Computers & Operations Research 154 (2023) 106202

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

A planar facility location–allocation problem with fixed and/or variable cost structures for rural electrification

Beste Akbas [a], Ayse Selin Kocaman [b,*]

[a] Department of Analytics, Marketing and Operations, Imperial College Business School, London, UK
[b] Department of Industrial Engineering, Bilkent University, Ankara, Turkey

ARTICLE INFO

Keywords:
Continuous location–allocation
Decentralized energy systems
Rural electrification
Mathematical modeling
Clustering

ABSTRACT

One major impediment to developing countries' economic growth is the lack of access to affordable, sustainable, and reliable modern energy systems. Even today, hundreds of millions of people live in rural areas and do not have access to essential electricity services. In this study, we present a planar facility location–allocation problem for planning decentralized energy systems in rural development. We consider nano-grid and micro-grid systems to electrify rural households. While micro-grids serve multiple households with a common generation facility, nano-grids are small-scale systems serving individual consumers. The households served by micro-grids are connected to the generation facilities with low-voltage cables, for which we employ a distance limit constraint due to technical concerns, including power loss and allowable voltage levels. In this problem, we minimize the total investment cost that consists of the facility opening and the low-voltage cable costs. In order to capture the diversity of cost structures in renewable energy investments, we consider three versions of the objective function where we incorporate different combinations of fixed and variable cost components for facilities. For this problem, we provide mixed-integer quadratically constrained problem formulations and propose model-based and clustering-based heuristic approaches. Model-based approaches are multi-stage, in which we solve the discrete counterparts of the problem and employ alternative selection methods for the candidate facility locations. Clustering-based approaches utilize faster clustering techniques to identify the type and location of the facilities. We conduct computational experiments on real-life instances from villages in Sub-Saharan Africa and perform a comparative analysis of the suggested heuristic approaches.

1. Introduction

In 2015 the United Nations member states adopted seventeen Sustainable Development Goals (SDGs) as a universal call "to action to end poverty, protect the planet and improve the lives and prospects of everyone, everywhere" along with a 15-year plan as a part of the 2030 Agenda for Sustainable Development (UNDP, 2023b). Among these goals, Sustainable Development Goal 7 (SDG7) includes specific targets to provide access to "affordable, reliable, sustainable and modern energy for all" (UNDP, 2023a) and has a direct impact on other SDGs such as no poverty (SDG1), quality education (SDG4), economic growth (SDG8) and climate action (SDG14).

The number of people without electricity access is reduced to a low record, 770 million, in 2019. This number, however, is set to increase again in 2020 due to the COVID 19-pandemic, reversing the progress towards SDG7 (IEA, 2022). The majority of the unelectrified populations reside in rural areas as these areas are sparsely populated and usually challenging to reach (World Bank, 2018). Therefore, in order to achieve global access to electricity, the countries must pay significant attention to rural areas, which often require new strategies specific to local geography and demographics (Batidzirai et al., 2021). The centralized grid may be an expensive and challenging option to expand electricity infrastructure to remote and hilly areas (Fobi et al., 2021; Bolukbasi and Kocaman, 2018; Kocaman, 2014; Adkins et al., 2017). Decentralized options such as nano-grid and micro-grids, on the other hand, can provide cost-effective solutions for rural electrification: Nano-grid systems are isolated systems designed to serve only one consumer, such as solar home systems and micro-grid systems serve closely located groups of consumers with a common generation facility such as solar systems, wind turbines, or diesel generators (Akbas et al., 2022).

In this study, we introduce a new problem that determines the locations and allocation of micro and nano-grid facilities on a greenfield in order to contribute to the rural electrification efforts. The lack of restricting infrastructure in greenfield areas allows the facilities

* Corresponding author.
E-mail address: selin.kocaman@bilkent.edu.tr (A.S. Kocaman).

https://doi.org/10.1016/j.cor.2023.106202
Received 26 February 2022; Received in revised form 25 February 2023; Accepted 25 February 2023
Available online 3 March 2023
0305-0548/

to be located at almost any point in continuous space, motivating site-generating facility location–allocation studies as opposed to site-selecting ones (Love et al., 1988). In this planar location and allocation problem, the consumers served by micro-grid facilities are connected to a generation facility with a low-voltage network. Due to technical constraints such as power loss and the maximum permitted voltage drop, these connections are distance limited. Our objective is to minimize the sum of the facility opening costs of micro and nano-grid facilities and the low-voltage connection costs in micro-grid systems.

In an earlier study, Gokbayrak and Kocaman (2017) considered a similar problem, where there is only a single type of facility (i.e., micro-grid) with a fixed opening cost regardless of the number of consumers served by the facility. In this study, we generalize this problem by assuming two different types of decentralized facility options: nano-grid and micro-grid systems. While economies of scale apply to the micro-grid facility costs, the additional cost of the local distribution network can force energy planners to choose between nano-grid and micro-grid systems (Chaurey and Kandpal, 2010), making the cost structure for the generation facility critically important. The cost structure of the decentralized systems can include different components depending on the technologies used to electrify rural households. For example, the cost of modular generators such as solar panels generally depends on the size of the installed module; wind turbines, on the other hand, usually require fixed upfront costs for the given bulk generation capacities. Similarly, the socio-demographics, geographical characteristics, and other country-specific attributes are also the key drivers affecting the cost structures and trends in different markets. Especially the unit costs of solar home systems and solar micro-grid components are observed to change in wide ranges depending on country-specific factors such as the logistics, material, and labor expenses (IRENA, 2016). Therefore, given the diversity in the cost components of different renewable technologies and the significant cost variations in the installation costs, in this study, we consider three different cost structures for the micro and nano-grid systems, (i) fixed costs for both facility types, (ii) variable costs for micro-grids depending on the number of consumers they serve, and (iii) fixed and variable cost for micro-grids.

The contributions of this paper are threefold: First, we introduce a problem that integrates multiple cost components into the decision-making process of the energy planners to compare nano-grid and micro-grid systems, including various renewable technologies. Second, we present mixed-integer quadratically constrained programming formulations for three cost structures to analyze the trade-off between two decentralized energy options and provide a cost-efficient network design for rural settlements. Third, we propose a mathematical modeling-based heuristic with five variants and three clustering-based algorithms for our planar facility location and allocation problem. In the model-based approaches, we make use of the relationship between the planar problems and their discrete counterparts. We first solve the discrete counterpart of the problem by considering the spatial locations of the augmented set of demand points with additional promising points as the candidate locations as in Gokbayrak and Kocaman (2017) and Brimberg et al. (2016). For the given discrete solution, we implement a modified version of Cooper's iterative algorithm (Cooper, 1964) to improve the facility locations on the continuous space and reallocate the households depending on the proposed locations of serving facilities. In addition to the augmentation method proposed in Gokbayrak and Kocaman (2017) for problems with single-type facilities, we propose four new augmentation methods for our problem to improve the solutions obtained from the discrete problem.

In addition to the model-based multi-stage heuristic approaches, we also developed faster heuristic algorithms using top-down and bottom-up clustering techniques. Employing the well-known partitioning algorithms such as agglomerative clustering (Müllner, 2011), DBSCAN (Han et al., 2011), and k-means algorithm (Likas et al., 2003), we provide hybrid solution methods that benefit from simple and practical approaches. For the agglomerative clustering, we also propose a new dissimilarity measure developed based on the problem structure. In the numerical experiments, the clustering-based methods are observed to provide competitive results within significantly shorter computational time and find solutions for large instances that the model-based methods cannot solve within the given time limit.

The sections of this paper are outlined as follows: In Section 2, we present a literature review of planar facility location–allocation problems and clustering algorithms. In Section 3, we introduce a planar facility location–allocation problem in the context of rural energy planning and provide mixed-integer quadratically constrained programming models considering different cost structures. Section 4 provides the details of the multi-stage model-based optimization frameworks and clustering-based heuristic algorithms. Section 5 includes the numerical results and the comparative analysis of the solution methods proposed in Section 4. Finally, Section 6 concludes the paper.

2. Literature review

The problem we propose in this study is related to two well-known problems in the literature: the multisource Weber problem (MWP) and the clustering problem. The MWP is one of the most widely studied problems. The aim is to locate a given number of facilities p, in continuous space while minimizing the weighted sum of the point-to-facility distances. For a given number and type of facilities and no distance limitation between the facility and demand points, our problem reduces to MWP, which is shown to be an NP-hard problem (Megiddo and Supowit, 1984). On the other hand, the clustering problem aims to group objects into different clusters based on their similarity. For a given set of demand points arbitrarily distributed on the Euclidean space, the similarity between two objects can be defined in terms of the Euclidean distances between them. If the households allocated to the same generation facility are considered a 'cluster', then the MWP could also be defined as a variant of the clustering problem. Therefore, clustering techniques can also be helpful for the solution of the MWP.

Brimberg et al. (2008) presented a detailed survey on the continuous location–allocation problems and examined the optimization frameworks, including exact methods, heuristics, and metaheuristics. Due to the non-polynomial nature of the problem and the summation of non-convex terms in the objective function, the literature on the MWP is prone to heuristic algorithms. Exact methods are generally implemented to the small-sized instances as in Ostresh (1975), Drezner (1984), Kuenne and Soland (1972), Rosing (1992) up to 100 customers. However, Krau (1999) presented a column generation approach and a branch-and-bound algorithm that finds optimal solutions for relatively larger instances with 287 demand points and up to 100 facilities.

Contrary to exact solution methods, heuristic approaches and metaheuristics can provide reasonable solutions for large-scale problems with higher computational efficiency. One of the well-known heuristics for MWP was presented by Cooper in 1964 (Cooper, 1964). The idea behind Cooper's iterative algorithm is to locate p facilities and reallocate customers to the nearest facility iteratively until there is no room for improvement. For the given customer-facility assignments, the problem is divided into p single facility location problems, which is much easier to solve by using Weiszfeld's famous iterative procedure (Weiszfeld, 1937). Weiszfeld's algorithm is used to identify the optimal facility location for each subproblem, and it generally converges to the geometric median unless the proposed location coincides with a demand point. However, a modified version of the Weiszfeld's algorithm proposed by Vardi and Zhang (2001) eliminates this necessity to guarantee convergence. Computational experiments indicate that Cooper's algorithm terminates with a local optimum at the end of a small number of iterations (Brimberg et al., 2008).

Unlike Cooper's iterative algorithm, where the location and allocation steps are performed iteratively, Murtagh and Niwattisyawong (1982) presented a large-scale non-linear programming approach to simultaneously decide facility locations and allocations. Similarly, Bongartz et al. (1994) developed an algorithm that simultaneously solves for location and allocation variables using a projection method. The authors derived projection formulas on the sub-spaces of the domain and used them to obtain descent directions. The authors also presented a multi-start version of their method by generating random initial solutions and reported promising results compared to Murtagh and Niwattisyawong (1982). Moreno et al. (1991) presented a constructive type heuristic starting with an initial solution having N clusters, where $p \leq N \leq 2p$. After the clusters are identified, surplus facilities are "dropped" one by one until exactly p facilities remain. An extensive discussion on different strategies based on "add" or "drop" decisions and the comparative analysis of heuristics is presented in Brimberg et al. (2000). Brimberg et al. (2000) also proposed a neighborhood structure, which focuses on the relocation of facilities rather than customer reallocations. This structure diversifies the solutions by expanding the scope of the local search with single moves and systematically examining the unexplored candidate locations. The facility locations are updated one at a time, and all possible single moves construct the neighborhood exchange. Cooper's iterative algorithm is then implemented with the selected portion of the one-exchange neighborhood. Another constructive algorithm is developed by Gamal and Salhi (2001) based on the furthest distance rule to avoid clustered facilities at specific regions.

Another stream of heuristic approaches for planar problems that we also aim to contribute to with this paper is built on the relation between the planar problems and their discrete counterparts. Hansen et al. (1998) proposed a heuristic where the discrete p−median problem is solved using the demand points as the candidate facility locations so as to divide the problem into p distinct single facility location problems. After the optimal solution to the discrete problem is obtained, facility locations are improved individually in the continuous space for each cluster. This one-step "continuous-space adjustment" idea is then improved in Brimberg et al. (2014). They suggested an approach that iteratively shifts between the discrete and continuous versions of the problems. This approach is revisited in Brimberg et al. (2016) to show the benefit of augmenting the candidate facility location set while solving the discrete problem with good injection points.

The prior studies that we have discussed so far primarily focused on the variations of multi-start Cooper's alternate heuristic or neighborhood structures. One of the first metaheuristic attempts is presented in Brimberg and Mladenovic (1996) using basic Tabu Search rules. Houck et al. (1996) suggested a different approach by using a genetic algorithm to solve the multi-source Weber problem. Similarly, Salhi and Gamal (2003) presented a genetic algorithm where the selection and removal process is based on groups of chromosomes rather than single entities. Hence, they introduced three categories of chromosomes, good, mediocre, and poor, to diversify the search process and avoid early convergence. However, the method fails for a large number of facilities. Drezner et al. (2015) proposed a variation of the genetic algorithm presented in Salhi and Gamal (2003) by using an effective merging process to generate off-springs. The new hybrid approach combining the distribution-based variable neighborhood search with the genetic algorithm is shown to obtain improved solutions compared to the basic variable neighborhood search and genetic algorithm.

Network design problems for rural electrification are usually subject to a distance constraint due to transmission loss concerns and maximum permitted voltage drop. Drezner et al. (1991) proposed a distance-limited version of the Weber problem where the service provided by a distant facility is considered useless if the cut-off distance is exceeded. Gokbayrak and Kocaman (2017) introduced a distance-limited continuous location–allocation problem, where the number of facilities is considered a decision variable. They proposed a multi-stage solution approach that involves developing promising candidate locations for the discrete counterpart of the problem and solving Cooper's iterative algorithm for fine-tuning on continuous space. Similarly, in Gokbayrak and Avci (2020), multi-stage methods that include the solution of the discrete problem and fine-tuning steps are presented for a similar problem with a multi-point low-voltage network. Kocaman et al. (2012) proposed an agglomerative clustering technique for a two-level power distribution network, where there is a distance limitation between transformers and the serving transformers. Assuming each customer is a singleton cluster initially, the algorithm searches for the two closest groups and locates a single facility to the center of mass in an iterative fashion. If the proposed site does not violate the distance limit, these two clusters are merged into a single cluster, and the current facility locations are replaced with the new one located at the centroid. This merging process continues until none of the clusters can be merged due to the distance limit. The algorithm provides the configuration having the least overall cost as the final solution.

The other well-known clustering techniques are also frequently used to solve multi-facility location problems. For instance, Esnaf and Küçükdeniz (2009) proposed a hybrid method using the fuzzy c-means clustering method. After the initial clusters are formed using the fuzzy c-means clustering algorithm, the locations of the facilities in each cluster are optimized individually as a single facility location problem. Similarly, Küçükdeniz and Esnaf (2018) also presented an approach where the initial clusters are formed by the fuzzy c-means clustering, and the existing configuration is optimized with the NM simplex algorithm. Geetha et al. (2009) developed a k-means-based solution approach to create k disjoint clusters and identify the optimal customer-facility allocations. Sahraeian and Kaveh (2010) presented another hybrid method combining k-means clustering with the fixed neighborhood search algorithm. Based on the initial locations that the k-means method yields, the fixed neighborhood algorithm improves the facility locations and the customer allocations. However, the final output of the k-means algorithm is susceptible to the randomly selected centroids selected as the initial solution. Therefore, the k-means algorithm is repeated multiple times to choose the best solution among several alternatives with different initial seedings.

Corigliano et al. (2021) also used the k-means and agglomerative hierarchical clustering algorithms to locate secondary substations as a part of the power distribution network design problem. Firstly, the k-means clustering algorithm is used for a predetermined number of substations. Then, the low-voltage connections are checked if the distance threshold is exceeded. The clusters violating the distance limitation are all subdivided iteratively until each one conforms to the distance threshold. Secondly, an agglomerative clustering-based approach is presented to merge consumers one by one until the maximum number of clusters is attained within the distance limitation. In the comparative analysis, the agglomerative clustering approach is found to be more effective than the other methods proposed in the study. Another transformer substation siting problem discussed by González-Sotres et al. (2013) utilized the k-means algorithm to divide rural settlements into small regions. The proposed algorithm starts with $k = 1$ and increments the number of transformers by one at each iteration. The algorithm records the total cost of the distribution network and the transformers for each cluster and calculates the medium voltage network cost between the transformers. The algorithm returns the least-cost configuration as the final solution.

Apart from the k-means and agglomerative clustering approaches, DBSCAN (Density-based spatial clustering of applications with noise) is another solution method for facility location problems. Sharma et al. (2014) proposed a two-stage solution methodology using the DBSCAN algorithm and affinity propagation. Similarly, Sharma and Jalal (2017) also presented a hybrid DBSCAN and linear programming-based approach to solving a facility location problem. After clustering the consumers using DBSCAN, the customer allocations to the facilities are optimized using a linear programming model.

The aforementioned literature on planar facility location–allocation problems could also be linked to the planar location-routing problems (PLRP), for which pioneer studies belong to Schwardt and Dethloff

Table 1
Literature review.

