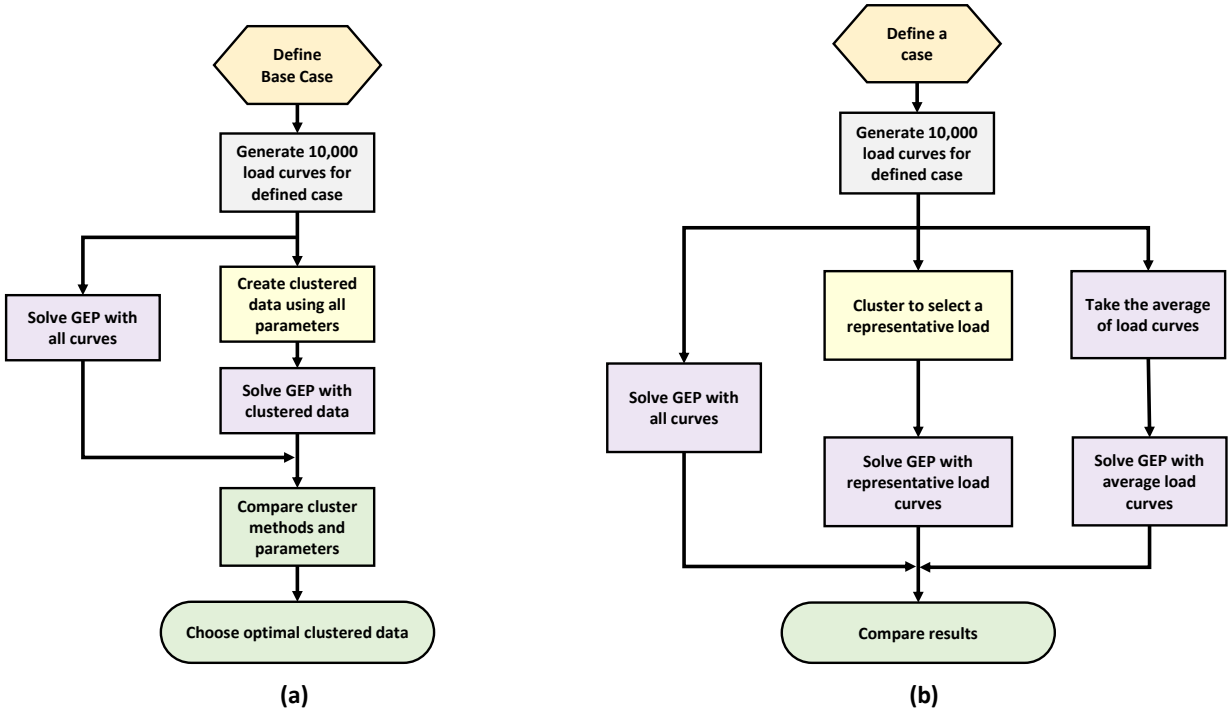

Figure 1. Flowchart for selecting optimal clustering and comparing GEP models.

random possibility for the EVs' demand. They are defined as scenarios in the Generation Expansion Planning (GEP) problem.

There are many factors affecting the EV's demand. For a detailed explanation of the factors such as driving characteristics, charging characteristics and so on, readers are referred to [28]. The main modification done is to make it possible to consider different charging policies at home and public stations.

Three charging locations are defined as charging centers for electric vehicles: home, work, and public stations to represent the different charging power levels. We consider different distributions for the preferences of charging locations. For the charging policies at homes, two different cases are considered: controlled (C_1) and uncontrolled (C_0). The controlled case is when people are not allowed to charge at peak hours, which are from 5 pm to 10 pm. The uncontrolled means they can charge at home at any time upon arrival.

For charging in public stations, different cases are defined: T_1, T_2, and T_3. Considering that people usually use public stations for charging after leaving their workplaces, these three cases are defined to see the effect of different patterns of public station charging. T_1 is defined as when people charge at public stations from 6 pm to 10 pm. The idea is that they charge the EVs after their work hours and not go home very late. T_2 is defined as when they charge the EVs from 10 pm to 2 am. This will show us the effect of late-night charging at public stations. T_3 is defined as no limit, meaning that public stations can be used any time of the day.

4. Generation expansion planning model

Generation expansion planning (GEP) problem is a mixed integer multilayer problem that targets identifying the best scheduling for investment, considering the locations and technologies. Here, we integrate the EVs

demand into GEP problems through scenarios. Each scenario is defined as a possible demand profile of the power system. The demand profile is obtained by adding the new required demand for charging EVs to the selected grid demand. The common approach for determining the power grid demand is to choose some days to represent the year instead of considering all days of the year. To illustrate the electricity demand of the grid, two days, one from the summer and one from the winter, with the highest peak load are selected. The required data set is collected from Turkish Electricity Transmission Corporation (TEİAŞ). The probability of occurring each of them is equal (p = 50%).

In the GEP model, the total cost is obtained by summing over scenarios. The adjustment factor is defined, corresponding to the number of days in the year represented by each scenario. We consider expanding the system with new wind turbines. The associated mathematical model is presented as follows:

$$\min \sum_{t,h,s,j} p^j x^j_{ths} w_s + \sum_{t,h,s,j} q^i y^i_{ths} w_s + \sum_{t,h} \text{penalty} * u_{ths} w_s + \sum_i (C^i z^i)$$

$$s.t. \sum_j x^j_{ths} + \sum_i y^i_{ths} + u_{ths} = d_{ths} \qquad \forall t,h,s$$

$$x^j_{ths} - K^j \leq 0 \qquad \forall t,h,s,j$$

$$y^i_{ths} - 0.3 L^i z^i \leq 0 \qquad \forall t,h,s,i$$

$$x^j_{ths}, y^i_{ths}, u_{ths} \geq 0 \qquad \forall t,h,j$$

$$z^i \in \{0,1\} \qquad \forall i$$

In the first term, the model's objective function is to minimize the cost of power generation with the existing system. The second term represents the generation cost of additional turbines, which is designed to minimize the penalty due to unmet demand with current and additional power capacity. Finally, it considers the capital required for investing in new turbines. The variable x represents the amount of generation using the current system, y is the amount of generation using new turbines, u is for unmet demand, and z is for whether a turbine is added or not, respectively. The indexes t, h, s indicate the day, the hour, and the scenario, where indices j and i show the current system and new turbines. Thus, p^j is the production cost of each MWh by the current system, while q^i is the unit production cost of new wind turbines. The capital investment for a new turbine is given by C^i. Note that w_s is an adjusting factor for scenario s to find the annual costs. It shows that all the values and numbers are for 1 year.