Authors	Facility type	Number of facilities	Distance limit	Cost structure	Solution methodology
Kuenne and Soland (1972)	Single	Predetermined	✗	–	Branch and Bound Algorithm
Murtagh and Niwattisyawong (1982)	Single	Predetermined	✗	–	NLP
Drezner et al. (1991)	Single	Predetermined	✓	–	Enumerative Algorithm
Moreno et al. (1991)	Single	Predetermined	✗	–	Drop and Add Heuristic
Rosing (1992)	Single	Predetermined	✗	–	LP
Bongartz et al. (1994)	Single	Predetermined	✗	–	Projection
Houck et al. (1996)	Single	Predetermined	✗	–	Genetic Algorithm
Hansen et al. (1998)	Single	Predetermined	✗	–	Discrete Problem + Projection
Krau (1999)	Single	Predetermined	✗	–	Branch and Bound Algorithm
Gamal and Salhi (2001)	Single	Predetermined	✗	–	Multi-start Cooper's Algorithm Furthest Distance Method Perturbation Heuristic + Cooper's Algorithm
Salhi and Gamal (2003)	Single	Predetermined	✗	–	Genetic Algorithm
Brimberg et al. (2004)	Single	Decision Variable	✗	Fixed	Discrete Problem + Cooper's Algorithm
Brimberg and Salhi (2005)	Multiple	Decision Variable	✗	Fixed	Discrete Problem + Cooper's Algorithm
Geetha et al. (2009)	Single	Predetermined	✗	–	K-Means Clustering
Esnaf and Küçükdeniz (2009)	Single	Predetermined	✗	–	Fuzzy C-means Clustering Algorithm with Nelder–Mead Simplex Algorithm
Sahraeian and Kaveh (2010)	Single	Predetermined	✗	–	K-Means Clustering, Fixed Neighborhood Search
Sharma et al. (2014)	Single	Decision Variable	✗	Fixed	DBSCAN Affinity Propagation
Drezner et al. (2015)	Single	–	✗	Fixed	Genetic Algorithm, Variable Neighborhood Search
Luis et al. (2015)	Multiple	Decision Variable	✗	Fixed	Region-rejection Algorithm, GRASP
Hosseininezhad et al. (2015)	Multiple	Decision Variable	✗	Fixed	Cross Entropy Meta-heuristic
Gokbayrak and Kocaman (2017)	Single	Decision Variable	✓	Fixed	Discrete Problem + Projection
Sharma and Jalal (2017)	Single	Decision Variable	✗	Fixed	DBSCAN MILP
Küçükdeniz and Esnaf (2018)	Single	–	✗	Fixed	Fuzzy C-means Clustering Algorithm
Gokbayrak and Avci (2020)	Single	Decision Variable	✓	Fixed	Discrete Problem + Projection + Esau-Williams Heuristic
Irawan et al. (2020)	Multiple	Decision Variable	✗	Fixed	Variable Neighborhood Search, Simulated Annealing
This Paper	**Multiple**	**Decision Variable**	✓	**Fixed, Variable, Fixed & Variable**	**Discrete Problem + Projection Clustering-based Heuristics**

(2005), Schwardt and Fischer (2009), and Salhi and Nagy (2009). Their work then led to the studies of Manzour-al Ajdad et al. (2012) and Irawan et al. (2022). In their recent study, Irawan et al. (2022) tackled an interesting variant of the PLRP and proposed a new optimization model and solution algorithms for the location and maintenance of offshore wind farms.

In Table 1, we summarize the studies on planar location–allocation problems reviewed in this section. Table 1 shows that only a small number of studies considered the number of facilities as a decision variable and included their costs into the problem setting. Brimberg et al. (2004) introduced the multi-source Weber problem with constant opening cost and Brimberg and Salhi (2005) considered zone-dependent fixed costs for the same problem. Hosseininezhad et al. (2015) considered the zone-based fixed cost for the capacitated multi-source Weber problem, and Luis et al. (2015) studied a similar problem with constant, zone-based, and continuous fixed cost functions. A single-source capacitated multi-facility Weber problem with fixed setup costs is recently investigated by Irawan et al. (2020).

Among the papers we have reviewed, Gokbayrak and Kocaman (2017) is the most related problem. Unlike MWP, Gokbayrak and Kocaman (2017) introduced a fixed opening cost and distance limit for facilities and defined the number of facilities as a decision variable. However, Gokbayrak and Kocaman (2017) only considers a single type of facility with a fixed cost structure. In this paper, we consider different types of facilities depending on their number of connected customers. Moreover, given the variations in the cost structure of different power generation technologies, it becomes necessary to consider different combinations of fixed and variable cost components for the design of energy systems. Because in addition to fixed upfront costs for the generation technologies with bulk capacities, highly modular generation technologies that could be adapted to small or large-scale systems may require variable cost components. Therefore, this study introduces a planar facility location–allocation problem with fixed and/or variable cost structures to design decentralized energy systems for rural electrification considering nano-grid and micro-grid options. We introduce new approaches based on mathematical modeling and clustering techniques to solve this problem.

3. Problem formulation

This study proposes a planar facility location–allocation problem with fixed and/or variable cost structures for rural and underdeveloped communities' electrification. The decentralized electrification options considered in this problem involve nano-grid and micro-grid systems. Micro-grids denote a small set of households electrified together by a single generation facility, whereas nano-grids are isolated standalone systems generating electricity for individual consumers. The households electrified by micro-grids are directly connected to the generation points with low-voltage cables. Therefore, we also design a single-level low-voltage network in a star topology to distribute the electricity to final consumers. As in Gokbayrak and Kocaman (2017), low-voltage connections cannot exceed a specific distance limit, $distLim$, due to technical constraints such as power loss and voltage drop limitations.

In a small rural settlement, the set of demand points is denoted by $\mathcal{N} = \{1, \ldots, N\}$. The coordinates of each demand point $i \in \mathcal{N}$ are given as (a_i, b_i), and N denotes the number of demand points that need to be electrified using an either nano-grid or micro-grid option. We assume that households are identical, and thus the consumption level is the same for all households. The index set of the candidate generation points is denoted by $\mathcal{G} = \{1, \ldots, N\}$ and each generation point $j \in \mathcal{G}$ is located at (c_j, d_j). Because the maximum number of generators is obtained when each household has a nano-grid, the upper bound on the number of generation points is equal to the number of households $|\mathcal{G}| = N$ in the continuous problem. The generation points can be located anywhere on the continuous space; therefore, (c_j, d_j), $\forall j \in \mathcal{G}$ are considered continuous decision variables in the mathematical formulations.

The objective of the problem is to minimize the total investment cost. In addition to the costs of the facilities, low-voltage network cost is also considered to calculate the total investment required. We assume that the distances between generation points and households are Euclidean. The low-voltage connection cost per unit distance is denoted by c_L, and the total connection cost is proportional to the total length of low-voltage cables in micro-grids.

The nano-grid and micro-grid systems have different facility opening costs. Depending on the type of generators in micro-grids and nano-grids, we use different cost structures reflecting economies of scale and making micro-grids an attractive alternative to nano-grids. Besides the fixed upfront cost of the systems with bulk generation capacities, we also incorporate variable-cost components for the highly modular technologies that could be easily scaled up to different sizes. Below, we first provide a mixed-integer quadratically constrained programming (MIQCP) model for the most general case, where we consider both fixed and variable cost components for micro-grid investments. We denote this formulation as PFLAP-Fixed&Var. Then we discuss the formulations of the cases where we have only fixed or variable cost components.

We use the following notations in the models:

Decision Variables:

$$x_{ij} = \begin{cases} 1, & \text{if household } i \in \mathcal{N} \text{ is served by generation facility } j \in \mathcal{G} \\ 0, & \text{otherwise} \end{cases}$$

$$v_j = \begin{cases} 1, & \text{if generation facility } j \in \mathcal{G} \text{ is open} \\ 0, & \text{otherwise} \end{cases}$$

$$k_j = \begin{cases} 1, & \text{if generation facility } j \in \mathcal{G} \text{ is a nano-grid} \\ 0, & \text{otherwise} \end{cases}$$

$$t_j = \begin{cases} 1, & \text{if generation facility } j \in \mathcal{G} \text{ is a micro-grid} \\ 0, & \text{otherwise} \end{cases}$$

γ_i : Distance between household $i \in \mathcal{N}$ and the facility it is assigned to

d_{ij} : Distance between household $i \in \mathcal{N}$ and facility $j \in \mathcal{G}$

d^x_{ij} : x-coordinate difference between household $i \in \mathcal{N}$ and facility $j \in \mathcal{G}$

d^y_{ij} : y-coordinate difference between household $i \in \mathcal{N}$ and facility $j \in \mathcal{G}$

(c_j, d_j) : Coordinates of the facility $j \in \mathcal{G}$

Parameters:

$F1$: Micro-grid facility cost

$F2$: Nano-grid facility cost

$F3$: Micro-grid cost per household

$distLim$: Distance limit

$L \times W$: Dimensions of the rectangular greenfield region

N : Number of households

c_L : Low-voltage connection cost per unit distance

(a_i, b_i) : Coordinates of the household $i \in \mathcal{N}$

Sets:

$\mathcal{N}$: The set of households

$\mathcal{G}$: The set of facility locations

PFLAP-Fixed&Var:

$$\min \sum_{i\in\mathcal{N}} c_L \gamma_i + \sum_{j\in\mathcal{G}} F1.t_j + \sum_{j\in\mathcal{G}} (F2 - F3).k_j + F3.N \tag{1}$$

$$\text{s.t.} \sum_{j\in\mathcal{G}} x_{ij} = 1, \quad i \in \mathcal{N} \tag{2}$$

$$x_{ij} \le v_j, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{3}$$

$$\sum_{i\in\mathcal{N}} x_{ij} \le k_j + N t_j, \quad j \in \mathcal{G} \tag{4}$$

$$k_j + t_j = v_j, \quad j \in \mathcal{G} \tag{5}$$

$$d^x_{ij} = a_i - c_j, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{6}$$

$$d^y_{ij} = b_i - d_j, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{7}$$

$$d^2_{ij} \ge (d^x_{ij})^2 + (d^y_{ij})^2, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{8}$$

$$\gamma_i \ge \sqrt{L^2 + W^2}(x_{ij} - 1) + d_{ij}, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{9}$$

$$\gamma_i \le distLim, \quad i \in \mathcal{N} \tag{10}$$

$$v_j \in \{0, 1\}, \quad j \in \mathcal{G} \tag{11}$$

$$k_j \in \{0, 1\}, \quad j \in \mathcal{G} \tag{12}$$

$$t_j \in \{0, 1\}, \quad j \in \mathcal{G} \tag{13}$$

$$x_{ij} \in \{0, 1\}, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{14}$$

$$c_j,\ d_j \in \mathbb{R}, \quad j \in \mathcal{G} \tag{15}$$

$$d^x_{ij},\ d^y_{ij} \in \mathbb{R}, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{16}$$

$$d_{ij} \ge 0, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{17}$$

$$\gamma_i \geq 0, \qquad i \in \mathcal{N} \qquad (18)$$

The objective function in (1) minimizes the total investment cost that involves the cost of deploying generation facilities and the total low-voltage connection cost. Constraints (2) guarantee that each household is served by exactly one generation facility. We ensure in (3) that closed facilities cannot provide electricity service. Constraints (4) determines if a facility serves more than one consumer. In other words, the type of facility, namely nano-grid and micro-grid, is identified by these constraints. The constraint set (5) indicates that an open facility is either a nano-grid or a micro-grid. The x-coordinate and y-coordinate differences are defined in constraints (6) and (7), respectively. Using the differences in x,y coordinates and the quadratic constraints (8), the Euclidean distances between the facilities and the households are determined. The distance between the households and the serving facilities is calculated in constraints (9). The constraint set (10) imposes a distance limit on the low-voltage connections between the micro-grid facilities and the households. Finally, the decision variables are defined in (11)–(18).

Based on our problem definition, $F3$ takes a non-negative value that is less than or equal to $F2$ so that a micro-grid system can be a cost-efficient alternative. One can see that the distance limit constraint can be rewritten as $min(distLim, (F2 - F3)/cL)$, since the longer connections would not be cost-efficient considering the trade-off between low-voltage connections and nano-grid facility costs.

Moreover, when $F3 = 0$, the same formulation can be used to solve the problem where there are only fixed cost components. However, note that when $F1 = 0$, the formulation above does not reduce to that of variable cost problem where the micro-grid facility costs are dependent only on the number of households it serves. When $F1 = 0$ and $F3 \leq F2$, the above formulation assigns even singleton clusters as micro-grids since building nano-grid is not cost-efficient. Therefore, for the variable cost problem, we replace the constraint sets (4)–(5) with (19)–(20) and propose the following formulation:

$$\min \quad \sum_{i \in \mathcal{N}} c_L \gamma_i + \sum_{j \in \mathcal{G}} (F2 - F3).k_j + F3.N$$

s.t. (2)–(3) & (6)–(18), *and*

$$\sum_{i \in \mathcal{N}} x_{ij} + k_j \geq 2v_j, \qquad j \in \mathcal{G} \qquad (19)$$

$$k_j \leq v_j, \qquad j \in \mathcal{G} \qquad (20)$$

In the variable cost formulation above, based on the constraint set (19), it is still feasible to assign clusters as nano-grid even if the facility serves more than one household. Note that we do not need additional constraints to eliminate such feasible solutions because the microgrid cost per household ($F3$) is less than or equal to the nano-grid cost ($F2$). The objective function forces $k_j = 0$ in this case, since classifying a micro-grid as a nano-grid ($k_j = 1$ instead of $k_j = 0$) would bring an additional cost and make the solution suboptimal.

4. Solution methodology

Our problem reduces to the MWP, which is shown to be NP-Hard (Megiddo and Supowit, 1984), for a predetermined number of single-type facilities and relaxed distance limitation. Therefore, in this study, we propose heuristic methods for the solution of our planar location–allocation problem. We classify these methods as model-based or clustering-based approaches.

4.1. Model-based heuristic approaches

Based on the discussions on optimal facility locations being very close to demand points (Brimberg et al., 2014), initializing the set of demand points as candidate facility locations may provide reasonable estimates for the continuous problem. Hence, following a similar approach that Hansen et al. (1998) presented, here we introduce five variants of a multi-stage approach in which we solve our facility location problem in the discrete space and then improve the existing configuration in the continuous space with projections. The model-based approaches differ in the first stage, where we determine the promising candidate facility locations for the discrete problem.

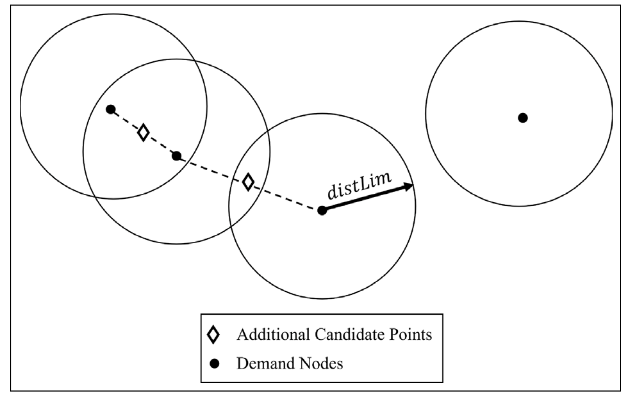

Fig. 1. Illustration of MB-I.

4.1.1. Stage-1: Identifying candidate facility locations

In the first stage of the multi-stage approach, we identify the candidate facility locations to be used in the discrete counterpart of the problem. We propose five approaches to identify additional promising candidate facility locations that will be used to augment the set of demand points.

MB-I: Mid-points as the candidate facility locations

Each demand point can be connected to a facility only if the facility is located within $distLim$ radius. Therefore, any two demand points served by the same facility must be located within at most $2 \times distLim$ distance. When the distance between two demand points is $2 \times distLim$, the facility should be located in the middle of these points. In the first model-based (MB-I) heuristic approach, we identify the pairs of demand points located less than $2 \times distLim$ distance from each other. We augment the set of demand points with the mid-point of the line segments connecting these pairs and use this augmented set as the set of candidate facility locations in the discrete problem. These candidate points are illustrated in Fig. 1.

MB-II: PCSP solutions as the candidate facility locations

In the second MB approach (MB-II), we determine the set of candidate facility locations by finding the solutions to the planar set covering problem (PSCP) as in Gokbayrak and Kocaman (2017). For this, we first find the set of circle intersection points drawing circles with a radius of $distLim$ around each demand point as shown in Fig. 2. These circle intersection points are suggested by Church and ReVelle (1974) to be used to find an optimal solution to the PSCP by solving a discrete set covering problem. We, then, solve the discrete set covering problem (SCP) using the demand points and the circle intersection points as the candidate facility locations and use the solution of this model to augment the set of demand points to be used in the discrete version of our original problem. For further details on finding the optimal solution of PSCP, the readers may refer to Gokbayrak and Kocaman (2017).

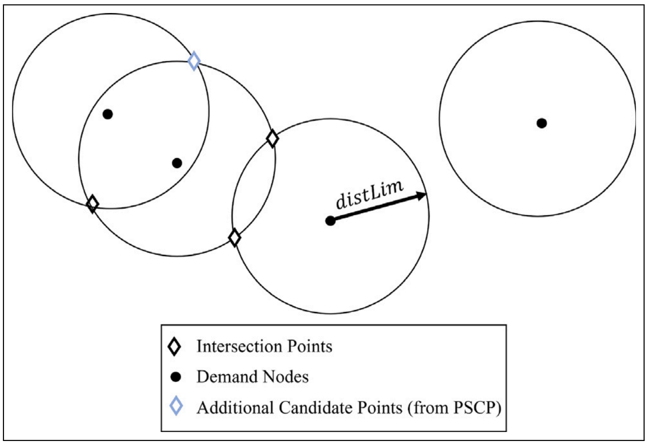

Fig. 2. Illustration of MB-II.

MB-III: Common intersection points as the candidate facility locations

During the computational experiments, it is observed that SCP might provide multiple optimal solutions, especially when several intersection points enclose the same demands nodes. However, MB-II considers only one of these solutions. The SCP results that we use in our discrete problems play a vital role in the optimal system configuration. Because the objective function accounts for the total distance between the demand points and the facilities, the optimal solution of our discrete problem could be significantly affected by the choices we make in this stage. In Fig. 3, we illustrate this argument with a small example. The total LV connections are observed to change considerably even in this small instance for the different intersection points, which are the alternative optimal solutions of the SCP.

As an alternative augmentation methodology, we identify other intersection points covering the same demand nodes that the optimal solution of the PSCP has covered. We refer to these points as common intersection points. As a result, instead of inserting only the optimal locations obtained from PSCP into the set of demand points, we consider these additional common intersection points as the candidate facility locations. At the expense of extra computational time in Stage-2 compared to MB-II, these additional points might provide benefits on the objective function..