To write the constraint, we defined the following parameters: K^j as the available capacity of unit j and L_i as the capacity of the new turbine. Moreover, d_{ths} represents the demand at day t, hour h, and scenario s. The first constraint ensures that all demand is satisfied. The second and third constraints ensure the production is less than the available installed capacity. For the wind turbines, 30% of installed capacity is defined as available capacity. Finally, we have nonnegativity and binary constraints.

4.1. GEP model with full data

Ten thousand (10,000) random EVs' demand profiles are generated for each case considered via Monte Carlo simulation. GEP model including all of them is defined as a GEP model with full data and it will be our reference to compare the results. For GEP with full data, the adjustment factor of each scenario is equal to the

product of the probability of the scenarios (*1/10,000*) by the number of days represented with selected days from a year.

4.2. GEP model with clustered load curves

The only change in the GEP model proposed before is that we have a smaller number of scenarios and also, the adjustment factor is different. As mentioned, the adjustment factor in GEP with clustered data is computed as the product of the probability of that cluster and the number of days represented by the selected days. The probability of clusters is obtained by the number of demand's profiles in the cluster divided by 10,000.

4.3. GEP model with average of load curves

For GEP with average EVs' demand, an average of 10,000 is obtained and added to the current grid demand. Therefore, the adjustment factor equals the number of days represented by the selected days from the year.

5. Data used and definition of cases

In the process of the generation of load curves, several characteristics and parameters are used. When people want to charge their electric vehicles, there are three charging stations that they can choose from, i.e. home, work, and public. The charger units used in each of them are uniformly distributed with 3.7 kW and 7.4 kW at home, 7.4 kW and 11 kW at work, and charger units with 11 kW and 22 kW power in the public stations. The charging location preference distributions that are considered in this paper can be observed in Table 1.

The distribution of electricity consumption of charging vehicles is also dependent on whether people are at home or work given a specific hour. The distribution for the percentage of people at home and at work presented in Figure 2 has been used.

Table 1. Charging place preference distribution.

Preference	Home	Work	Public
1	15%	15%	70%
2	30%	40%	30%
3	0%	0%	100%
4	0%	100%	0%
5	100%	0%	0%

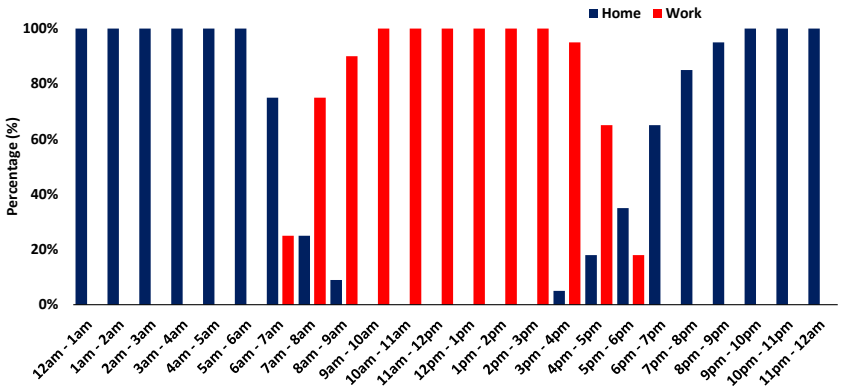

Figure 2. Distribution of people at home and work throughout a day.

Since the actual EV penetration level in İstanbul is uncertain, two different penetration levels are considered for this paper: 10% and 30%. The GEP problems are usually important for the decision-making of long-term problems. Even though the current penetration level in İstanbul is less than 10%, it is assumed that the number of electric vehicles would grow rapidly sooner or later. Thus, higher penetration level can help us in future decision-making.

All these different scenarios and assumptions led to 156 cases; for each case, 10,000 EVs' demand curves are generated by using the Monte Carlo simulation. This caused many data sets with similar characteristics. To resolve this issue, 26 most impactful and diverse cases are selected in the paper which are presented in Table 2.

Table 2. List of the datasets.

Case	Penetration level	Peak hour restriction	Public time	Home %	Work %	Public %
1	10%	No	All day	0%	0%	100%
2	10%	No	6 pm to 10 pm	0%	0%	100%
3	10%	No	10 pm to 2 am	0%	0%	100%
4	10%	No	All day	100%	0%	0%
5	10%	Yes	All day	100%	0%	0%
6	30%	No	All day	0%	0%	100%
7	30%	No	6 pm to 10 pm	0%	0%	100%
8	30%	No	10 pm to 2 am	0%	0%	100%
9	10%	No	All day	15%	15%	70%
10	10%	No	6 pm to 10 pm	15%	15%	70%
11	10%	No	10 pm to 2 am	15%	15%	70%
12	10%	No	All day	80%	10%	10%
13	10%	Yes	All day	80%	10%	10%
14	10%	No	6 pm to 10 pm	0%	100%	0%
15	30%	No	6 pm to 10 pm	0%	100%	0%
16	30%	No	All day	100%	0%	0%
17	30%	Yes	All day	100%	0%	0%
18	30%	No	All day	15%	15%	70%
19	30%	No	6 pm to 10 pm	15%	15%	70%
20	30%	No	10 pm to 2 am	15%	15%	70%
21	10%	No	All day	30%	40%	30%
22	10%	No	6 pm to 10 pm	30%	40%	30%
23	10%	No	10 pm to 2 am	30%	40%	30%
24	10%	Yes	All day	30%	40%	30%
25	10%	Yes	6 pm to 10 pm	30%	40%	30%
26	10%	Yes	10 pm to 2 am	30%	40%	30%

These cases are solved by the GAMS with CPLEX solver on three sets of data; full data set, clustered data set, and averaged data set to compare the efficiency of both methods with respect to the full data set. GEP is mixed-integer programming (MIP) that finds the number of additional wind turbines while minimizing the cost.

The capacity of the existing power production is assumed to be 40,000 MWh. As EVs are cleaner than regular cars and the increasing number of them is a result of the need for a cleaner approach, a green energy resource like the wind is supposed as a new resource in this paper.

Due to the stochastic nature of wind power, wind turbine does not produce with its full capacity as it has a 24% capacity, so this was considered in our formulation. During the day, there are hours when the system's capacity cannot satisfy the full demand; to address this, we penalize the unmet demand, which could result in outsourcing or power outage. Other parameters such as power generation cost and investment cost are presented in Tables 3 and 4. The data is collected from Turkish Electricity Transmission Corporation (TEİAŞ). In the final solution of the model, the number of new turbines will be determined.

Table 3. Variable cost of different energy sources.

Energy source	Installed power (MW)	Availability factor (AF)	Energy loss	Available energy after applying AF and loss (MW)	Variable cost ($/MWh)
Hydro	3270	50%	8%	1504	0.002
Geothermal	172	75%	8%	119	0.011
Solar	3	50%	8%	1	0.003
Biomass	73	75%	8%	50	5.272
Waste	22	85%	8%	17	0.001
Natural gas	4299	85%	8%	3362	3.602
Oil	111	70%	8%	71	3.602
Coal	3187	85%	8%	2,492	4.474
Stream	1243	50	8%	572	0.001

Table 4. Variable cost of wind turbine.

Energy Source	Installation Cost(M$)	Lifetime (Years)	Installation Cost($/Year)	Availability Factor (AF)	Variable Cost ($/MWh)	Capacity (MW)
Wind	3.5	20	175,000	24%	0.002	2

6. Hierarchical clustering

Hierarchical clustering is an approach to partitioning clustering to categorize groups in any given data set. The main advantage of applying hierarchical clustering in this paper is that it does not need to select or choose the number of clusters at the beginning. The final result of hierarchical clustering is a tree-based graph, which is a good representation of the objects, also known as dendrogram. Data points can be subdivided into groups by cutting the dendrogram at any desired similarity level as shown in Figure 3.

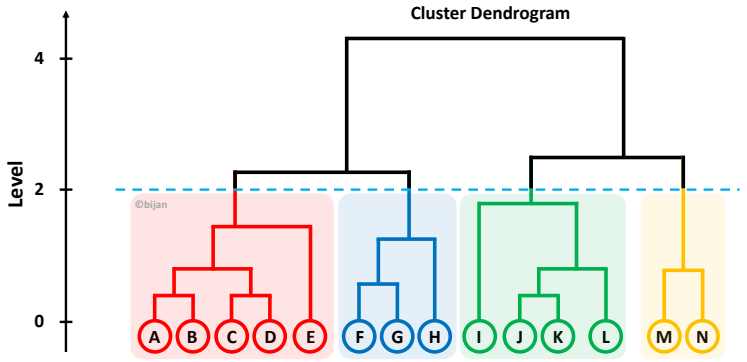

Figure 3. Dendrogram diagram.

One of the most important necessities of having an efficient hierarchical clustering is to figure out the best number of clusters. As you can see in the picture above, this clustering technique goes forward until it reaches one cluster. We will have a representative point for each cluster and use that instead of all the clusters.

In this paper, three different points are considered as the representative point: min, max and average. When clustering reaches one cluster and if the average method is used for the representative point, it will be the same as taking the average for all the data. Thus, it is very important to choose the best number of clusters.

In this paper, two different methods are used for each of five distance measurement methods to figure out the best number of clusters: the silhouette and elbow methods.

6.1. Elbow method

In clustering, the elbow method is a heuristic to figure out the best number of clusters in a data set. The method consists of creating the variation graph with respect to the number of clusters and picking the elbow of the graph as shown in Figure 4a.

6.2. Silhouette method

In this method, as shown in Figure 4b, which is popular for both cohesion and separation, a silhouette coefficient is defined for an individual point in three steps as follows:

1. For the i^{th} point, calculate the average distance to all other points in the same cluster, and call it a_i.
2. For the i^{th} point and any cluster not containing this point, calculate the average distance to all points in the given cluster and call it b_i.
3. For the i^{th} point, the silhouette coefficient is $s_i = (b_i - a_i)/max(a_i, b_i)$

The value of this coefficient changes from -1 to 1. A negative value means the average distance of the point in the same cluster is more than its average distance to points in other clusters, which is not desirable. We want this coefficient to be positive and as close to 1 as possible since, in the case of 1, it means that a_i was equal to zero.

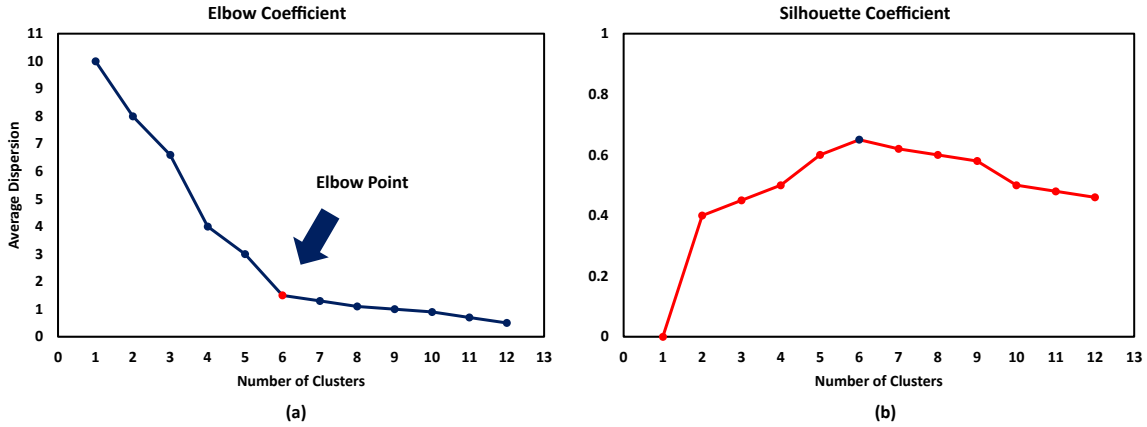

Figure 4. Elbow and silhouette methods for selection optimal clusters.