MB-IV: Candidate facility locations from the convex hull of the common intersection points

As an alternative augmentation approach, here we identify the groups of common intersection points as in MB-III. Then we find the convex hulls covering the points in each group and add the centroids and the corner points of these convex hulls to the set of candidate locations. Fig. 4 illustrates the identification of the additional candidate sites.

MB-V: Candidate facility locations from the convex hull of the demand points

In the final approach, we propose to solve the PSCP as in MB-II and obtain the least number of facilities required to cover all demand nodes (n^*). Then, we select the same number of random points from the convex hull covering all demand points and add them to the set of candidate facilities (see Fig. 5). Please note that this method includes a random process. It can be expected to obtain better results by repeating the random selection process multiple times at the expense of the significantly increased solution time.

4.1.2. Stage-2: Determining the facilities and initial household allocations

In this stage, we find the least-cost solution for the discrete space using the candidate facility locations obtained in the previous stage. The notations in the discrete counterpart are the same as in the continuous problem. However, unlike the continuous formulation, $\mathcal{G}$ will denote the augmented set of demand points to be used as the candidate facility locations in the discrete problem.

Below we provide the formulation for the discrete facility location–allocation problem with the most general cost structure (i.e., fixed and variable cost), which is denoted by DFLAP-Fixed&Var.

DFLAP-Fixed&Var:

$$\min \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{G}} c_L x_{ij} d_{ij} + + \sum_{j\in\mathcal{G}} F1.t_j + \sum_{j\in\mathcal{G}} (F2 - F3).k_j + F3.N$$

s.t. (2)-(5)

$$\sum_{j\in\mathcal{G}} x_{ij} d_{ij} \le distLim, \quad i \in \mathcal{N} \tag{21}$$

$$v_j \in \{0,1\}, \quad j \in \mathcal{G} \tag{22}$$

$$k_j \in \{0,1\}, \quad j \in \mathcal{G} \tag{23}$$

$$t_j \in \{0,1\}, \quad j \in \mathcal{G} \tag{24}$$

$$x_{ij} \in \{0,1\}, \quad i \in \mathcal{N},\ j \in \mathcal{G} \tag{25}$$

The discrete formulation is almost equivalent to the continuous counterparts except for the domain constraints and the objective functions. Constraints(21) ensure that each household will be electrified by a facility within the distance threshold. The remaining constraint sets (22)–(25) define the decision variables. Note that we use (19)–(20) instead of (4)–(5) in the variable cost problem.

In the following stage of the heuristic approach, we use the initial configuration attained in the solution of DFLAP. Considering the initial facility locations and household allocations, we implement a modified version of Cooper's iterative algorithm with a new allocation step to improve the total investment cost and provide closer estimates of the solution to the continuous problem.

4.1.3. Stage-3: Cooper's iterative algorithm with a new allocation strategy

In the third stage, we perform a modified version of Cooper's well-known heuristic, which iteratively improves the existing configuration while alternating between location and allocation steps. Each cluster is treated as a Weber problem to optimize the facility locations for the given household allocations obtained in the previous stage. Since we assume that the households are identical in terms of their demand, the solutions of the Weber problem converge to the geometric median point, which minimizes the total household-to-facility distances for each cluster. Nevertheless, no straightforward formulation in the literature calculates the geometric median. Weber (1929) proved that an explicit formulation could not be derived to find the geometric median. Therefore, convergence algorithms are extensively used in order to generate closer estimates. Considering the convexity property of the summation of convex functions (distance function), these approximation algorithms generally do not fall into local optimality traps with one exception (Brimberg, 2016). Weiszfeld's algorithm presented in Weiszfeld (1937) converges to the geometric median unless the proposed facility location overlaps a demand point in one of the iterations. However, an extension of the algorithm proposed by Vardi and Zhang (2001) eliminates this problem. Hence, in the location step of Cooper's heuristic, we implement the modified version of Weiszfeld's algorithm (Vardi and Zhang, 2001) in order to project the facility locations to the geometric median points considering our distance limit.

As Vardi and Zhang's convergence algorithm does not consider any distance limitation, the suggested facility location does not necessarily obey this limit. After the algorithm converges to a location, we project the proposed location to a feasible point that conforms to the distance

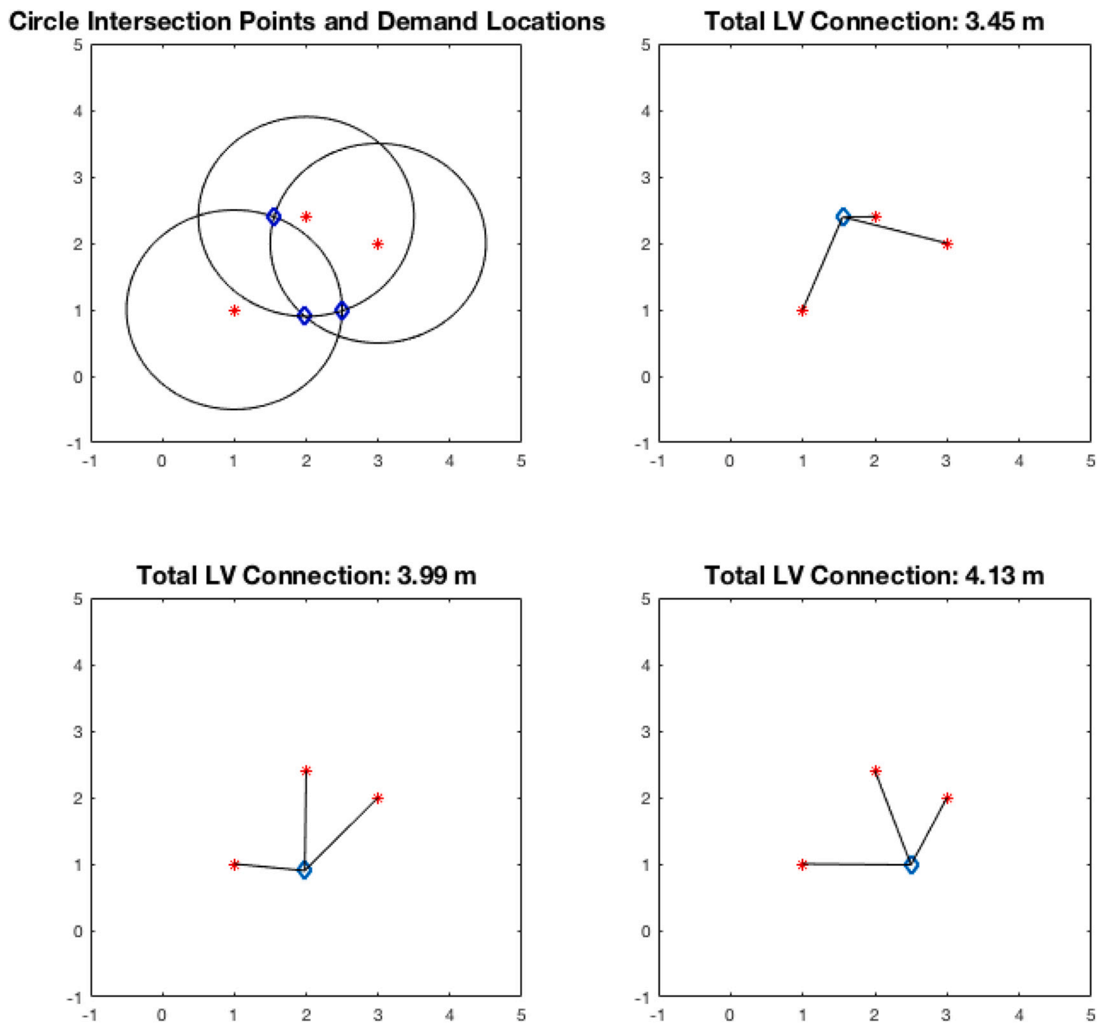

Fig. 3. The Effect of the SCP solution on the low-voltage (LV) cable costs.

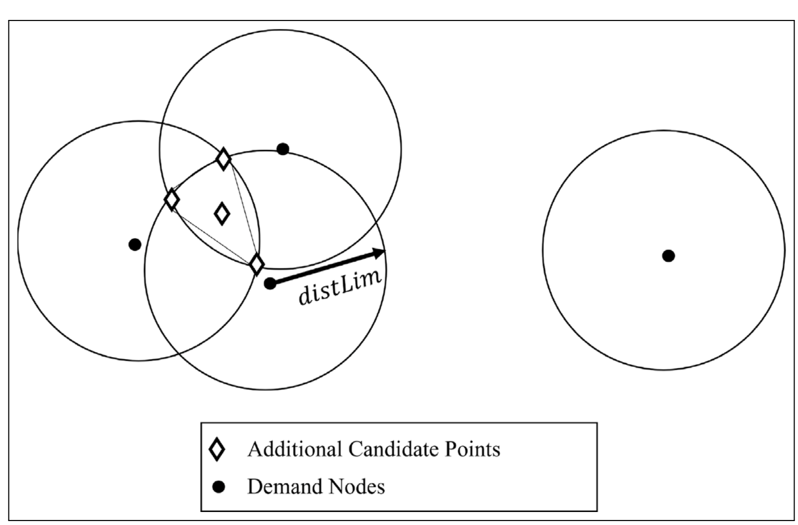

Fig. 4. Illustration of MB-IV.

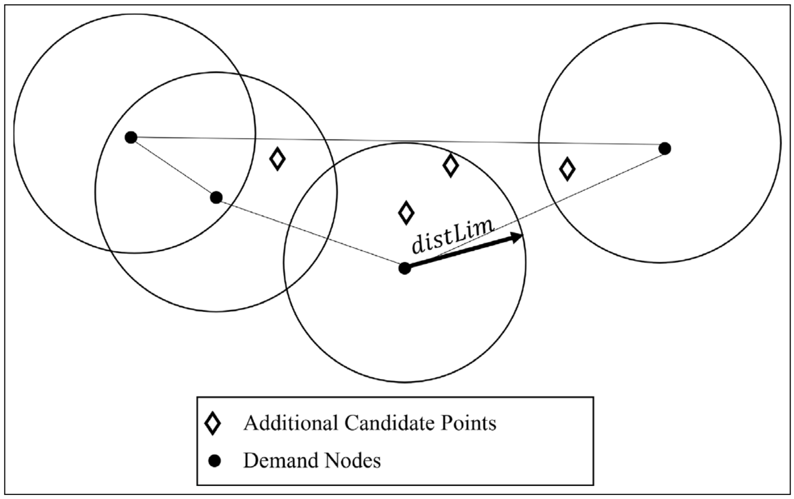

Fig. 5. Illustration of MB-V.

limit constraint for each demand point in the cluster, using the projection methodology proposed in Gokbayrak and Kocaman (2017). We call this new facility location the distance-limited geometric median.

Once the facilities' locations are determined in the location step of Cooper's approach, households should be reallocated to the facilities in a cost-efficient way. Unlike the original allocation step of Cooper's algorithm, allocating the households to the nearest facility may not be a cost-efficient move in our case since the decentralized systems have different facility costs. One of the following two cases might occur: i) One member of a micro-grid connecting at least three households can be reallocated to a former nano-grid. This refers to a case when a micro-grid has at least three households, and if one of its members is reassigned to a nano-grid, the facility type of the latter changes while the former remains the same. In this case, as the former nano-grid transforms into a micro-grid, the cost of this new micro-grid cluster needs to be recalculated accordingly. ii) The transfer of a household from a micro-grid connecting two households to another micro-grid. This refers to the case in which the first micro-grid transforms into a nano-grid while the other remains as a micro-grid. In this case, one might expect a reduction in the facility costs since the nano-grid facility's cost is less than the micro-grid's. This highlights the necessity of recalculating the facility costs, as the generator types can change after each reallocation. Therefore, we propose to allocate the households one by one to their nearest facility, considering the two cases above and calculating the potential cost improvements. If the nearest facility for a particular demand point in one iteration is different from its already serving facility, we perform the reallocation only if it is cost-efficient. At the end of consumer allocations, we also check each cluster to see if the current micro-grid configurations yield a lower cost than electrifying the households individually. If it is more cost-efficient when each household is electrified by individual stand-alone systems (nano-grid), we break this particular cluster into singleton clusters. Note that location and allocation steps are repeated until the change in the objective value is less than a threshold value.

4.2. Clustering-based heuristic approaches

The computational experiments have demonstrated that model-based heuristic approaches may have difficulty solving the mathematical models optimally within a predetermined time limit. Hence, we also propose clustering-based heuristic algorithms to provide immediate solutions without needing a commercial solver for energy planners to make rapid assessments. Thus, we present three heuristic approaches based on fast clustering techniques, including density-based spatial clustering applications with noise (DBSCAN), k-means algorithm, and agglomerative clustering. Since clustering problems and MWP have similarities by definition, clustering analysis has been frequently used as a solution methodology for planar facility location–allocation problems (e.g., Corigliano et al. (2020),(Corigliano et al., 2021; González-Sotres et al., 2013; Yu et al., 2018)).

4.2.1. Agglomerative clustering-based heuristic approach

Our first clustering-based approach involves hierarchical agglomerative clustering, a well-known technique for grouping objects into clusters based on their proximity. It is a bottom-up approach, where each entity is initially considered an individual cluster (i.e., each demand point is a singleton cluster) and merged to create larger clusters based on their dissimilarity.

In Fig. 6, we provide a flowchart of the proposed method. In this approach, we use our distance limit as the stopping criterion for agglomeration. Hence, agglomeration is performed iteratively based on the proximity of the clusters until the most similar pairs cannot be merged anymore due to the distance threshold. At each step of the clustering, we identify the least dissimilar pair of clusters based on a specific dissimilarity measure and locate a new single facility to the distance-limited geometric median of the merged clusters. To determine the facility location for each cluster, we employ Vardi and Zhang's modified Weiszfeld algorithm and the projection methodology to shift the suggested facility location to the distance-limited geometric median, which is mentioned in Section 4.1.3 previously. While the number of clusters is reduced iteratively, we record the total cost. We report the configuration with the least-cost design as the final solution. However, depending on the trade-off between facility costs, some micro-grid clusters may not reach an ideal number of households, making nano-grids a cost-competitive alternative. Hence, we apply a fine-tuning step to disintegrate the expensive micro-grid clusters into singleton clusters or assign some individual stand-alone systems to the closest micro-grid within the distance limit if it is cost-efficient.

The dissimilarity measure plays a vital role in the agglomeration process. Various measures have been proposed in the literature to define the dissimilarity between clusters. The comparative analysis of these measures has indicated no clear evidence showing that a specific measure is superior to others. Hence, we investigate well-known dissimilarity measures such as single, complete (Johnson, 1967), average (Sneath and Sokal, 1963), minimax (Bien and Tibshirani, 2011), centroid, Hausdorff distance (Hausdorff, 1957), and Ward's minimum variance (Ward, 1963) to identify the best option for each problem among different alternatives.

In addition to these existing measures, we propose an additional measure called GeomDiff. Inspired by Ward's measure (Ward, 1963),

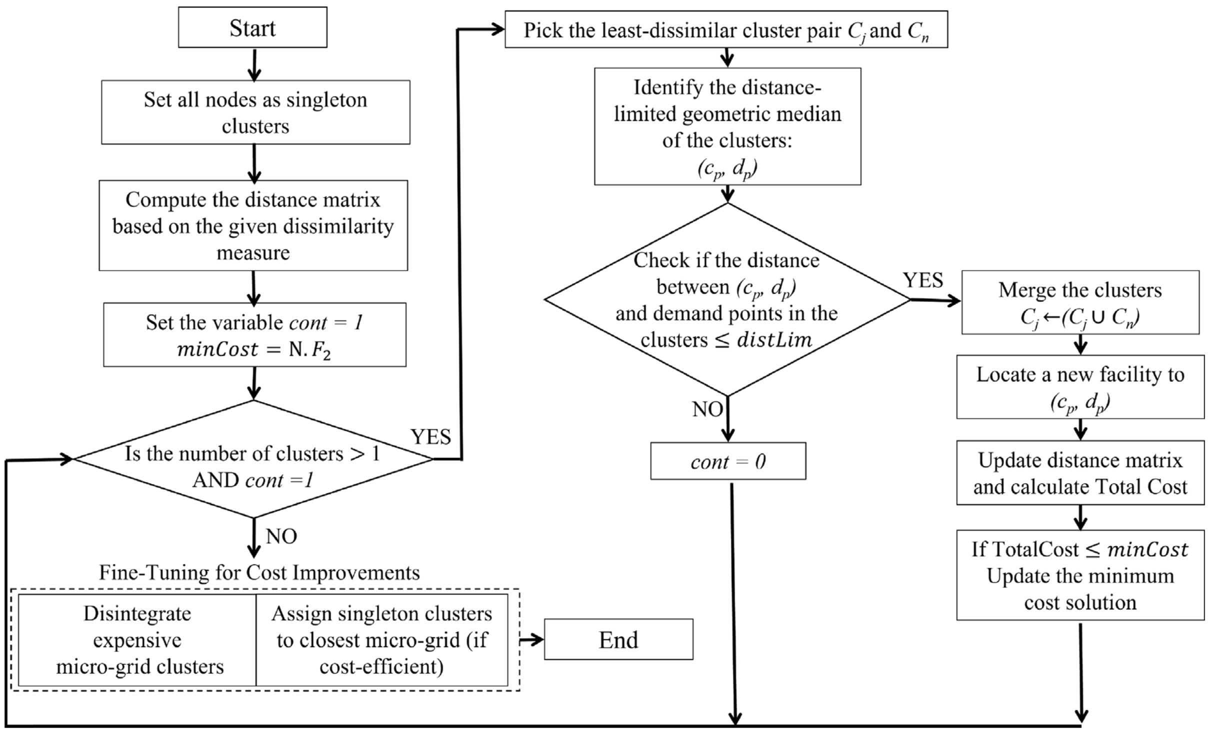

Fig. 6. Flowchart of agglomerative clustering-based heuristic approach.

which calculates the dissimilarity between clusters as the increase in the total within-cluster sum of squares due to agglomeration, this new measure considers the increase in the sum of point-to-facility distance due to agglomeration. In other words, we consider the potential increase in the total length of the low-voltage connections to measure the dissimilarity between the two clusters. Let us denote the total sum of demand point-to-facility distances of the cluster C_j as $Geom(C_j)$. Then, the objective of the GeomDiff is formulated as follows:

$$d(j,n) = Geom(C_j \cup C_n) - Geom(C_j) - Geom(C_n) \tag{26}$$

4.2.2. A hybrid heuristic approach based on DBSCAN and agglomerative clustering

In this heuristic approach, we propose to implement density-based spatial clustering applications with noise (DBSCAN) (Ester et al., 1996) to form micro-grid and nano-grid clusters. DBSCAN is one of the most extensively studied clustering approaches in the literature to group closely situated points in a given region. The algorithm has been frequently used due to several distinguishing characteristics making the DBSCAN an attractive way for spatial clustering. Firstly, the algorithm does not require a predetermined value for the number of clusters, unlike the other well-known clustering approaches such as the k-means algorithm. Secondly, the algorithm can detect the remote points (outliers) to be individually clustered.