7. Comparison of hierarchical clustering methods with different parameters

Choosing the number of clusters is included in the code. The results show that the optimal number of clusters is six. By using five different distance measurement methods (Euclidean, city block, Sorensen, Canberra, and Minkowski) and three different representative points (average, min, and max), we investigated a list of different methods on a base case (penetration level = 10%, peak hour restriction = no, public time = all day, home = 0%, work = 0%, public = 100%) to find the optimal clustering method. This method is later used on the other 25 cases defined in the GEP part as well. By comparing the percentage error of this list, we figure the best method for our data set. The results of these methods can be found in Table 5.

Table 5. Clustering method comparison.

Method number	Distance method	Centroid criteria	Parameter	Optimal number of clusters	Total cost (cluster)	Total cost (full)	Error
1	Euclidean	Average	N/A	6	$356,011,974	$356,061,539	0.01392%
2	Euclidean	Min	N/A	6	$355,894,828	$356,061,539	0.04682%
3	Euclidean	Max	N/A	6	$355,875,765	$356,061,539	0.05217%
4	City	Average	N/A	6	$355,410,661	$356,061,539	0.18280%
5	City	Min	N/A	6	$354,876,285	$356,061,539	0.33288%
6	City	Max	N/A	6	$354,306,418	$356,061,539	0.49293%
7	Soren	Average	N/A	6	$354,734,371	$356,061,539	0.37274%
8	Soren	Min	N/A	6	$354,590,094	$356,061,539	0.41326%
9	Soren	Max	N/A	6	$353,740,043	$356,061,539	0.65199%
10	Canberra	Average	N/A	6	$354,190,199	$356,061,539	0.52557%
11	Canberra	Min	N/A	5	$353,376,945	$356,061,539	0.75397%
12	Canberra	Max	N/A	6	$352,940,333	$356,061,539	0.87659%
13	Minkowski	Average	Power = 4	6	$355,003,301	$356,061,539	0.29721%
14	Minkowski	Min	Power = 4	6	$354,131,082	$356,061,539	0.54217%
15	Minkowski	Max	Power = 4	6	$353,485,269	$356,061,539	0.72355%
16	Minkowski	Average	Power = 5	6	$355,812,163	$356,061,539	0.07004%
17	Minkowski	Min	Power = 5	6	$355,738,310	$356,061,539	0.09078%
18	Minkowski	Max	Power = 5	6	$355,413,883	$356,061,539	0.18189%
19	Minkowski	Average	Power = 6	5	$355,533,802	$356,061,539	0.14822%
20	Minkowski	Min	Power = 6	6	$354,974,035	$356,061,539	0.30543%
21	Minkowski	Max	Power = 6	6	$354,089,460	$356,061,539	0.55386%

It is understood from this table that the best similarity method for our data set is Euclidean, and the best representative point is the average. To get a better understating of the results of the table, the two charts in Figures 5 and 6 are provided.

Figure 5 indicates the percentage error of the implemented methods. As shown, the first three bars have the lowest error rates, which are for Euclidean measurement. For each group of three bars, the first one is for the Average of the specified similarity method, the second one is for min, and the third one is for max. For example, the first bar number represents average Euclidean, the second one represents min Euclidean, and the third one represents max Euclidean. The fourth bar is for the average city block, the fifth one is for the min city block, and the sixth bar is for the max city block. It goes on for all methods. It is shown that in all the cases, the error for average is the least, and then min and max are after that. Other than percentage error,

another chart for the total cost is provided in Figure 6.

Figure 7 indicates the results of clustering on one of the base cases under average Euclidean method. As shown, the first cluster represents 2150 demand profiles with average demand 101 MWh (lowest electricity demand), while the third cluster with 1482 demand profiles and average demand about 140 MWh has the highest rate.

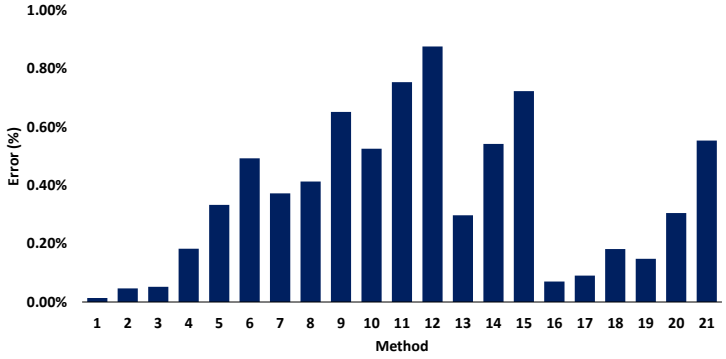

Figure 5. Error rates of implemented methods.

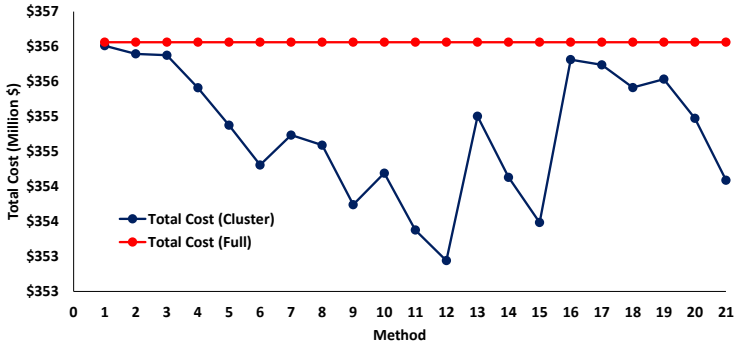

Figure 6. Cost comparison of different clustering methods with the actual cost.

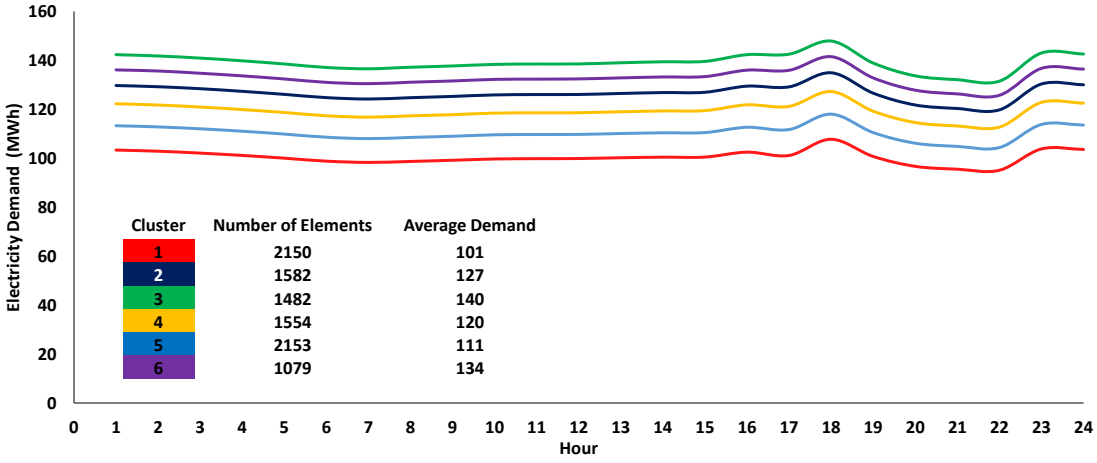

Figure 7. Results of clustering on one of the base cases.

8. GEP problems for different cases

Using GAMS, the mathematical problem has been solved for all the cases presented in Table 2. The clustering method selected in the previous section is used to obtain the representative load curves for these cases. In Table 6, the final results are shown. There are different parameters to be considered, the final cost of each method, the number of suggested turbines, and the time of solving the problems. As shown in the last two columns, the difference between the suggested number of turbines between using the full data and the clustering method is very low. In most cases, they are identical, but the difference between using the average and using full data is much higher. Comparing the costs of each method tells us the same thing; using the clustering method is not giving a huge error compared to using full data, rather than using the average which has a much higher error.

Another parameter that explains the difference is the unmet demand. By using the full data, the amount of unmet demand is the percentage of total demand that is not satisfied, meaning that it is not generated in our network. In other words, unmet demand is the difference between total demand and maximum generation of our network in percentage. For example, as seen in the results table, for the first case, unmet demand is 0.00082% for the full data. It is the same for clustering data, but for the average, the number is 0.01866%. This shows that in the average method the model is ignoring a lot of demand by not satisfying them and will not suggest investment in new resources for them. That results in less total cost compared to the full data case. This huge difference with the full data, which we know is the actual and correct answer, is causing the final error to be much higher. As shown in Figure 8, in some cases, the difference between unmet demand for cluster and average is not so much, whereas in other cases, there is a huge difference. By looking out for total cost comparison cases, we can see that there is a relationship between the difference in unmet demand and the percentage error in total cost, which means that the cases with a huge difference between their unmet demands will cause higher total cost error in their average method compared to the clustered method.

By comparing the calculation time of each problem, it is shown that the time of using average and clustering methods are very close to each other, whereas they are less than five percent of using full data in most cases. In comparing these methods, reducing run time is one of the most important benefits. Figure 9a indicates a visual comparison of calculation time.

As shown in this chart, the run time for the clustering method and average are so close that they can hardly be distinguished from each other, while the full method line is much higher. To have a better understanding, a logarithmic graph is given below for the same purpose in Figure 9b. More importantly, the comparison between the total costs is our focus since the goal of the model is to minimize total cost. In Table 7, the total cost for each method and also the percentage error for clustering and average technique are given.

Figure 10 is given to have a visual sense of the differences. As shown, it is evident that the error from using the averaging method is so drastic, whereas the error from the clustering method is relatively negligible compared to using the full data. Since this is big data, we have seen in Figure 9 that the run time for the full data is rather long. Thus, for real-time analysis, using full data becomes rather impractical. This issue caused the necessity for a more practical method to solve the GEP problems on a large scale.

Table 6. Results obtained for defined cases.