The algorithm incorporates two parameters, $minPts$ and ϵ, that could affect the algorithm's performance. In the algorithm, $minPts$ denotes the minimum number of points required in the $\epsilon-$neighborhood of a node to define it as a core point. On the other hand, the neighborhood radius is denoted by the parameter ϵ. Each cluster consists of the connected core points and the non-core (border) points covered by the core nodes in the cluster. The nodes having zero points in the $\epsilon-$neighborhood are considered outliers, as shown in Fig. 7. The steps of the DBSCAN algorithm are also provided in Fig. 8.

In the algorithm, we used $\epsilon = distLim$ to determine the nodes in the $\epsilon-$neighborhood. The choice of $MinPts$, on the other hand, is not as trivial as the choice of the ϵ parameter. For our problem, we chose $MinPts = 2$ intuitively, as the mini-grid requires at least two

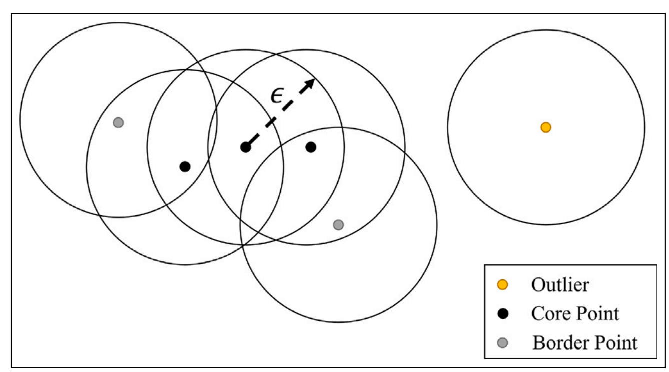

Fig. 7. Classification of the points in DBSCAN algorithm ($minPts = 3$).

consumers. We also performed a sensitivity analysis to determine the ideal value for the $MinPts$ parameter. The sensitivity analysis on the dense and dispersed data sets has shown that the algorithm's ability to form micro-grid clusters decreases as we enforce higher values on $MinPts$. Therefore, the algorithm yields an expensive investment cost as the micro-grid option is used infrequently.

The flowchart of the hybrid DBSCAN-agglomerative clustering method is provided in Fig. 9. This hybrid solution approach implements the DBSCAN algorithm to identify the nano-grids and form the initial micro-grid clusters. We also calculate the radius of the smallest circle surrounding the demand locations for each cluster serving more than one consumer. Although we specify the ϵ parameter as the coverage threshold, the distance between the demand points and the cluster center may not necessarily be within the distance limit. If the radius of the circle does not violate the distance limit constraint (i.e, there is a feasible solution satisfying $distLim$, we locate a facility to the geometric median by using the modified version of Weiszfeld's algorithm and project the suggested location by using the projection method provided in Section 4.1.3. Otherwise, we perform the agglomerative

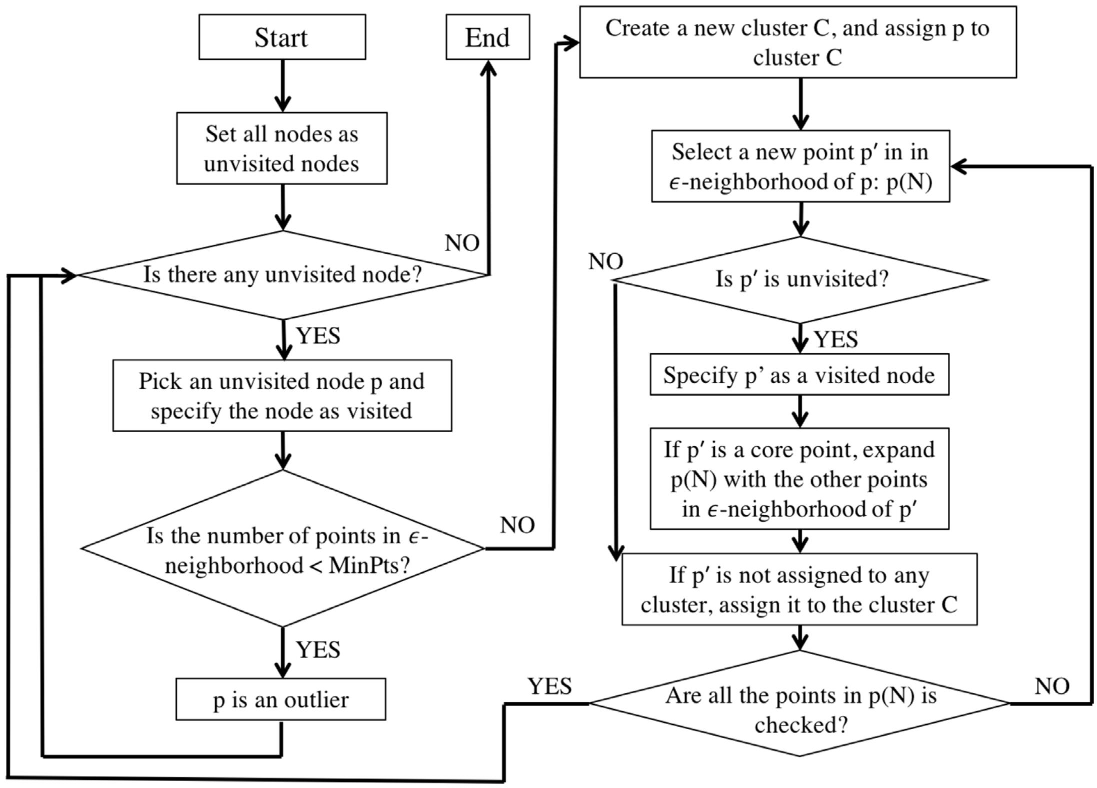

Fig. 8. Flowchart of DBSCAN Clustering.

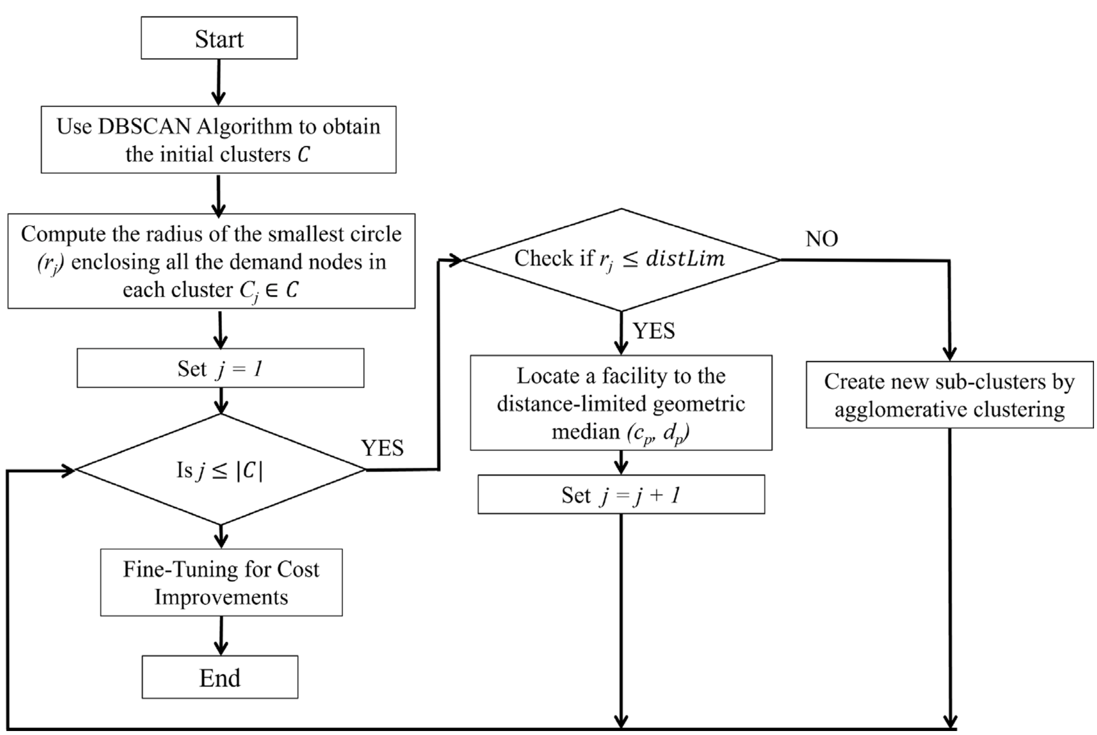

Fig. 9. Flowchart of the Hybrid DBSCAN-Agglomerative Clustering Method.

hierarchical clustering algorithm for the same cluster, assuming that each node within the cluster is a singleton cluster, and create new sub-clusters conforming to the distance threshold. After obtaining the final configuration, we apply the fine-tuning step for cost improvements as in Section 4.2.1. In this step, we detach the households from the expensive microgrids having an insufficient number of households (i.e., if total investment cost is cheaper when nano-grids electrify all consumers in the cluster) and allocate the singleton clusters to the closest micro-grid within the coverage threshold if it improves the total infrastructure cost.

4.2.3. A hybrid heuristic approach based on DBSCAN and K-means algorithms

This heuristic method follows a similar approach to the hybrid methodology proposed in the previous section. We implement the DBSCAN algorithm using the same system parameters to create initial nano-grid and micro-grid clusters. Then, we find the minimum circle covering all the demand nodes in the cluster for each micro-grid. If the radius of the minimum circle exceeds the distance threshold, the cluster is iteratively divided into two sub-clusters using the k-means clustering technique until each sub-group conforms to the distance limit constraint. For each cluster satisfying the distance limit, we locate the facility to the distance-limited geometric median and apply the same fine-tuning steps at the end of the algorithm. The major steps of this hybrid approach are summarized in Fig. 10.

K-means algorithm is one of the most well-known partitioning techniques in the literature. The algorithm is considered a practical way of partitioning as it provides the final solution with high computational efficiency and requires only a single parameter k (the number of clusters). The k-means clustering technique consists of the location and allocation steps as in Cooper's iterative heuristic. First, the algorithm picks k random locations and forms k clusters after assigning the points to the nearest location. Then, it computes the centroid of each cluster and reassigns the demand points to the closest center of mass. The algorithm repeatedly recalculates the centroids and performs the reallocation step until no further change is observed in two consecutive steps. However, it has been stated that the final configuration is susceptible to the starting clusters, and trying different initial assignments would improve the performance of the algorithm (Cabrera-Celi et al., 2017). Therefore, we repeat k-means clustering 100 times to ensure a higher clustering accuracy.

5. Computational results

In this section, we analyze the performance of the solution methodologies by conducting experiments with data sets from sub-Saharan Africa. Our data sets are obtained from Millennium Villages sites located in Sub-Saharan Africa. Millennium Villages Project aimed to achieve rural development towards the Millennium Development Goals of the United Nations, addressing issues such as poverty, hunger, disease, and gender inequality (Sanchez et al., 2007; Zvoleff et al., 2009). In this study, we use three sites with different settlement patterns to test our solution methodologies using realistic cost parameters. These sites (Tiby, Mali - Mbola, Tanzania - Potou, Senegal) consist of 1545, 1168, and 1781 households, respectively. These data sets are obtained by Gokbayrak (2022) using the open buildings data set of Google,[1] and can be accessed via the following link: https://osf.io/k9des. As the mathematical modeling-based approaches could be computationally inefficient for such large instances, we selected two 2 × 2 km^2 sample sites[2] with different densities from the villages as shown in Figs. 11–13 so that we can compare the model-based approaches with the clustering-based methods.

We assume that households are identical and consume the same amount of electricity for all instances. We consider realistic cost parameters and distance limits for low-voltage cable connections. The cost of low-voltage cables between the generation points and the consumers (c_L) is set to \$4/m, consistent with the parameters used in Energy Data Platform (2019). The maximum allowable length of these single-phase cables can take values from the set $distLim = \{50, 100, 200\}$ as in Papathanassiou et al. (2005), Stephen et al. (2013), Short (2003). The nano-grid generation capacity is considered 100 W, as the capacity of solar home systems generally ranges between 20 W and 100 W in the rural areas of Sub-Saharan Africa (IRENA, 2016). Hence, the cost of deploying nano-grid is selected from the set $F2 = \{750, 1000\}$ based on the stand-alone PV costs provided in Energy Data Platform (2019), IRENA (2016, 2018). Since solar systems with higher scales can benefit from the economies of scale, we assume that the cost of micro-grid facilities is either \$3000 or \$4000 in our experiments for the fixed cost problem as suggested in Energy Data Platform (2019), Korkovelos et al. (2019), World Bank (2019), ESMAP (2019) ($F1 = \{3000, 4000\}$). The cost of stand-alone PV systems with capacities less than 1 kW ranges from \$4/W to \$16/W across Sub-Saharan Africa. However, the cost reduces to \$2/W to \$8/W for the solar systems with more than 1 kWp installed capacities. Therefore, the additional charge of connecting households to micro-grids is selected from the set $F3 = \{500, 600\}$ in variable cost and fixed & variable cost problems. For fixed & variable cost problems, we set the fixed facility costs to \$1000 and \$2000 in our experiments so that the micro-grids connecting two demand points are still economically viable ($F1 = \{1000, 2000\}$).

The computational experiments are conducted on a dual 2.4 GHz Intel XeonE5-2630 v3 CPU server with 64 GB RAM. The heuristic methods are implemented in Matlab 2020a, and the optimization methods are solved using CPLEX 12.10. We enforce a CPU time limit of three hours for all optimization models. The convergence threshold ε in Cooper's iterative algorithm is 0.1 for the multi-stage heuristic approaches.

5.1. Household coverage index (HCI)

In order to measure the dispersion of households and classify the spatial distribution of the sample populations, we propose a new indicator metric called Household Coverage Index (HCI). In accordance with our problem definition, this metric represents the average number of demand nodes within the $distLim$ distance from each demand point. In Table 2, the regions with a smaller HCI are considered dispersed settlements, whereas higher HCI refers to densely populated regions. Therefore, all sample sites from Tiby (T-1 and T-2) and M-2 from Mbola are considered densely populated areas, while M-1 from Mbola, P-1 and P-2 from Potou have dispersed settlement structures. The performances of the best performing model-based and clustering-based approaches are compared in terms of running time and objective value for each sample site in Table 3, Tables 4 and 5 for the fixed, variable, and fixed&variable cost problems, respectively. When there is more than one approach listed under the best-performing method columns, it means that more than one method was able to find the best solution. However, in that case, the solution time of the fastest method is listed under the time column and the method is highlighted in bold. In the last columns of these tables, we present the percentage difference between the model-based and the clustering-based approaches. Notice that when this percentage is negative, the clustering-based algorithms report better solutions than the model-based ones. We also provide detailed tables where we compare the model-based and clustering-based methods among themselves in Appendix.

The computational results indicate that the solution time of the model-based heuristic methods for densely populated areas is significantly longer than that of the sites with dispersed settlements. While our model-based methods could obtain solutions in seconds in sparsely populated sites, in densely populated areas, they exceed the CPU time limit of three hours, especially when the distance limit is enforced to be greater than 50 m. As the number of candidate facility locations increases, we observe longer computational time in accordance with the

[1] Can be downloaded from https://sites.research.google/open-buildings/.

[2] The smaller sample sites can be found at https://osf.io/39a78/

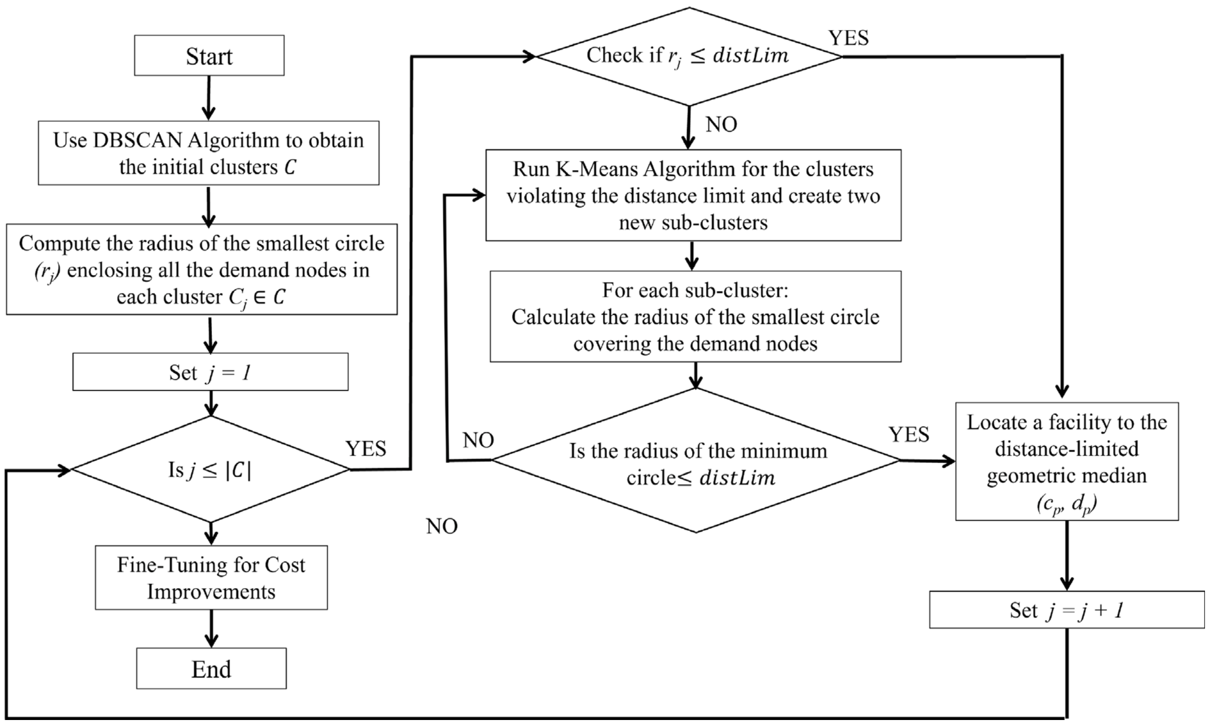

Fig. 10. Flowchart of DBSCAN — K-Means Clustering-based Heuristic Approach.

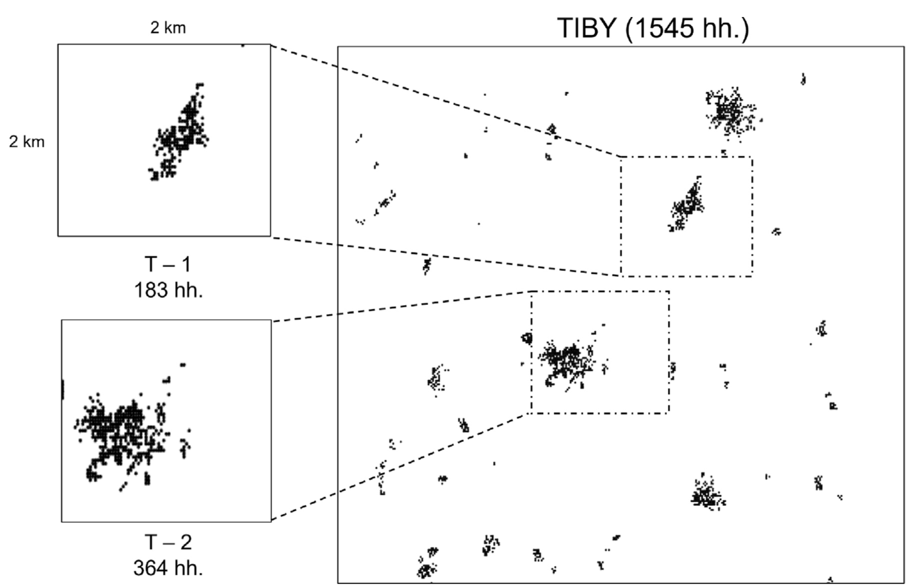

Fig. 11. Tiby.

increasing computational complexity. MB-I is observed to be the most computationally intensive method among the alternatives, especially in dense areas. Contrary to the model-based approaches, clustering-based heuristics yield the final output within a significantly shorter computational time regardless of the spatial distribution of the demand points.

5.2. Analysis of the results based on the problem types

5.2.1. Results for the problem with fixed facility costs

Under the fixed cost problem formulation, the MB-I approach is observed to outperform the other model-based heuristics regardless of the settlement pattern for the small distance limits. For larger distance limits, on the other hand, MB-II-III-IV provide competitive results as shown in Table 3. In dispersed areas, MB-I, MB-III, and MB-IV provide a better solution than MB-II.

Regarding the performance of the clustering-based approaches, it is observed that the majority of the least-cost solutions can be obtained by using the agglomeration-based heuristic algorithms (Agg, DB-Agg) for the fixed cost problem. However, DBSCAN-Kmeans (DB-KM) surpasses the other alternatives when we impose smaller distance limits. For the heuristics, including an agglomeration step, the new dissimilarity measure, GeomDiff, works successfully when the distance limit is less than 100 m. GeomDiff also performs better than the other 2 dissimilarity measures for larger distance limits if the sample has dispersed

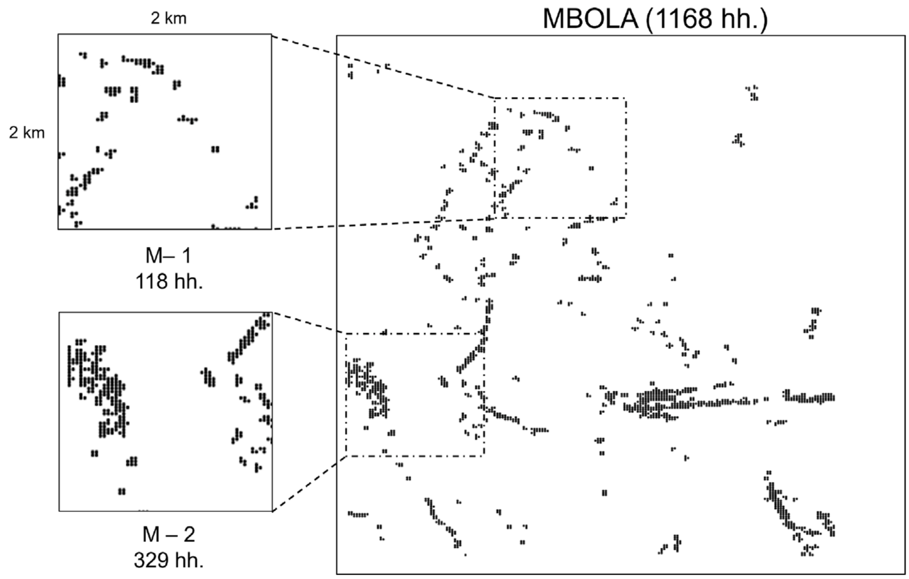

Fig. 12. Mbola.

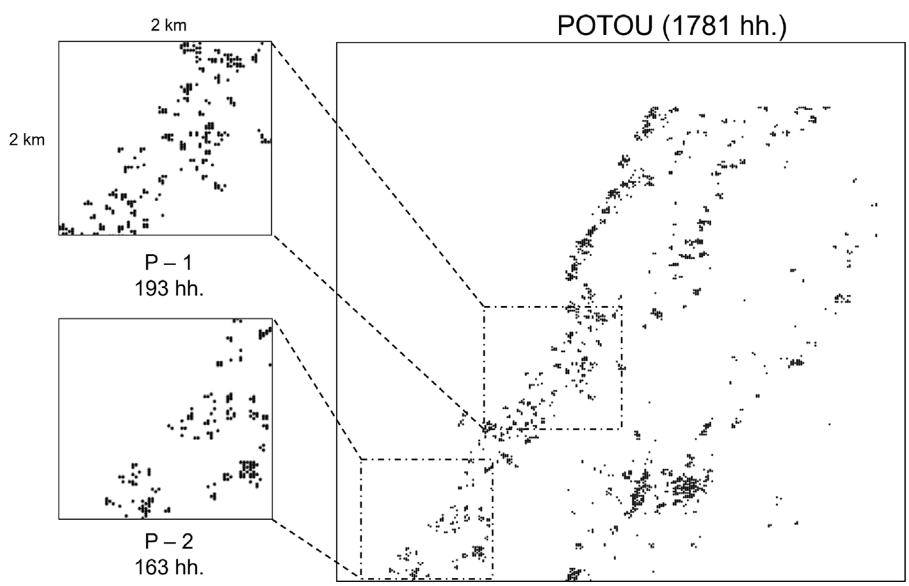

Fig. 13. Potou.

Table 2
Household Coverage Index.

	Distance limit	HCI		Distance limit	HCI		Distance limit	HCI
M-1	50	3.4	T-1	50	3.8	P-1	50	2.5
	100	6.2		100	11.6		100	5.3
	200	11.5		200	37.2		200	13.1
M-2	50	4.4	T-2	50	4.5	P-2	50	2.4
	100	11.6		100	14.2		100	5.3
	200	34.9		200	51.4		200	13.7

settlement patterns. On the other hand, Ward is observed to provide the best configuration in dense areas under 200 distance limit.

Although the model-based heuristic approaches dominate the clustering-based algorithms in most instances based on cost, the results demonstrate that the clustering algorithms can provide approximate solutions within a few CPU seconds. While the model-based methods can provide a feasible solution for dispersed areas within the time limit, the solution time more likely exceeds the given limit for the densely populated samples. In this case, clustering-based approaches can provide approximate solutions to the model-based methods, and

Table 3
Fixed cost problem results.

	Dist Lim	F1	F2	Comparison of multi-stages			Comparison of clustering techniques			Diff.
				Cost	Time	Method	Cost	Time	Method	
M-1	50	3000	750	76 451	2.9	MB-I-**II**-III-IV	80 618	0.03	DB-Agg	5.2%
	50	3000	1000	85 499	1.4	MB-I-**II**-III-IV	88 500	0.02	Agg	3.4%
	50	4000	750	84 408	15.9	MB-I-IV	87 175	0.3	DB-KM	3.2%
	50	4000	1000	98 701	2.4	MB-I-**II**-III-IV	105 368	0.02	DB-Agg	6.3%
	100	3000	750	66 505	10.5	MB-I-IV	66 876	0.03	Agg	0.6%
	100	3000	1000	70 505	2.6	MB-I-**III**-IV	72 577	0.19	DB-KM	2.9%
	100	4000	750	75 625	100.3	MB-I	71 291	0.03	Agg	−6.1%
	100	4000	1000	81 505	9.1	MB-I-**III**-IV	82 876	0.03	Agg	1.7%
	200	3000	750	66 071	890.1	MB-I	66 270	0.19	Agg	0.3%
	200	3000	1000	69 423	427.3	MB-I	69 423	0.06	DB-Agg	0.0%
	200	4000	750	73 815	29.2	MB-I-II-III-IV-V	69 423	0.19	Agg	−6.3%
	200	4000	1000	78 620	13.7	MB-I-**II**-III-IV	78 929	0.05	DB-Agg	0.4%
M-2	50	3000	750	199 108	1024.7	MB-I	220 429	7.77	DB-KM	9.7%
	50	3000	1000	211 236	10885.0	MB-I	243 190	0.05	DB-Agg	13.1%
	50	4000	750	228 840	195.5	MB-I	242 754	8.12	DB-KM	5.7%
	50	4000	1000	254 608	4251.9	MB-I	285 009	0.05	DB-Agg	10.7%
	100	3000	750	154 333	2866.6	MB-**III**-IV	163 176	0.27	DB-Agg	5.4%
	100	3000	1000	155 958	543.8	MB-**III**-IV	164 917	0.12	DB-Agg	5.4%
	100	4000	750	176 358	10816.3	MB-**III**-IV	165 384	0.16	Agg	−6.6%
	100	4000	1000	182 119	769.6	MB-**II**-III	197 246	0.22	Agg	7.7%
	200	3000	750	149 396	9633.7	MB-IV	150 545	1.05	Agg	0.8%
	200	3000	1000	150 125	5339.6	MB-IV	151 269	0.22	DB-Agg	0.8%
	200	4000	750	167 999	10822.4	MB-II	151 269	0.42	Agg	−11.1%
	200	4000	1000	170 178	10819.1	MB-II	171 830	0.48	Agg	1.0%
T-1	50	3000	750	112 100	1049.2	MB-I	127 690	6.52	DB-KM	12.2%
	50	3000	1000	120 769	10840.1	MB-I	137 250	0.08	Agg	12.0%
	50	4000	750	127 423	33.3	MB-I	136 913	7.05	DB-KM	6.9%
	50	4000	1000	143 699	8557.6	MB-I	167 153	7.23	DB-KM	14.0%
	100	3000	750	80 513	705.8	MB-II	86 482	0.03	Agg	6.9%
	100	3000	1000	80 591	301.3	MB-II	86 482	0.11	DB-Agg	6.8%
	100	4000	750	91 513	1692.5	MB-II	86 482	0.14	Agg	−5.8%
	100	4000	1000	92 591	620.8	MB-II	101 982	0.03	Agg	9.2%
	200	3000	750	77 846	10810.5	MB-II	79 632	0.17	Agg	2.2%
	200	3000	1000	77 852	3903.4	MB-III	79 632	0.17	DB-Agg	2.2%
	200	4000	750	86 562	10809.1	MB-III	79 632	0.16	Agg	−8.7%
	200	4000	1000	86 434	10810.6	MB-II	89 077	0.12	Agg	3.0%
T-2	50	3000	750	219 961	10899.9	MB-I	253 062	13.66	DB-KM	13.1%
	50	3000	1000	236 865	10885.9	MB-I	273 000	0.11	Agg	13.2%
	50	4000	750	252 482	645.5	MB-I	270 403	13.53	DB-KM	6.6%
	50	4000	1000	281 032	10872.3	MB-I	330 242	13.67	DB-KM	14.9%
	100	3000	750	165 063	10818.9	MB-II	175 241	6.03	DB-KM	5.8%
	100	3000	1000	167 931	10821.4	MB-III	179 011	6.00	DB-KM	6.2%
	100	4000	750	187 723	10825.9	MB-IV	185 738	0.20	Agg	−1.1%
	100	4000	1000	193 182	10818.2	MB-II	210 853	6.61	DB-KM	8.4%
	200	3000	750	160 904	10 844.3	MB-III	163 254	0.69	Agg	1.4%
	200	3000	1000	161 850	10837.1	MB-V	163 994	0.55	DB-Agg	1.3%
	200	4000	750	177 289	10838.3	MB-V	164 293	0.55	Agg	−7.9%
	200	4000	1000	180 276	10846.1	MB-III	183 370	0.72	DB-Agg	1.7%
P-1	50	3000	750	135 409	21.0	MB-**III**-IV	137 536	0.45	DB-KM	1.5%
	50	3000	1000	155 357	31.8	MB-I	144 750	0.03	Agg	−7.3%
	50	4000	750	143 106	10.6	MB-I-II-III-IV-V	143 805	0.44	DB-KM	0.5%
	50	4000	1000	176 409	28.7	MB-I-**III**-IV	182 272	0.70	DB-KM	3.2%
	100	3000	750	117 626	357.4	MB-I	119 474	0.02	Agg	1.5%
	100	3000	1000	125 267	14.0	MB-**II**-III-IV	129 483	0.02	DB-Agg	3.3%
	100	4000	750	131 999	806.0	MB-I	129 220	0.02	Agg	−2.2%
	100	4000	1000	146 259	362.8	MB-I	149 724	0.02	Agg	2.3%
	200	3000	750	112 790	74.1	MB-II-III-IV-V	113 417	0.06	Agg	0.6%
	200	3000	1000	115 135	46.4	MB-V	116 716	0.08	DB-Agg	1.4%
	200	4000	750	127 407	242.3	MB-II-III-V	116 513	0.05	Agg	−9.3%
	200	4000	1000	133 284	196.3	MB-III-IV	134 915	1.33	DB-KM	1.2%

(*continued on next page*)

the difference between the results reduces as larger distance limits are imposed. Similarly, the clustering-based techniques are shown to perform better when the difference between micro-grid and nano-grid costs is high. In fact, these methods could provide up to 11% better results than model-based heuristics in some instances.

Table 3 (*continued*).

	Dist Lim	F1	F2	Comparison of multi-stages			Comparison of clustering techniques			Diff.
				Cost	Time	Method	Cost	Time	Method	
P-2	50	3000	750	115 167	17.6	MB-I-II-**III**-IV	118 231	1.14	DB-KM	2.6%
	50	3000	1000	134 443	9.1	MB-I-**II**-III-IV	122 250	0.02	Agg	−10.0%
	50	4000	750	121 084	11.7	MB-I-**II**-III-IV-V	122 250	1.16	DB-KM	1.0%
	50	4000	1000	150 461	11.4	MB-I-**II**-III-IV	155 743	1.17	DB-KM	3.4%
	100	3000	750	101 383	278.3	MB-I	102 881	0.58	DB-KM	1.5%
	100	3000	1000	108 502	24.7	MB-III	113 105	0.06	DB-Agg	4.1%
	100	4000	750	110 673	252.1	MB-I	112 574	0.02	Agg	1.7%
	100	4000	1000	125 474	37.5	MB-**III**-IV	130 884	0.05	Agg	4.1%
	200	3000	750	96 751	63.6	MB-**II**-III-IV-V	99 194	0.03	Agg	2.5%
	200	3000	1000	99 589	56.5	MB-**III**-IV	100 513	0.16	DB-Agg	0.9%
	200	4000	750	106 916	114.4	MB-II-III-IV-V	100 513	0.05	Agg	−6.4%
	200	4000	1000	113 811	100.1	MB-**II**-III-IV-V	115 010	0.08	DB-Agg	1.0%

5.2.2. Results for the problem with variable facility costs

Under the variable cost problem structure, the solution times reduce significantly for both clustering-based and model-based heuristics, as shown in Table 4. The model-based heuristics are able to provide a solution within the 3 h time limit, and the maximum solution time for the clustering-based methods is 0.9 CPU seconds. In densely populated areas (T-1, T-2 and M-2), MB-I outperforms the other heuristics with a slightly high computational time in 11 out of 12 instances with a 50 m distance limit. However, when the distance limit is larger, there is no clear superiority between the model-based methods.

The clustering-based approaches can provide very close and, in some cases, better solutions than the model-based approaches within a second. The solutions deviate from the model-based ones by less than 5% except for one instance, and the deviation is between 0.7%–3.1% in densely populated areas. The highest difference is observed in dispersed samples. This result can be associated with the fact that the algorithms prefer the nano-grid option more frequently in dispersed settlements as the distance limit constraint restricts micro-grid connections. The results also show that the Agglomerative Clustering approach dominates the other clustering methods in most instances, and Ward's variance method is also a prominent dissimilarity measure for this problem type.