Data	Cost_full	Unmet	Penalty	Turbine	Time (s)	Cost_clust	Unmet	Penalty	Turbine	Time (s)	Cost_ave	Unmet	Penalty	Turbine	Time (s)
1	$356,061,539	0.00082%	$0.08	23	262	$356,011,974	0.00082%	$0.08	23	14.7	$319,476,645	0.01866%	$2	19	14.5
2	$298,939,678	0.00172%	$0.17	15	163	$295,987,439	0.00340%	$0.34	14	14.8	$264,559,189	0.00860%	$1	12	14.6
3	$209,802,157	0.00003%	$0.00	5	400	$209,718,442	0.00003%	$0.00	5	16.03	$208,031,329	0.00219%	$0	4	15.64
4	$217,450,381	0.00000%	$0.00	6	363	$216,762,847	0.00126%	$0.13	5	15.55	$209,632,048	0.00126%	$0	5	16.65
5	$191,881,404	0.00091%	$0.09	2	331	$193,799,731	0.00091%	$0.09	2	16.22	$186,146,980	0.00091%	$0	2	15.75
6	$916,461,670	0.00191%	$0.19	92	157	$911,432,334	0.00191%	$0.19	92	14.93	$799,558,246	0.10576%	$11	78	15.08
7	$850,851,917	0.00413%	$0.41	82	148	$855,617,605	0.00066%	$0.07	86	15.29	$737,281,386	0.03122%	$3	71	14.95
8	$308,817,397	0.00060%	$0.06	16	327	$307,563,464	0.00060%	$0.06	16	15.27	$291,505,591	0.00576%	$1	14	15.05
9	$275,477,733	0.00049%	$0.05	13	347	$272,251,564	0.00049%	$0.05	13	15.47	$248,877,499	0.01834%	$2	10	15.03
10	$212,271,809	0.00041%	$0.04	5	328	$212,185,941	0.00041%	$0.04	5	15.88	$186,287,673	0.00689%	$1	2	15.5
11	$210,643,885	0.00016%	$0.02	5	292	$210,484,040	0.00016%	$0.02	5	16	$209,638,795	0.00016%	$0	5	15.28
12	$221,635,937	0.00066%	$0.07	6	281	$221,618,905	0.00066%	$0.07	6	15.26	$217,464,269	0.00066%	$0	6	15.13
13	$196,437,383	0.00039%	$0.04	3	255	$194,778,303	0.00039%	$0.04	3	15.39	$193,962,750	0.00039%	$0	3	15.05
14	$307,528,885	0.00059%	$0.06	17	270	$309,367,401	0.00059%	$0.06	17	17.41	$291,628,100	0.00530%	$1	15	15.58
15	$660,403,666	0.00262%	$0.26	59	299	$660,896,564	0.00262%	$0.26	59	15.47	$603,910,553	0.01545%	$2	54	15.76
16	$441,561,116	0.00052%	$0.05	33	300	$441,330,787	0.00052%	$0.05	33	15.68	$398,926,270	0.01746%	$2	28	15.64
17	$481,487,763	0.00188%	$0.19	37	234	$483,306,871	0.00189%	$0.19	37	15.88	$429,978,942	0.01317%	$1	32	15.36
18	$703,407,142	0.00077%	$0.08	66	158	$699,061,761	0.00077%	$0.08	66	15.57	$611,726,796	0.06449%	$6	55	15.37
19	$588,863,746	0.00256%	$0.26	50	137	$587,674,976	0.00153%	$0.15	51	15.85	$517,110,255	0.01791%	$2	43	15.12
20	$313,852,229	0.00017%	$0.02	17	347	$313,661,547	0.00017%	$0.02	17	16.07	$297,064,583	0.00489%	$0	15	15.28
21	$242,984,099	0.00031%	$0.03	9	237	$242,868,189	0.00031%	$0.03	9	15.74	$233,178,199	0.00234%	$0	8	15.28
22	$243,008,187	0.00032%	$0.03	9	273	$242,205,308	0.00032%	$0.03	9	15.69	$233,178,043	0.00232%	$0	8	15.49
23	$217,602,934	0.00002%	$0.00	6	400	$217,399,734	0.00002%	$0.00	6	15.38	$209,652,440	0.00254%	$0	5	15.7
24	$219,494,245	0.00032%	$0.03	6	257	$219,582,988	0.00032%	$0.03	6	19.4	$211,272,549	0.00400%	$0	5	19.37
25	$219,586,805	0.00033%	$0.03	6	383	$220,817,979	0.00033%	$0.03	6	16.75	$211,462,016	0.00404%	$0	5	15.17
26	$218,005,609	0.00009%	$0.01	6	312	$217,688,956	0.00009%	$0.01	6	15.9	$209,644,613	0.00189%	$0	5	15.45

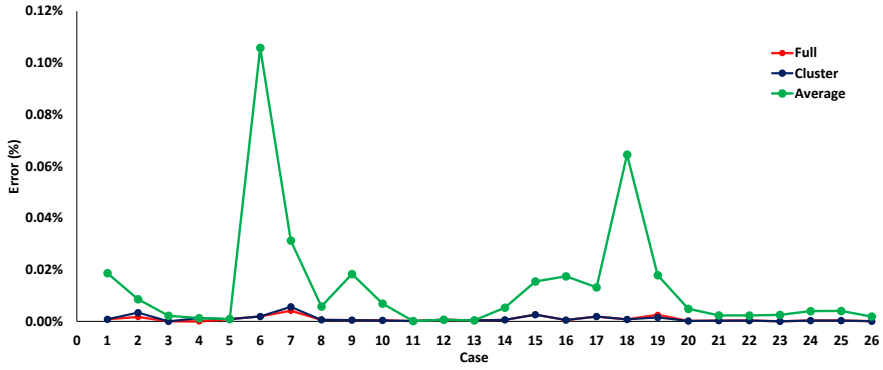

Figure 8. Unmet demand percent comparison.

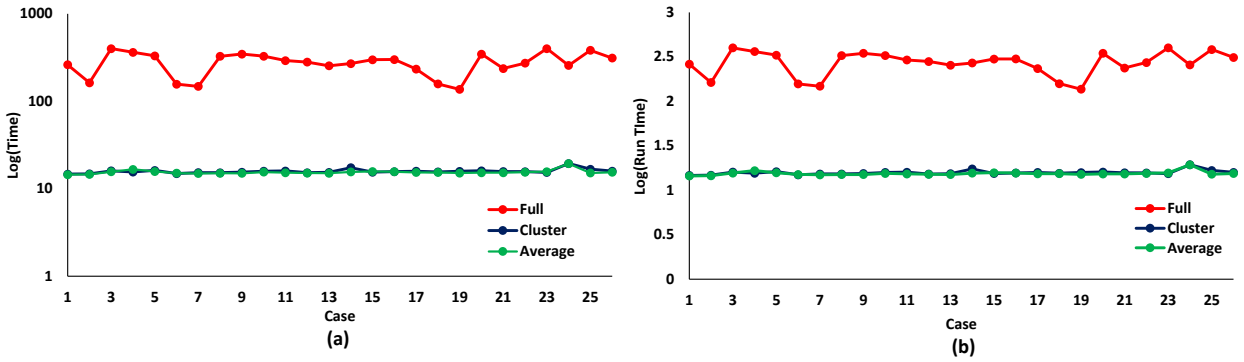

Figure 9. Run time versus log run time.

The presented results in this section confirm our hypothesis that using hierarchical clustering, while not losing much information and accuracy, is extremely advantageous regarding the run time of the GEP problem-solving. It has been figured that when most of the charging or all of them happen at one charging place (like home or public stations), new peaks occur in our demand curves, meaning it requires more investment. In these cases, the error of taking the average is much higher than using the suggested clustering method. It is the same for all other limitations. For example, by applying the charging time limitations for public stations, which are defined as T_1, T_2, and T_3 in this study, we can see that for T_1 and T_2 cases (with limitations) the error difference between the average and clustering method is relatively higher than T_3 case (with no limitation).