5.2.3. Results for the problem with fixed & variable facility costs

Table 5 summarizes the performances of modeling and clustering-based approaches for the fixed & variable cost problem. Similar to the variable cost problem, MB-I is the best-performing approach for densely populated areas under a 50 m distance limit. In the rest of the instances, the results demonstrate that all methods are equally likely to provide a cost-efficient solution since they yield similar configurations. Since we have two different cost components for micro-grids in this problem, the resulting configuration is more likely to include a higher number of nano-grids, which is the key driver behind the similarity of outcomes. The heuristics with the agglomeration steps are prominent in the fixed & variable cost problem. In the agglomeration steps, the GeomDiff measure can attain the best result in the majority of instances regardless of the settlement pattern, and the second best-performing measure is the Complete measure.

5.3. Generalization of the results

The results for all problem types demonstrate that the model-based heuristics achieve most of the least-cost solutions. Although the cost difference between the model-based and clustering-based methods generally ranges between 0% to 5%, this percentage is observed to increase in densely populated regions. In the regions with the dispersed settlement pattern, the cost difference reduces below 0% for some instances, meaning that the clustering-based approach outperforms the model-based heuristics. It should also be highlighted that GeomDiff, the new dissimilarity we introduced in this study, gives promising solutions, especially for the smaller distance limits. The computational experiments indicate that GeomDiff can attain the least cost configuration in most instances in the fixed cost and fixed & variable cost problem settings. For the variable cost problems, we demonstrate that the cost difference between the model-based and clustering-based approaches is generally less than 4%, even in the densely populated sample sites. Another observation is that the agglomerative clustering method dominates the other two clustering-based approaches in most instances. Ward's variance method is the prominent measure for densely populated settlements in the agglomeration steps.

We observe that the performances of the model-based and clustering-based methods differ slightly in most instances. Two main reasons could explain the reduction in the cost differences. Firstly, the discrete model in the model-based approach may not obtain the optimal solution within the specified CPU time limit for the dense sample sites. Hence, as the optimal solution cannot be attained in three hours, the cost difference between the clustering algorithms and the optimization model-based heuristic decreases. Secondly, creating micro-grid clusters that can cover an ideal number of households within the distance limit in the fixed & variable cost formulation is more challenging. Given that the investment cost of deploying a micro-grid facility is composed of the fixed facility cost, low-voltage connections, and a variable cost component, micro-grid clusters are required to cover a higher number of households in order for the nano-grid option to be discarded. However, only a minority of the potential micro-grid clusters can reach a sufficient number of households under the given distance thresholds. Accordingly, we observe that the nano-grid option becomes more prevalent as micro-grids are relatively expensive compared to the fixed-cost formulation.

The aim of Stage-1 and Stage-2 in the model-based approaches is to provide good initial points for the modified Cooper's iterative algorithm in Stage-3. In Tables 6 and 7, we summarize the performances of the augmentation methods on the results of Stage-2 (i.e., the solution of the discrete models) and Stage-3 (the final solution). We observe that a better solution of the discrete model in Stage-2 is not necessarily a better initial solution for Stage-3. For all problem types, however, it is possible to say that the centroid method (MB-I) is observed to outperform other augmentation methods in terms of cost, including the MB-II, especially in dense samples under smaller distance limits. Table 7 shows that MB-I is the best-performing model-based method on all problem types, and the method can provide the least-cost solution for at least half of the instances. This method is also observed to be most effective when the problem includes both the fixed cost and variable cost components. Moreover, common intersection points (in MB-III) and the centroid of the convex hull of the common intersection points (in MB-IV) are shown to improve the results, especially for the fixed cost problem. Table 8 summarizes the performances of the clustering-based methods by problem type. Although the agglomerative clustering approach is observed to perform better, especially under the variable

Table 4
Variable cost problem results.

	Dist Lim	F2	F3	Comparison of multi-stages			Comparison of clustering techniques			Diff.
				Cost	Time	Method	Cost	Time	Method	
M-1	50	750	500	67 669	6.6	MB-I-II	68 422	0.03	Agg	1.1%
	50	750	600	79 369	4.5	MB-I-II	79 722	0.16	Agg	0.4%
	50	1000	500	67 919	4.4	MB-I-II	68 836	0.05	Agg	1.3%
	50	1000	600	79 619	4.9	MB-I-II	80 536	0.05	Agg	1.1%
	100	750	500	67 669	5.2	MB-III	67 860	0.11	DB-KM	0.3%
	100	750	600	79 369	8.6	MB-I-II-IV-V	72 577	0.19	DB-KM	−9.4%
	100	1000	500	67 919	4.4	MB-I-**II**-III	68 836	0.09	Agg	1.3%
	100	1000	600	79 619	4.5	MB-I-**II**-III	80 536	0.05	Agg	1.1%
	200	750	500	67 669	7.6	MB-IV	68 422	0.14	Agg	1.1%
	200	750	600	79 369	9.7	MB-II-V	72 498	0.19	DB-KM	−9.5%
	200	1000	500	67 919	8.7	MB-I-III-IV	68 836	0.12	Agg	1.3%
	200	1000	600	79 619	8.6	MB-I-III-IV	80 025	0.22	DB-KM	0.5%
M-2	50	750	500	186 204	74.1	MB-I	187 461	0.33	Agg	0.7%
	50	750	600	219 104	72.1	MB-I	219 961	0.25	Agg	0.4%
	50	1000	500	186 204	70.4	MB-I	188 461	0.44	Agg	1.2%
	50	1000	600	219 112	27.1	MB-I-II-III-V	220 961	0.33	Agg	0.8%
	100	750	500	186 204	36.9	MB-I-III-IV-V	187 461	0.58	Agg	0.7%
	100	750	600	219 104	27.8	MB-I-**II**-III-IV	219 961	0.48	Agg	0.4%
	100	1000	500	186 204	33.7	MB-I-III-IV-V	188 461	0.56	Agg	1.2%
	100	1000	600	219 104	33.2	MB-I-III-IV-V	220 961	0.64	Agg	0.8%
	200	750	500	186 204	44.8	MB-I-IV-V	187 461	0.73	Agg	0.7%
	200	750	600	219 104	35.3	MB-I-II-III-V	219 961	0.56	Agg	0.4%
	200	1000	500	186 204	36.0	MB-I-II-V	188 461	0.53	Agg	1.2%
	200	1000	600	219 104	36.4	MB-I-II-V	220 961	0.66	Agg	0.8%
T-1	50	750	500	104 769	74.7	MB-I	106 619	0.12	Agg	1.7%
	50	750	600	122 829	52.3	MB-I	123 717	0.14	Agg	0.7%
	50	1000	500	104 789	26.6	MB-I	108 119	0.09	Agg	3.1%
	50	1000	600	123 097	43.0	MB-I	125 819	0.14	Agg	2.2%
	100	750	500	104 769	145.1	MB-I	106 619	0.27	Agg	1.7%
	100	750	600	122 829	62.8	MB-I	123 717	0.27	Agg	0.7%
	100	1000	500	104 789	39.3	MB-**II**-IV	108 119	0.25	Agg	3.1%
	100	1000	600	123 089	44.3	MB-I-**II**-IV	125 819	0.06	Agg	2.2%
	200	750	500	104 769	28.5	MB-V	106 619	0.16	Agg	1.7%
	200	750	600	122 838	36.8	MB-I-II-III-IV-V	123 717	0.19	Agg	0.7%
	200	1000	500	104 789	92.1	MB-I-II-V	108 119	0.19	Agg	3.1%
	200	1000	600	123 089	67.8	MB-II	125 819	0.38	Agg	2.2%
T-2	50	750	500	208 102	168.1	MB-I	211 043	0.14	Agg	1.4%
	50	750	600	243 740	51.8	MB-II-IV-V	245 343	0.27	Agg	0.7%
	50	1000	500	209 167	174.8	MB-I	214 138	0.41	Agg	2.3%
	50	1000	600	245 159	203.4	MB-I	249 438	0.39	Agg	1.7%
	100	750	500	208 094	59.2	MB-V	213 011	0.72	Agg	2.3%
	100	750	600	243 740	62.3	MB-I-III	246 543	0.83	Agg	1.1%
	100	1000	500	208 683	10963.5	MB-I	216 603	0.89	Agg	3.7%
	100	1000	600	244 990	83.1	MB-II-III	251 603	0.5	Agg	2.6%
	200	750	500	208 094	101.5	MB-I-II-IV-V	211 302	0.72	Agg	1.5%
	200	750	600	243 740	93.7	MB-I-II-V	245 457	0.86	Agg	0.7%
	200	1000	500	208 652	138.6	MB-**III**-IV	215 052	0.86	Agg	3.0%
	200	1000	600	244 991	152.5	MB-IV-V	249 952	1.06	Agg	2.0%
P-1	50	750	500	112 461	16.0	MB-I	114 466	0.5	Agg	1.8%
	50	750	600	130 971	19.8	MB-I	131 802	0.05	Agg	0.6%
	50	1000	500	113 420	16.1	MB-I	118 966	0.03	Agg	4.7%
	50	1000	600	132 397	15.3	MB-I	136 466	0.06	Agg	3.0%
	100	750	500	112 452	9.5	MB-**II**-IV	114 466	0.09	Agg	1.8%
	100	750	600	130 962	11.2	MB-I-II-III-V	131 802	0.19	Agg	0.6%
	100	1000	500	113 135	11.0	MB-IV	118 966	0.16	Agg	4.9%
	100	1000	600	132 261	10.4	MB-I-IV	136 466	0.05	Agg	3.1%
	200	750	500	112 452	7.1	MB-V	114 466	0.11	Agg	1.8%
	200	750	600	130 962	8.3	MB-**II**-III	131 802	0.11	Agg	0.6%
	200	1000	500	113 144	81.8	MB-I	118 966	0.11	Agg	4.9%
	200	1000	600	132 269	9.4	MB-III-V	136 466	0.08	Agg	3.1%

(*continued on next page*)

cost problem setting, introducing a pre-clustering step with DBSCAN is shown to perform effectively for the fixed & variable cost problems.

Note that MB-V has a random selection process while determining candidate facility locations for the discrete models. The results we reported are based on a single run because we aim to compare the performances of the model-based heuristics by solving the discrete mathematical models just once. We also repeated the random process 10 and 100 times for each instance to investigate the effect of the number of repetitions on the solution quality. The objective values

Table 4 (*continued*).

	Dist Lim	F2	F3	Comparison of multi-stages			Comparison of clustering techniques			Diff.
				Cost	Time	Method	Cost	Time	Method	
P-2	50	750	500	96 328	13.4	MB-V	98 603	0.03	Agg	2.3%
	50	750	600	111 636	52.3	MB-V	112 798	0.03	Agg	1.0%
	50	1000	500	97 046	12.8	MB-V	101 901	0.03	Agg	4.8%
	50	1000	600	113 146	11.9	MB-V	116 961	0.02	Agg	3.3%
	100	750	500	96 302	77.4	MB-V	98 603	0.03	Agg	2.3%
	100	750	600	109 016	217.1	MB-I	112 798	0.03	Agg	3.4%
	100	1000	500	96 588	12.1	MB-IV	101 892	0.05	Agg	5.2%
	100	1000	600	112 891	15.7	MB-IV	116 955	0.03	Agg	3.5%
	200	750	500	96 302	15.3	MB-II-**III**-IV-V	98 470	0.06	DB-Agg	2.2%
	200	750	600	99 601	10829.1	MB-I	112 798	0.06	Agg	11.7%
	200	1000	500	96 600	17.9	MB-**III**-V	101 447	0.19	DB-Agg	4.8%
	200	1000	600	112 900	15.0	MB-**II**-III	116 654	0.14	DB-Agg	3.2%

Table 5
Fixed & variable cost problem results.

	Dist Lim	F1	F2	F3	Comparison of multi-stages			Comparison of clustering techniques			Diff.
					Cost	Time	Method	Cost	Time	Method	
M-1	50	1000	750	500	88 486	3.7	MB-I-II-III-IV-**V**	88 500	0.28	Agg	0.0%
	50	1000	750	600	88 500	2.8	MB-I-II-III-IV-**V**	88 500	0.14	Agg	0.0%
	50	1000	1000	500	97 680	2.7	MB-I-**II**-III-IV	99 325	0.06	DB-Agg	1.7%
	50	1000	1000	600	108 499	3.2	MB-I-**II**-III-IV	110 521	0.28	Agg	1.8%
	50	2000	750	500	88 500	3.5	MB-I-II-III-IV-**V**	88 500	0.09	Agg	0.0%
	50	2000	750	600	88 500	3.8	MB-I-II-III-IV-**V**	88 500	0.05	Agg	0.0%
	50	2000	1000	500	112 888	5.4	MB-I-**II**-III-IV	114 952	0.09	DB-Agg	1.8%
	50	2000	1000	600	117 376	11.1	MB-I-IV	117 855	0.97	DB-KM	0.4%
	100	1000	750	500	88 486	11.7	MB-I-**II**-III-IV-V	88 500	0.06	Agg	0.0%
	100	1000	750	600	88 500	4.0	MB-I-II-III-IV-**V**	88 500	0.05	Agg	0.0%
	100	1000	1000	500	95 867	6.1	MB-I-IV	95 867	0.09	Agg	0.0%
	100	1000	1000	600	107 245	5.5	MB-I-II-III-IV-**V**	107 245	0.09	Agg	0.0%
	100	2000	750	500	88 500	2.4	MB-I-II-III-IV-**V**	88 500	0.09	Agg	0.0%
	100	2000	750	600	88 500	2.1	MB-I-II-III-IV-**V**	88 500	0.05	Agg	0.0%
	100	2000	1000	500	109 179	41.2	MB-I	109 179	0.09	Agg	0.0%
	100	2000	1000	600	116 175	7.1	MB-II-III-IV-V	116 258	0.12	DB-Agg	0.1%
	200	1000	750	500	88 486	16.4	MB-I-II-III-IV-**V**	88 500	0.09	Agg	0.0%
	200	1000	750	600	88 500	1.9	MB-I-II-III-IV-**V**	88 500	0.12	Agg	0.0%
	200	1000	1000	500	95 867	113.4	MB-I	95 867	0.14	Agg	0.0%
	200	1000	1000	600	107 245	12.6	MB-I-II-III-IV-**V**	107 245	0.12	Agg	0.0%
	200	2000	750	500	88 500	4.9	MB-I-II-III-IV-**V**	88 500	0.14	Agg	0.0%
	200	2000	750	600	88 500	2.3	MB-I-II-III-IV-**V**	88 500	0.09	Agg	0.0%
	200	2000	1000	500	109 160	17.6	MB-I-II-III-IV-**V**	109 160	0.14	Agg	0.0%
	200	2000	1000	600	116 175	32.2	MB-II-III-IV-V	116 925	0.14	DB-Agg	0.6%
M-2	50	1000	750	500	246 708	18.4	MB-I-II-III-IV-**V**	246 750	0.83	Agg	0.0%
	50	1000	750	600	246 750	15.6	MB-I-II-III-IV-**V**	246 750	0.97	Agg	0.0%
	50	1000	1000	500	263 366	165.9	MB-I	271 482	0.08	DB-Agg	3.0%
	50	1000	1000	600	294 842	385.7	MB-I	302 990	36.00	DB-Agg	2.7%
	50	2000	750	500	246 750	15.5	MB-I-II-III-IV-**V**	246 750	0.80	Agg	0.0%
	50	2000	750	600	246 750	15.9	MB-I-II-III-IV-**V**	246 750	0.78	Agg	0.0%
	50	2000	1000	500	307 802	286.4	MB-I	319 099	16.20	DB-KM	3.5%
	50	2000	1000	600	326 198	95.9	MB-I	328 855	15.28	DB-KM	0.8%
	100	1000	750	500	246 703	1563.5	MB-I	246 750	1.36	Agg	0.0%
	100	1000	750	600	246 750	14.4	MB-I-II-III-IV-**V**	246 750	0.53	Agg	0.0%
	100	1000	1000	500	257 079	7086.6	MB-I	258 415	0.39	Agg	0.5%
	100	1000	1000	600	289 920	293.6	MB-III	291 220	0.66	Agg	0.4%
	100	2000	750	500	246 750	17.6	MB-I-II-III-IV-**V**	246 750	0.64	Agg	0.0%
	100	2000	750	600	246 750	14.1	MB-I-II-III-IV-**V**	246 750	0.59	Agg	0.0%
	100	2000	1000	500	289 091	1817.4	MB-II	291 896	0.77	Agg	1.0%
	100	2000	1000	600	316 195	1591.2	MB-IV	318 211	0.64	Agg	0.6%
	200	1000	750	500	246 703	10 952.9	MB-I	246 750	0.73	Agg	0.0%
	200	1000	750	600	246 750	33.1	MB-I-II-III-IV-**V**	246 750	0.77	Agg	0.0%
	200	1000	1000	500	257 107	264.2	MB-II	258 401	0.55	Agg	0.5%
	200	1000	1000	600	289 918	618.2	MB-IV	291 207	0.48	Agg	0.4%
	200	2000	750	500	246 750	112.9	MB-I-II-III-IV-**V**	246 750	1.02	Agg	0.0%
	200	2000	750	600	246 750	20.6	MB-I-II-III-IV-**V**	246 750	0.84	Agg	0.0%
	200	2000	1000	500	288 424	5616.5	MB-II	290 557	0.59	Agg	0.7%
	200	2000	1000	600	316 210	7587.1	MB-**II**-III	318 587	0.73	Agg	0.7%

(*continued on next page*)

Table 5 (*continued*).

	50	1000	750	500	137 208	12.3	MB-I-II-III-IV-**V**	137 250	0.19	Agg	0.0%
	50	1000	750	600	137 250	7.0	MB-I-II-III-IV-**V**	137 250	0.38	Agg	0.0%
	50	1000	1000	500	149 127	118.1	MB-I	155 490	0.14	DB-Agg	4.1%
	50	1000	1000	600	165 416	387.8	MB-I	171 279	29.00	DB-Agg	3.4%
	50	2000	750	500	137 250	7.4	MB-I-II-III-IV-**V**	137 250	0.27	Agg	0.0%
	50	2000	750	600	137 250	8.8	MB-I-II-III-IV-**V**	137 250	0.16	Agg	0.0%
	50	2000	1000	500	171 738	164.3	MB-I	180 625	12.25	DB-KM	4.9%
	50	2000	1000	600	181 526	15.7	MB-I-II-III-IV-**V**	183 000	0.12	Agg	0.8%
	100	1000	750	500	137 203	1542.0	MB-I	137 250	0.36	Agg	0.0%
	100	1000	750	600	137 250	6.9	MB-I-II-III-IV-**V**	137 250	0.36	Agg	0.0%
	100	1000	1000	500	143 545	77.8	MB-IV	145 047	0.06	Agg	1.0%
T-1	100	1000	1000	600	161 609	10 815.6	MB-I	163 347	0.06	Agg	1.1%
	100	2000	750	500	137 250	10.8	MB-I-II-III-IV-**V**	137 250	0.09	Agg	0.0%
	100	2000	750	600	137 250	6.5	MB-I-II-III-IV-**V**	137 250	0.12	Agg	0.0%
	100	2000	1000	500	158 497	1076.5	MB-II	161 584	11.00	DB-Agg	1.9%
	100	2000	1000	600	174 963	1032.7	MB-**II**-III-IV	177 024	0.12 2	Agg	1.2%
	200	1000	750	500	137 203	5617.3	MB-I	137 250	0.22	Agg	0.0%
	200	1000	750	600	137 250	20.8	MB-I-II-III-IV-**V**	137 250	0.27	Agg	0.0%
	200	1000	1000	500	143 474	213.1	MB-IV	144 660	0.19	Agg	0.8%
	200	1000	1000	600	161 614	199.7	MB-IV	162 960	0.16	Agg	0.8%
	200	2000	750	500	137 250	68.6	MB-I-II-III-IV-**V**	137 250	0.22	Agg	0.0%
	200	2000	750	600	137 250	8.9	MB-I-II-III-IV-**V**	137 250	0.20	Agg	0.0%
	200	2000	1000	500	158 081	4033.5	MB-III	159 919	0.17	Agg	1.1%
	200	2000	1000	600	174 962	2666.4	MB-**II**-III-IV	177 188	0.22	Agg	1.3%
	50	1000	750	500	272 903	28.3	MB-I-II-III-IV-**V**	273 000	0.69	Agg	0.0%
	50	1000	750	600	273 000	23.9	MB-I-II-III-IV-**V**	273 000	1.03	Agg	0.0%
	50	1000	1000	500	295 412	4044.8	MB-I	309 165	0.27	DB-Agg	4.4%
	50	1000	1000	600	328 309	7637.6	MB-I	340 849	80.00	DB-Agg	3.7%
	50	2000	750	500	273 000	24.0	MB-I-II-III-IV-**V**	273 000	0.67	Agg	0.0%
	50	2000	750	600	273 000	25.0	MB-I-II-III-IV-V	273 000	0.53	Agg	0.0%
	50	2000	1000	500	340 601	6451.3	MB-I	356 786	23.88	DB-KM	4.5%
	50	2000	1000	600	360 573	258.9	MB-I	363 855	23.42	DB-KM	0.9%
	100	1000	750	500	272 903	115.0	MB-**II**-III-IV-V	273 000	0.78	Agg	0.0%
	100	1000	750	600	273 000	24.6	MB-I-II-III-IV-**V**	273 000	0.83	Agg	0.0%
	100	1000	1000	500	286 394	392.7	MB-IV	288 163	1.25	DB-Agg	0.6%
T-2	100	1000	1000	600	321 470	1196.8	MB-III	323 381	1.06	DB-Agg	0.6%
	100	2000	750	500	273 000	63.3	MB-I-II-III-IV-**V**	273 000	1.00	Agg	0.0%
	100	2000	750	600	273 000	24.7	MB-I-II-III-IV-V	273 000	0.95	Agg	0.0%
	100	2000	1000	500	317 411	10 816.9	MB-II	321 548	10.11	DB-KM	1.3%
	100	2000	1000	600	348 226	10 826.6	MB-III	350 986	1.06	DB-Agg	0.8%
	200	1000	750	500	272 903	308.7	MB-**II**-III-IV-V	273 000	1.81	Agg	0.0%
	200	1000	750	600	273 000	68.9	MB-I-II-III-IV-**V**	273 000	1.56	Agg	0.0%
	200	1000	1000	500	286 144	2042.2	MB-II	288 411	0.92	Agg	0.8%
	200	1000	1000	600	321 473	3461.6	MB-IV	324 031	0.75	Agg	0.8%
	200	2000	750	500	273 000	157.8	MB-I-II-III-IV-**V**	273 000	0.95	Agg	0.0%
	200	2000	750	600	273 000	40.1	MB-I-II-III-**IV**-**V**	273 000	0.97	Agg	0.0%
	200	2000	1000	500	317 127	10 839.8	MB-IV	320 934	0.75	Agg	1.2%
	200	2000	1000	600	348 527	10 847.0	MB-III	351 426	0.86	Agg	0.8%
	50	1000	750	500	144 750	8.5	MB-I-II-III-IV-**V**	144 750	0.22	Agg	0.0%
	50	1000	750	600	144 750	9.1	MB-I-II-III-IV-**V**	144 750	0.12	Agg	0.0%
	50	1000	1000	500	167 333	15.1	MB-I	170 781	0.11	DB-Agg	2.0%
	50	1000	1000	600	182 357	20.8	MB-I	184 751	1.38	DB-KM	1.3%
	50	2000	750	500	144 750	8.8	MB-I-II-III-IV-**V**	144 750	0.09	Agg	0.0%
	50	2000	750	600	144 750	10.1	MB-I-II-III-IV-**V**	144 750	0.09	Agg	0.0%
	50	2000	1000	500	189 883	26.0	MB-**I**-II-III-IV	191 602	1.89	DB-KM	0.9%
	50	2000	1000	600	192 855	22.5	MB-**I**-II-III-IV-V	192 855	1.48	DB-KM	0.0%
	100	1000	750	500	144 750	37.8	MB-I-II-III-**IV**-V	144 750	0.16	Agg	0.0%
	100	1000	750	600	144 750	23.2	MB-I-**II**-III-IV-V	144 750	0.14	Agg	0.0%
	100	1000	1000	500	162 116	97.3	MB-I	162 856	0.12	Agg	0.5%
P-1	100	1000	1000	600	179 404	130.7	MB-I	179 811	0.12	DB-Agg	0.2%
	100	2000	750	500	144 750	9.1	MB-I-II-III-IV-**V**	144 750	0.09	Agg	0.0%
	100	2000	750	600	144 750	9.4	MB-I-II-III-IV-**V**	144 750	0.16	Agg	0.0%
	100	2000	1000	500	183 866	417.8	MB-I	184 146	2.53	DB-KM	0.2%
	100	2000	1000	600	191 923	202.1	MB-I	192 217	1.91	DB-KM	0.2%
	200	1000	750	500	144 750	14.7	MB-I-II-III-IV-**V**	144 750	0.09	Agg	0.0%
	200	1000	750	600	144 750	6.7	MB-I-**II**-III-IV-V	144 750	0.17	Agg	0.0%
	200	1000	1000	500	161 971	997.4	MB-I	162 852	0.11	Agg	0.5%
	200	1000	1000	600	179 404	1053.0	MB-I	180 029	0.12	Agg	0.3%
	200	2000	750	500	144 750	9.7	MB-I-**II**-III-IV-V	144 750	0.08	Agg	0.0%
	200	2000	750	600	144 750	10.2	MB-I-II-**III**-IV-V	144 750	0.06	Agg	0.0%
	200	2000	1000	500	183 663	10 844.0	MB-I	184 727	0.11	Agg	0.6%
	200	2000	1000	600	191 923	10 846.6	MB-I	193 000	0.12	Agg	0.6%

(*continued on next page*)

Table 5 (*continued*).

	50	1000	750	500	122 250	8.5	MB-I-II-III-IV-**V**	122 250	0.03	Agg	0.0%
	50	1000	750	600	122 250	8.1	MB-I-II-III-IV-**V**	122 250	0.02	Agg	0.0%
	50	1000	1000	500	143 123	31.3	MB-I	144 856	1.61	DB-KM	1.2%
	50	1000	1000	600	154 843	69.0	MB-**I**-II-III-IV	156 309	0.02	DB-Agg	0.9%
	50	2000	750	500	122 250	9.8	MB-I-II-III-IV-**V**	122 250	0.09	Agg	0.0%
	50	2000	750	600	122 250	9.8	MB-I-II-III-IV-**V**	122 250	0.05	Agg	0.0%
	50	2000	1000	500	160 638	97.1	MB-**I**-II-III-IV	161 998	3.02	DB-KM	0.8%
	50	2000	1000	600	162 908	8.7	MB-I-II-III-IV-**V**	163 000	0.09	Agg	0.1%
	100	1000	750	500	122 250	21.1	MB-I-II-III-IV-**V**	122 250	0.11	Agg	0.0%
	100	1000	750	600	122 250	8.6	MB-I-II-III-IV-**V**	122 250	0.09	Agg	0.0%
	100	1000	1000	500	138 855	74.7	MB-I	139 310	9.00	DB-Agg	0.3%
P-2	100	1000	1000	600	152 906	107.3	MB-I	153 407	52.00	DB-Agg	0.3%
	100	2000	750	500	122 250	39.2	MB-**I**-II-III-IV-V	122 250	0.19	Agg	0.0%
	100	2000	750	600	122 250	12.6	MB-**I**-II-III-IV-V	122 250	0.17	Agg	0.0%
	100	2000	1000	500	156 041	253.8	MB-I	156 800	92.00	DB-Agg	0.5%
	100	2000	1000	600	161 699	306.2	MB-I	162 107	2.66	DB-KM	0.3%
	200	1000	750	500	122 250	15.5	MB-I-II-III-IV-**V**	122 250	0.19	Agg	0.0%
	200	1000	750	600	122 250	5.1	MB-I-**II**-III-IV-V	122 250	0.14	Agg	0.0%
	200	1000	1000	500	138 678	16.4	MB-I-II-**III**-IV	139 331	6.00	DB-Agg	0.5%
	200	1000	1000	600	152 906	733.6	MB-I	153 184	44.00	DB-Agg	0.2%
	200	2000	750	500	122 250	7.3	MB-I-**II**-III-IV-V	122 250	0.11	Agg	0.0%
	200	2000	750	600	122 250	5.3	MB-I-**II**-III-IV-V	122 250	0.11	Agg	0.0%
	200	2000	1000	500	155 865	59.2	MB-I-**II**-III-IV	159 139	129.00	DB-Agg	2.1%
	200	2000	1000	600	161 699	4496.4	MB-I	163 000	0.06	Agg	0.8%

Table 6
The performances of the model-based methods in Stage-2.

	MB-I	MB-II	MB-III	MB-IV	MB-V
Fixed & variable cost problem	115/144 *(80%)*	83/144 *(58%)*	89/144 *(62%)*	94/144 *(65%)*	84/144 *(58%)*
Fixed cost problem	33/72 *(46%)*	13/72 *(18%)*	19/72 *(26%)*	26/72 *(36%)*	12/72 *(17%)*
Variable cost problem	56/72 *(78%)*	42/72 *(58%)*	47/72 *(65%)*	48/72 *(67%)*	50/72 *(69%)*

Table 7
The performances of the model-based methods in Stage-3.

	MB-I	MB-II	MB-III	MB-IV	MB-V
Fixed & variable cost problem	118/144 *(82%)*	92/144 *(64%)*	91/144 *(63%)*	93/144 *(65%)*	86/144 *(60%)*
Fixed cost problem	36/72 *(50%)*	27/72 *(38%)*	33/72 *(46%)*	30/72 *(42%)*	11/72 *(15%)*
Variable cost problem	46/72 *(64%)*	27/72 *(38%)*	21/72 *(29%)*	22/72 *(31%)*	28/72 *(39%)*

Table 8
The performances of the clustering-based methods by problem type.

	Agglomerative	DBSCAN - KMeans.	DBSCAN - Agglom.
Fixed & variable cost problem	106/144 (74%)	92/144 (64%)	121/144 (84%)
Fixed cost problem	33/72 (46%)	21/72 (29%)	28/72 (39%)
Variable cost problem	65/72 (90%)	4/72 (6%)	25/72 (35%)

are observed to improve by 0.065% and 0.079% on average with 10-repeats and 100-repeats, respectively. However, increasing the number of repetitions also increases the solution times as many times as the number of repeats.

Moreover, we performed further experiments with those instances for which the discrete model in a model-based approach cannot find the optimal solution within our three-hour time CPU limit. We increased our time limit to 24 h and investigated if more time improves the solution quality of the methods. We repeated 44 runs with an increased time limit and observed that 29 of them still terminated with an optimality gap of 5.7%, on average. For all experiments, the average improvement obtained in the solution quality is 0.24%. Therefore, it is possible to conclude that although not very significant, there is some room for improvement in the results of the model-based approaches at the expense of significantly increased solution time.

5.4. Comparison of the results with benchmarks

To be able to analyze the contribution of the new methods, we measure the performance of our heuristics on smaller samples that can also be solved by continuous models presented in Section 3 using

Table 9
Comparison of model-based heuristics with the continuous models (F1 = F2=1000, F3=0).

$\lvert\mathcal{N}\rvert$	distLim	Continuous model		MB-I		MB-II		MB-III		MB-IV		MB-V	
		Cost	Time	Cost	Time	Cost	Time	Cost	Time	Cost	Time	Cost	Time
10	50	3239.32	682.95	**3239.32**	0.48	3242.65	1.28	**3239.32**	1.11	**3239.32**	0.55	3242.65	2.22
10	50	3254.44	1115.66	**3254.45**	0.42	3288.09	0.55	3288.09	0.52	**3254.45**	0.48	4193.25	2.34
10	100	3348.81	515.23	**3348.82**	0.44	3375.01	0.7	**3348.82**	0.55	3375.01	0.48	3375.01	2.09
10	100	2484.92	69.98	**2489.67**	0.53	2616.34	0.69	2616.34	0.56	2616.34	0.8	3355.59	2.42
15	100	2850.51	925.6	**2853.96**	0.61	2873.55	0.66	2873.55	0.48	2873.55	0.75	2875.81	2.39
15	100	2726.3	3126.2	**2728.55**	0.75	2734.2	0.66	2734.2	0.67	2734.2	0.55	2734.2	2.83
20	200	6328.81	43%[a]	**6239.39**	0.44	6306.72	0.81	**6239.39**	0.56	**6239.39**	0.56	7562.51	2.86
20	200	7108.26	48%[a]	**7117.2**	0.53	7348.19	0.48	**7117.2**	0.52	**7117.2**	0.59	9201.04	2.08

[a]The instance could not be solved within 24 h time limit and an optimality gap is obtained.

Table 10
Comparison of model-based and clustering-based approaches (F1=F2=1000, F3=0).

$\lvert\mathcal{N}\rvert$	distLim	Continuous model		Best clustering-based method					Best model-based method			
		Cost	Time	Cost	Time	Method	Diss. Measure	Diff	Cost	Time	Method	Diff.
10	50	3239.32	682.95	3242.65	0.05	Agg, DB-Agg	GeomDiff	0.1%	3239.32	0.48	MB I-III-IV	0.0%
10	50	3254.44	1115.66	3310.45	0.03	Agg.	GeomDiff	1.7%	3254.45	0.42	MB I-IV	0.0%
10	100	3348.81	515.23	3348.82	0.004	Agg.	Ward	0.0%	3348.82	0.44	MB I-III	0.0%
10	100	2484.92	69.98	2616.34	0.02	Agg,., DB-Agg., DB-KM	Ward	5.3%	2489.67	0.53	MB-I	0.19%
15	100	2850.51	925.6	2853.96	0.05	Agg,., DB-Agg., DB-KM	GeomDiff	0.12%	2853.96	0.61	MB-I	0.12%
15	100	2726.3	3126.2	2728.55	0.002	DB-KM	–	0.08%	2728.55	0.75	MB-I	0.08%
20	200	6238.81	43%[a]	6789.88	0.05	Agg, DB-Agg	GeomDiff	8.8%	6239.39	0.44	MB I-III-IV	0.01%
20	200	7108.26	48%[a]	7117.2	0.004	Agg.	All	0.13%	7117.2	0.53	MB I-III-IV	0.13%

[a]The instance could not be solved within 24 h time limit and an optimality gap is obtained.

commercial solvers. Additionally, we set the micro-grid and nano-grid costs equal to each other, and aim to better demonstrate the contributions of the new methods using the same problem setting in Gokbayrak and Kocaman (2017). When we set F1=F2 and F3=0, our MB-II approach also reduces to the solution method presented in Gokbayrak and Kocaman (2017). Therefore, MB-II is considered a benchmark case in this special setting. The results for these smaller samples are provided in Tables 9 and 10. Table 9 provides the results of the model-based methods, whereas Table 10 summarizes the best-performing model-based and clustering-based methods and compares them in terms of the objective value, solution time, and their deviation from the optimal solution if applicable.

In Table 9, one can observe that the newly proposed methods (MB-I-III-IV) outperform the benchmark model (MB-II) when there is no distinction between nano-grid and macro-grid costs. The additional candidate points, common circle intersection points, and the centroids of their convex hulls are shown to improve the results that could be obtained by the benchmark model. Moreover, the mid-points of every demand node pair that can be covered by the same facility (MB-I) are shown to make the most significant improvement. For all samples, MB-I is observed to provide the least cost configuration by less than 0.2% deviation from the optimal solution. It is also shown to provide the final result faster than the other model-based approaches in these small instances.