9. Conclusion

The electrified transportation sector would have a huge effect on electricity usage and the supply network. This magnifies the importance of deep studies on this matter. Because of the inevitable increasing number of electric vehicles (EVs) in coming years, it is necessary to consider their new load on expansion planning of the current network. This study integrates EVs demand into the GEP problem via clustering. For this matter, the new load is added to two different current day loads. One of them as the representative of summer and the other one for winter. The Monte Carlo simulation method is used to generate new load curves with defined assumptions.

Firstly, three different places for charging the EVs are assumed: home, work, and public stations. For each of them, we assign different characteristics, which we think may be highly possible when the number of EVs is increased. For home charging, we assume controlled and uncontrolled charging, which means whether

Table 7. Cost and error comparison between clustering and average.

Cost_full	Cost_clustering	Cost_average	Clustering	Average
$356,061,539	$356,011,974	$319,476,645	0.01%	10.27%
$298,939,678	$295,987,439	$264,559,189	0.99%	11.50%
$209,802,157	$209,718,442	$208,031,329	0.04%	0.84%
$217,450,381	$216,762,847	$209,632,048	0.32%	3.60%
$191,881,404	$193,799,731	$186,146,980	1.00%	2.99%
$916,461,670	$911,432,334	$799,558,246	0.55%	12.76%
$850,851,917	$855,617,605	$737,281,386	0.56%	13.35%
$308,817,397	$307,563,464	$291,505,591	0.41%	5.61%
$275,477,733	$272,251,564	$248,877,499	1.17%	9.66%
$212,271,809	$212,185,941	$186,287,673	0.04%	12.24%
$210,643,885	$210,484,040	$209,638,795	0.08%	0.48%
$221,635,937	$221,618,905	$217,464,269	0.01%	1.88%
$196,437,383	$194,778,303	$193,962,750	0.84%	1.26%
$307,528,885	$309,367,401	$291,628,100	0.60%	5.17%
$660,403,666	$660,896,564	$603,910,553	0.07%	8.55%
$441,561,116	$441,330,787	$398,926,270	0.05%	9.66%
$481,487,763	$483,306,871	$429,978,942	0.38%	10.70%
$703,407,142	$699,061,761	$611,726,796	0.62%	13.03%
$588,863,746	$587,674,976	$517,110,255	0.20%	12.19%
$313,852,229	$313,661,547	$297,064,583	0.06%	5.35%
$242,984,099	$242,868,189	$233,178,199	0.05%	4.04%
$243,008,187	$242,205,308	$233,178,043	0.33%	4.05%
$217,602,934	$217,399,734	$209,652,440	0.09%	3.65%
$219,494,245	$219,582,988	$211,272,549	0.04%	3.75%
$219,586,805	$220,817,979	$211,462,016	0.56%	3.70%
$218,005,609	$217,688,956	$209,644,613	0.15%	3.84%

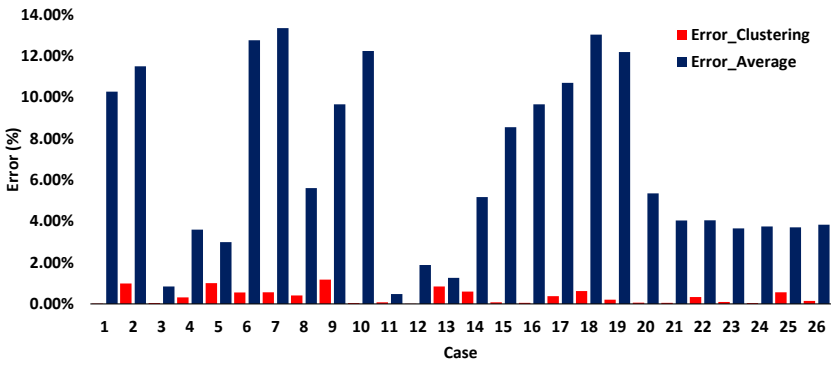

Figure 10. Error comparison between clustering and average.

people can charge during peak hours or not. For public stations, three different scenarios are considered. In one case, they can charge from 6 pm to 10 pm, which means that people will charge their cars after work but not very late. The other case is charging from 10 pm to 2 am to show the effect of charging at late hours, and the last case is charging without any limitation.

For penetration level, 10% and 30% are considered to evaluate the effect of higher demand on the supply network. In each scenario, different percentages are given to each of the charging places to have a comprehensive comparison and reliable results. Monte Carlo simulation is used to generate different load curves as a result of increasing the number of EVs. Then, by using hierarchical clustering, the final huge data set is reduced to 6 points instead of 10,000 points which gives a much faster solving time. This method is compared to simply taking the average and using just one point instead of 10,000. It is shown that while the solving time for both average and clusters are very close and much less than full data (around 5 percent), the results for clustered data are much more accurate and closer to full data, which is the definite and optimum answer.

The comparison is made for the total cost of the problem, including new investment and penalty cost for unmet demands, the percentage of unmet demand, and the amount of new needed investments. It is figured that when charging places are well distributed, the difference between using the average and using clusters is not so much. Although solving times of both are very close, it is recommended to use the clustering method even in these cases, but especially in extreme cases, when the charging percent of one of the places is much higher than the others, the difference is very high, and it is strongly recommended to use hierarchical clustering instead of taking the average.

We utilize a small GEP problem to show the efficiency of clustering. We choose it so that the problem with full data could be solved. If we consider multiple years, different investment options, or more complicated problems, GEP with full data will not be solved. However, GEP with cluster has very little time for this simple case; it tells us that we can use it for more complicated and large problems.