On the other hand, the clustering-based approaches are observed to provide similar configurations within remarkably shorter computational time. The new dissimilarity measure GeomDiff is found to perform better than the other existing measures in the agglomeration process, as shown in Table 10. Given that the continuous model can have difficulty in reaching an optimal solution within 24 hr time limit, even for small samples with 20 nodes, both clustering-based and model-based methods can be useful to obtain approximate solutions. For large samples with thousands of demand nodes, clustering-based algorithms can be more convenient for obtaining cost estimates.

6. Conclusion

In this paper, we present a new planar facility location–allocation problem in the context of rural electrification. In this problem, we consider two different facility types: nano-grids and micro-grids. The planar location–allocation problem we propose for designing rural electrification systems cannot be solved optimally, even for the small samples. Therefore, we present model-based and clustering-based approaches from which energy planners can benefit. Given that each household could be electrified by individual stand-alone systems (nano-grid) in the most trivial solution, the potential micro-grid clusters could lead to remarkable cost reductions based on the trade-off between the decentralized facilities.

The computational results indicate that the model-based approaches are more capable of identifying micro-grid clusters. Contrary to the model-based heuristics, the bottom-up or top-down clustering approaches may overlook some possible micro-grid opportunities as they construct the solutions iteratively. However, one should also note that while these iterative processes can negatively impact the solutions' quality, it helps the algorithm to provide a final solution within a significantly shorter computational time.

Although the results indicate that the model-based approaches can generally provide lower-cost solutions compared to the clustering-based heuristics, we also highlight that these experiments are conducted on relatively smaller samples with less than 400 nodes. Since the samples include only a small portion of the villages, the model-based approaches were able to attain the solution without any optimality gap in most instances. However, once we work on the larger samples or the entire village, we are more likely to observe significant optimality gaps, which may negatively impact the performance of the model-based approaches. Therefore, it can be concluded that while the model-based approaches can work effectively on small-sized samples, the clustering-based methods can find approximate solutions in a shorter time, even for very large samples. Moreover, the clustering-based algorithms can enable energy planners to evaluate the trade-off between two decentralized systems and make rapid assessments without the need for a commercial solver. When the solvers suffer from computational complexity, clustering-based methods can even find better solutions than model-based approaches in some instances. Thus, energy planners could select the most convenient method that aligns with the user's needs, including the choice of the computational environment and the allocated computational time.

Future research may extend our study to consider the grid option as well. In the most simplified case, the electricity generated at centralized facilities can be first distributed via a medium voltage backbone to the transformers, which drop down the voltage and allow low-voltage cables to connect to final consumers. Therefore, a two-level network design can also be introduced to enrich the electrification options. Another extension would be to consider capacitated facilities and cables.

CRediT authorship contribution statement

Beste Akbas: Conceptualization, Methodology, Formal analysis, Software, Investigation, Writing – original draft, Writing – review & editing. **Ayse Selin Kocaman:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision.

Data availability

We have shared the link for the data sets.

Acknowledgments

We gratefully acknowledge the insightful comments and suggestions of three anonymous reviewers and the editors, which led to several improvements in our manuscript. Dr. Kocaman was supported by the BAGEP Award of the Science Academy. Beste Akbas was a graduate student at Bilkent University when this research was initiated and she receives funding from Imperial College London under the President's PhD Scholarships.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cor.2023.106202.

References

Adkins, J.E., Modi, V., Sherpa, S., Han, R., Kocaman, A.S., Zhao, N., Natali, C., Carbajal, J., 2017. A geospatial framework for electrification planning in developing countries. In: 2017 IEEE Global Humanitarian Technology Conference. GHTC, IEEE, pp. 1–10.

Manzour-al Ajdad, S., Torabi, S.A., Salhi, S., 2012. A hierarchical algorithm for the planar single-facility location routing problem. Comput. Oper. Res. 39 (2), 461–470.

Akbas, B., Kocaman, A.S., Nock, D., Trotter, P.A., 2022. Rural electrification: An overview of optimization methods. Renew. Sustain. Energy Rev. 156, 111935.

Batidzirai, B., Trotter, P.A., Brophy, A., Stritzke, S., Moyo, A., Twesigye, P., Puranasamriddhi, A., Madhlopa, A., 2021. Towards people-private-public partnerships: An integrated community engagement model for capturing energy access needs. Energy Res. Soc. Sci. 74, 101975.

Bien, J., Tibshirani, R., 2011. Hierarchical clustering with prototypes via minimax linkage. J. Amer. Statist. Assoc. 106 (495), 1075–1084.

Bolukbasi, G., Kocaman, A.S., 2018. A prize collecting Steiner tree approach to least cost evaluation of grid and off-grid electrification systems. Energy 160, 536–543.

Bongartz, I., Calamai, P.H., Conn, A.R., 1994. A projection method for l_p norm location-allocation problems. Math. Program. 66 (1), 283–312.

Brimberg, J., 2016. Further notes on convergence of the weiszfeld algorithm. Yugosl. J. Oper. Res. 13 (2), 199–206.

Brimberg, J., Drezner, Z., Mladenović, N., Salhi, S., 2014. A new local search for continuous location problems. European J. Oper. Res. 232 (2), 256–265.

Brimberg, J., Drezner, Z., Mladenović, N., Salhi, S., 2016. Using injection points in reformulation local search for solving continuous location problems. Yugosl. J. Oper. Res. 27 (3), 291–300.

Brimberg, J., Hansen, P., Mladenovic, N., Salhi, S., 2008. A survey of solution methods for the continuous location-allocation problem. Int. J. Oper. Res. 5 (1), 1–12.

Brimberg, J., Hansen, P., Mladenović, N., Taillard, E.D., 2000. Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem. Oper. Res. 48 (3), 444–460.

Brimberg, J., Mladenovic, N., 1996. Solving the continuous location-allocation problem with tabu search. Stud. Locat. Ann. 8 (1), 23–32.

Brimberg, J., Mladenovic, N., Salhi, S., 2004. The multi-source Weber problem with constant opening cost. J. Oper. Res. Soc. 55 (6), 640–646.

Brimberg, J., Salhi, S., 2005. A continuous location-allocation problem with zone-dependent fixed cost. Ann. Oper. Res. 136 (1), 99–115.

Cabrera-Celi, G.C., Novoa-Guaman, E.G., Vasquez-Miranda, P.F., 2017. Design of secondary circuits of distribution networks using clustering and shortest path algorithms. In: 2017 IEEE PES Innovative Smart Grid Technologies Conference-Latin America. ISGT Latin America, IEEE, pp. 1–6.

Chaurey, A., Kandpal, T.C., 2010. A techno-economic comparison of rural electrification based on solar home systems and PV microgrids. Energy Policy 38 (6), 3118–3129.

Church, R., ReVelle, C., 1974. The maximal covering location problem. Pap. Reg. Sci. Assoc. 32, 101–118.

Cooper, L., 1964. Heuristic methods for location-allocation problems. SIAM Rev. 6 (1), 37–53.

Corigliano, S., Carnovali, T., Edeme, D., Merlo, M., 2020. Holistic geospatial data-based procedure for electric network design and least-cost energy strategy. Energy Sustain. Dev. 58, 1–15.

Corigliano, S., Rosato, F., Ortiz Dominguez, C., Merlo, M., 2021. Clustering techniques for secondary substations siting. Energies 14 (4), 1028.

Drezner, Z., 1984. The planar two-center and two-median problems. Transp. Sci. 18 (4), 351–361.

Drezner, Z., Brimberg, J., Mladenović, N., Salhi, S., 2015. New heuristic algorithms for solving the planar p-median problem. Comput. Oper. Res. 62, 296–304.

Drezner, Z., Mehrez, A., Wesolowsky, G.O., 1991. The facility location problem with limited distances. Transp. Sci. 25 (3), 183–187.

Energy Data Platform, 2019. Global electrification platform (GEP) modeling parameters. https://energydata.info/dataset/uganda-global-electrification-platform-gep. (Accessed 13 January 2023).

ESMAP, 2019. Mini grids for half a billion people: Market outlook and handbook for decision makers. https://openknowledge.worldbank.org/bitstream/handle/10986/31926/Mini-Grids-for-Half-a-Billion-People-Market-Outlook-and-Handbook-for-Decision-Makers-Executive-Summary.pdf. (Accessed 13 January 2023).

Esnaf, Ş., Küçükdeniz, T., 2009. A fuzzy clustering-based hybrid method for a multi-facility location problem. J. Intell. Manuf. 20 (2), 259–265.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In: KDD'96 Proc. Second Int. Conf. Knowl. Discov. Data Mining, Vol. 96, No. 34. pp. 226–231.

Fobi, S., Kocaman, A.S., Taneja, J., Modi, V., 2021. A scalable framework to measure the impact of spatial heterogeneity on electrification. Energy Sustain. Dev. 60, 67–81.

Gamal, M., Salhi, S., 2001. Constructive heuristics for the uncapacitated continuous location-allocation problem. J. Oper. Res. Soc. 52 (7), 821–829.

Geetha, S., Poonthalir, G., Vanathi, P., 2009. Improved k-means algorithm for capacitated clustering problem. INFOCOMP J.Comput. Sci. 8 (4), 52–59.

Gokbayrak, K., 2022. A two-level off-grid electric distribution problem on the continuous space. Comput. Oper. Res. 144, 105853.

Gokbayrak, K., Avci, H., 2020. A voltage drop limited decentralized electric power distribution network. Comput. Oper. Res. 118, 104907.

Gokbayrak, K., Kocaman, A.S., 2017. A distance-limited continuous location-allocation problem for spatial planning of decentralized systems. Comput. Oper. Res. 88, 15–29.

González-Sotres, L., Domingo, C.M., Sánchez-Miralles, Á., Miró, M.A., 2013. Large-scale MV/LV transformer substation planning considering network costs and flexible area decomposition. IEEE Trans. Power Deliv. 28 (4), 2245–2253.

Han, J., Kamber, M., Pei, J., 2011. The Morgan Kaufmann Series in Data Management Systems, No. 4, vol. 5, pp. 83–124.

Hansen, P., Mladenović, N., Taillard, E., 1998. Heuristic solution of the multisource Weber problem as a p-median problem. Oper. Res. Lett. 22 (2–3), 55–62.

Hausdorff, F., 1957. Set Theory. American Mathematical Soc., Providence.

Hosseininezhad, S.J., Salhi, S., Jabalameli, M.S., 2015. A cross entropy-based heuristic for the capacitated multi-source Weber problem with facility fixed cost. Comput. Ind. Eng. 83, 151–158.

Houck, C.R., Joines, J.A., Kay, M.G., 1996. Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems. Comput. Oper. Res. 23 (6), 587–596.

IEA, 2022. SDG7: Data and projections. https://www.iea.org/reports/sdg7-data-and-projections/access-to-electricity. (Accessed 13 January 2023).

Irawan, C.A., Salhi, S., Chan, H.K., 2022. A continuous location and maintenance routing problem for offshore wind farms: Mathematical models and hybrid methods. Comput. Oper. Res. 144, 105825.

Irawan, C.A., Salhi, S., Soemadi, K., 2020. The continuous single-source capacitated multi-facility Weber problem with setup costs: formulation and solution methods. J. Global Optim. 78 (2), 271–294.

IRENA, 2016. Solar PV in africa: Costs and markets. https://www.irena.org/publications/2016/Sep/Solar-PV-in-Africa-Costs-and-Markets. (Accessed 24 February 2023).

IRENA, 2018. Renewable Power Generation Costs in 2017. Report, International Renewable Energy Agency, Abu Dhabi.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 32 (3), 241–254.

Kocaman, A.S., 2014. Essays on Infrastructure Design and Planning for Clean Energy Systems (Ph.D. thesis). Columbia University.

Kocaman, A.S., Huh, W.T., Modi, V., 2012. Initial layout of power distribution systems for rural electrification: A heuristic algorithm for multilevel network design. Appl. Energy 96, 302–315.
Korkovelos, A., Khavari, B., Sahlberg, A., Howells, M., Arderne, C., 2019. The role of open access data in geospatial electrification planning and the achievement of SDG7. An onsset-based case study for malawi. Energies 12 (7), 1395.
Krau, S., 1999. Extensions du probleme de Weber (Ph.D. thesis). École Polytechnique de Montréal.
Küçükdeniz, T., Esnaf, Ş., 2018. Hybrid revised weighted fuzzy c-means clustering with Nelder-Mead simplex algorithm for generalized multisource Weber problem. J. Enterp. Inf. Manag. 31 (6), 908–924.
Kuenne, R.E., Soland, R.M., 1972. Exact and approximate solutions to the multisource Weber problem. Math. Program. 3 (1), 193–209.
Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognit. 36 (2), 451–461.
Love, R.F., Morris, J.G., Wesolowsky, G.O., 1988. Facility Location: Models and Methods. North-Holland, Amsterdam.
Luis, M., Salhi, S., Nagy, G., 2015. A constructive method and a guided hybrid grasp for the capacitated multi-source Weber problem in the presence of fixed cost. J. Algorithms Comput. Technol. 9 (2), 215–232.
Megiddo, N., Supowit, K.J., 1984. On the complexity of some common geometric location problems. SIAM J. Comput. 13 (1), 182–196.
Moreno, J., Rodríguez, C., Jiménez, N., 1991. Heuristic cluster algorithm for multiple facility location-allocation problem. RAIRO Rech. Oper. 25 (1), 97–107.
Müllner, D., 2011. Modern hierarchical, agglomerative clustering algorithms. Preprint arXiv:1109.2378.
Murtagh, B., Niwattisyawong, S., 1982. An efficient method for the multi-depot location—Allocation problem. J. Oper. Res. Soc. 33 (7), 629–634.
Ostresh, Jr., L.M., 1975. An efficient algorithm for solving the two center location-allocation problem. J. Reg. Sci. 15 (2), 209–216.
Papathanassiou, S., Hatziargyriou, N., Strunz, K., et al., 2005. A benchmark low voltage microgrid network. In: Proceedings of the CIGRE Symposium: Power Systems with Dispersed Generation. CIGRE, pp. 1–8.
Rosing, K.E., 1992. An optimal method for solving the (generalized) multi-Weber problem. European J. Oper. Res. 58 (3), 414–426.
Sahraeian, R., Kaveh, P., 2010. Solving capacitated p-median problem by hybrid k-means clustering and fixed neighborhood search algorithm. In: Proceedings of the 2010 International Conference on Industrial Engineering and Operation Management. pp. 1–6.
Salhi, S., Gamal, M.D.H., 2003. A genetic algorithm based approach for the uncapacitated continuous location–allocation problem. Ann. Oper. Res. 123 (1–4), 203–222.
Salhi, S., Nagy, G., 2009. Local improvement in planar facility location using vehicle routing. Ann. Oper. Res. 167 (1), 287–296.
Sanchez, P., Palm, C., Sachs, J., Denning, G., Flor, R., Harawa, R., Jama, B., Kiflemariam, T., Konecky, B., Kozar, R., et al., 2007. The african millennium villages. Proc. Natl. Acad. Sci. 104 (43), 16775–16780.
Schwardt, M., Dethloff, J., 2005. Solving a continuous location-routing problem by use of a self-organizing map. Int. J. Phys. Distrib. Logist. Manage. 35 (6), 390–408.
Schwardt, M., Fischer, K., 2009. Combined location-routing problems—a neural network approach. Ann. Oper. Res. 167 (1), 253–269.
Sharma, A., Jalal, A.S., 2017. Hybrid algorithm for the facility location problem based on density based clustering and profit maximization. Int. J. Future Gener. Commun. Netw. 10 (11), 47–54.
Sharma, A., Kant, K., Jalal, A.S., 2014. A density based model for facility location problem. In: 2014 Annual IEEE India Conference. INDICON, IEEE, pp. 1–5.
Short, T.A., 2003. Electric Power Distribution Handbook. CRC Press, Boca Raton.
Sneath, P.H.A., Sokal, R.R., 1963. In: Freeman, W.H. (Ed.), Principles of Numerical Taxonomy. W.H. Freeman and Company, San Francisco.
Stephen, B., Mutanen, A.J., Galloway, S., Burt, G., Järventausta, P., 2013. Enhanced load profiling for residential network customers. IEEE Trans. Power Deliv. 29 (1), 88–96.
UNDP, 2023a. Goal 7: Affordable and clean energy. https://www.undp.org/sustainable-development-goals/affordable-and-clean-energy. (Accessed 13 January 2023).
UNDP, 2023b. Sustainable development goals. https://www.undp.org/content/undp/en/home/sustainable-development-goals.html. (Accessed 13 January 2023).
Vardi, Y., Zhang, C.H., 2001. A modified weiszfeld algorithm for the Fermat–Weber location problem. Math. Program. 90 (3), 559–566.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. 58 (301), 236–244.
Weber, A., 1929. Über den Standort der Industrien. Tübingen: JCB Mohr. English translation: The Theory of the Location of Industries. Chicago: Chicago University Press.
Weiszfeld, E., 1937. Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Math. J. (1) 43, 355–386.
World Bank, 2018. Access to electricity. https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?view=map. (Accessed 13 January 2023).
World Bank, 2019. Benchmarking study of solar PV mini-grids investment costs. http://documents1.worldbank.org/curated/en/569621512389752401/pdf/121829-ESM-PVHybridminigridsCostingbenchmarkTTAESMAPConfEdtemplateDecv-PUBLIC.pdf. (Accessed 13 January 2023).
Yu, L., Shi, D., Guo, X., Jiang, Z., Xu, G., Jian, G., Lei, J., Jing, C., 2018. An efficient substation placement and sizing strategy based on GIS using semi-supervised learning. CSEE J. Power Energy Syst. 4 (3), 371–379.
Zvoleff, A., Kocaman, A.S., Huh, W.T., Modi, V., 2009. The impact of geography on energy infrastructure costs. Energy Policy 37 (10), 4066–4078.