References

[1] Shaukat N, Khan B, Ali SM, Mehmood CA, Khan J et al. A survey on electric vehicle transportation within smart grid system. Renewable and Sustainable Energy Reviews 2018; 81: 1329–1349. doi:10.1016/j.rser.2017.05.092

[2] US Energy Information Administration (EIA), International Energy Outlook 2017 Overview; vol. IEO2017: 143

[3] Bibak B, Tekiner-Mogulkoc H. Influences of vehicle to grid (V2G) on power grid: An analysis by considering associated stochastic parameters explicitly. Sustainable Energy, Grids and Networks 2021; 26. doi:10.1016/j.segan.2020.100429

[4] Bibak B, Tekiner-Mogulkoc H. The parametric analysis of the electric vehicles and vehicle to grid system's role in flattening the power demand. Sustainable Energy, Grids and Networks 2022; 30. doi:10.1016/j.segan.2022.100605

[5] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys 1999; 31 (3): 264–323. doi:10.1145/331499.331504

[6] Augustson JG, Minker J. An Analysis of Some Graph Theoretical Cluster Techniques. Journal of the ACM 1970; 17 (4):571–588. doi:10.1145/321607.321608

[7] Bajcsy P. Hierarchical segmentation and clustering using similarity analysis. PhD, University of Illinois at Urbana-Champaign. ProQuest Dissertations Publishing, USA, 1997.

[8] Kriegel HP , Kröger P, Sander J, Zimek A. Density-based clustering. WIREs Data Mining and Knowledge Discovery 2011; 1 (3):231-240. doi:10.1002/widm.30

[9] Bibak B, Tekiner-Moğulkoç H. A comprehensive analysis of Vehicle to Grid (V2G) systems and scholarly literature on the application of such systems. Renewable Energy Focus 2021; 36:1–20. doi:10.1016/j.ref.2020.10.001

[10] Hannan MA, Azidin FA, Mohamed A. Hybrid electric vehicles and their challenges: A review. Renewable and Sustainable Energy Reviews 2014; 29:135–150. doi:10.1016/j.rser.2013.08.097

[11] Thounthong P, Raël S, Davat B. Control strategy of fuel cell/supercapacitors hybrid power sources for electric vehicle. Journal of Power Sources 2006; 158 (1):806–814. doi:10.1016/j.jpowsour.2005.09.014

[12] Hadley SW, Tsvetkova AA. Potential Impacts of Plug-in Hybrid Electric Vehicles on Regional Power Generation. The Electricity Journal 2009; 22 (10):56–68. doi:10.1016/j.tej.2009.10.011

[13] Momtazpour M, Butler P, Hossain MS, Bozchalui MC, Ramakrishnan N et al. Coordinated Clustering Algorithms to Support Charging Infrastructure Design for Electric Vehicles. Association for Computing Machinery 2012; 126–133. doi:10.1145/2346496.2346517

[14] Shukla A, Verma K, Kumar R. Consumer perspective based placement of electric vehicle charging stations by clustering techniques. 2016 National Power Systems Conference (NPSC) 2016; 1–6. doi:10.1109/NPSC.2016.7858946

[15] Koltsaklis NE, Dagoumas AS. State-of-the-art generation expansion planning: A review. Applied Energy 2018; 230:563–589. doi:10.1016/j.apenergy.2018.08.087

[16] Tekiner H, Coit DW, Felder FA. Multi-period multi-objective electricity generation expansion planning problem with Monte-Carlo simulation. Electric Power Systems Research 2010; 80 (12):1394–1405. doi:10.1016/j.epsr.2010.05.007

[17] Ramírez PJ, Papadaskalopoulos D, Strbac G. Co-Optimization of Generation Expansion Planning and Electric Vehicles Flexibility. IEEE Transactions on Smart Grid 2016; 7 (3):1609-1619. doi:10.1109/TSG.2015.2506003

[18] Hu J, Yan Z, Chen S, Xu X, Ma H. Distributionally Robust Optimization for Generation Expansion Planning Considering Virtual Inertia from Wind Farms. Electric Power Systems Research 2022; 210:1–11. doi:10.1016/j.epsr.2022.108060

[19] Abdin AF, Caunhye A, Zio E, Cardin MA. Optimizing generation expansion planning with operational uncertainty: A multistage adaptive robust approach. Applied Energy 2022; 306(Part A):1–18. doi:10.1016/j.apenergy.2021.118032

[20] Yu X. Impacts assessment of PHEV charge profiles on generation expansion using national energy modeling system. 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century 2008; 1–5. doi:10.1109/PES.2008.4596189

[21] Foley A, Gallachóir BÓ . Analysis of electric vehicle charging using the traditional generation expansion planning analysis tool WASP-IV. Journal of Modern Power Systems and Clean Energy 2015; 3 (2):240–248. doi:10.1007/s40565-015-0126-y

[22] Yao W, Zhao J, Wen F, Dong Z, Xue Y et al. A Multi-Objective Collaborative Planning Strategy for Integrated Power Distribution and Electric Vehicle Charging Systems. IEEE Transactions on Power Systems 2014; 29 (4):1811–1821. doi:10.1109/TPWRS.2013.2296615

[23] Tekiner Mogulkoc H . A Methodology for Explicit Representation Of The Stochastic Demand Due To Electric Vehicles in Generation Expansion Planning Problems. Academic Platform - Journal of Engineering and Science 2021; 9 (2):257–263. doi:10.21541/apjes.821861

[24] Guo C, Bodur M, Papageorgiou DJ. Generation expansion planning with revenue adequacy constraints. Computers Operations Research 2022; 142. doi:10.1016/j.cor.2022.105736

[25] Castro LM, González-Cabrera N. Short-term generation capacity expansion planning considering multi-terminal VSCHVDC links using a linear programming framework improved by shift factors. Electric Power Systems Research 2022; 206. doi:10.1016/j.epsr.2022.107819

[26] Singh B, Kumar Dubey P. Distributed power generation planning for distribution networks using electric vehicles: Systematic attention to challenges and opportunities. Journal of Energy Storage 2022; 48. doi:10.1016/j.est.2022.104030

[27] Borozan S, Giannelos S, Strbac G. Strategic network expansion planning with electric vehicle smart charging concepts as investment options. Advances in Applied Energy 2022; 5. doi:10.1016/j.adapen.2021.100077

[28] Ediz S. Evaluation of the Impacts of Plug-in Hybrid Electric Vehicles on Electricity Load Curve for İstanbul. MSc, İstanbul Şehir Üniversitesi, Turkey, 2017.