energies

MDPI

Article

Clustering Techniques for Secondary Substations Siting

Silvia Corigliano [1,*], Federico Rosato [1,2], Carla Ortiz Dominguez [1] and Marco Merlo [1]

[1] Department of Energy, Politecnico di Milano, 20156 Milano, Italy; federico.rosato@polimi.it (F.R.); carla.ortiz@mail.polimi.it (C.O.D.); marco.merlo@polimi.it (M.M.)
[2] Dipartimento Ambiente Costruzioni e Design, SUPSI, 6952 Canobbio, Switzerland
* Correspondence: silvia.corigliano@polimi.it

Abstract: The scientific community is active in developing new models and methods to help reach the ambitious target set by UN SDGs7: universal access to electricity by 2030. Efficient planning of distribution networks is a complex and multivariate task, which is usually split into multiple subproblems to reduce the number of variables. The present work addresses the problem of optimal secondary substation siting, by means of different clustering techniques. In contrast with the majority of approaches found in the literature, which are devoted to the planning of MV grids in already electrified urban areas, this work focuses on greenfield planning in rural areas. K-means algorithm, hierarchical agglomerative clustering, and a method based on optimal weighted tree partitioning are adapted to the problem and run on two real case studies, with different population densities. The algorithms are compared in terms of different indicators useful to assess the feasibility of the solutions found. The algorithms have proven to be effective in addressing some of the crucial aspects of substations siting and to constitute relevant improvements to the classic K-means approach found in the literature. However, it is found that it is very challenging to conjugate an acceptable geographical span of the area served by a single substation with a substation power high enough to justify the installation when the load density is very low. In other words, well known standards adopted in industrialized countries do not fit with developing countries' requirements.

Keywords: rural electrification; secondary substations; clustering; sustainable development; optimization

Citation: Corigliano, S.; Rosato, F.; Ortiz Dominguez, C.; Merlo, M. Clustering Techniques for Secondary Substations Siting. *Energies* **2021**, *14*, 1028. https://doi.org/10.3390/en14041028

Academic Editor: Laia Ferrer-Martí

Received: 13 January 2021
Accepted: 8 February 2021
Published: 16 February 2021

1. Introduction

In 2019, 770 million people in the world lacked access to electricity. Approximately 75% of them are found in Sub-Saharan Africa alone, where access to electricity rate in rural areas is only 30% [1,2]. In a global framework where the goal of the 7th SDG is to ensure access to affordable, reliable, sustainable and modern energy for all, rural electrification planning is needed. Due to the high upfront investment costs of power infrastructure, a proper design of the network and ability to satisfy the demand while respecting electrical constraints and pursuing the minimization of costs, is needed [3]. Distribution network planning is composed of several steps such as location of feeders, location of substations, allocation of loads, and transformer sizing [4–6]. This work focuses on the problem of distribution substations siting, and is inserted in a wider project, called Gisele (Geographic Information Systems for electrification) [7], for the development of a tool for optimal rural electrification planning. The problem of optimal location and number of secondary substations is complex and multivariate. Distribution feeders and substations should respect geographical and technical constraints, i.e., voltage drops and loading of the conductors, while connecting costumers with adequate value of security and reliability.

Significant research has been carried out on methods for planning substation locations and LV networks in general, with interest arising as computation instruments began to be diffused in electrical design [8–13]. These methods typically introduce a simplified structure of the problem that is then used for formulating a tractable mathematical optimization

problem or offer rule-based approaches. Some older methods require candidate substation locations as input.

Currently proposed planning models to site secondary substations can be divided into three main categories [14]:

- Mathematical or numerical methods (e.g., integer or mixed-integer programming, branch and bound and, network-flow programming algorithm): the models allow to reach a global optimum of the solution but often incur in convergence problems when the size and complexity of the system grows substantially [15–17].
- Heuristic and metaheuristic algorithms: several techniques have been developed to solve problems with different sizes and characteristics. They do not guarantee the optimality of the solution and the convergence and depend on a variety of different input parameters, but when properly set could allow to solve complex problems with good accuracy. Genetic Algorithms [18], Imperialist Competitive Algorithm [19], Particle Swarm Optimization [20], are some examples of the wide variety of algorithms and methods present in the literature.
- Unsupervised learning techniques: data analysis techniques, and in particular clustering techniques, could be an effective way to find the optimal location of distribution substations on the basis of little input data, usually just the population density [21]. The most commonly used algorithm of this category is K-means. In [14], the authors use a combination of K-means and Dijkstra's algorithm to design the LV grid. In [22] a GIS-based and Semi-Supervised Learning Algorithm based on K-means is developed, in [23] K-means with post processing techniques are used for large scale planning.

Some literature works also mix different approaches within iterative procedures in order to increase the accuracy of the results and speed up the computational process [24,25].

When dealing with rural electrification of low income countries, some specific characteristics must be taken into account. Population is dispersed and the load per capita is very low, mainly composed by residential load consumption given the scarcity of industries and production sites [26]. Usually lack of reliable data and historical trends make it difficult to accurately estimate and forecast load consumption and expansion trends [27,28]. Moreover, the budget for electrification projects is generally low, requiring simplified approaches that can guarantee the necessary flexibility. Finally, in most cases the problem to be faced is greenfield planning, since there are wide areas where the national grid is not present at all [29].

The approach followed in the present paper to site secondary substations is based on population clustering due to several reasons. It is a scalable approach, able to find solution of problems of different sizes, usually with reasonable computational effort. There is only a little data and few parameters to be set in input and population data is available on online free databases. Electrical modeling, which could have several unknowns due to the greenfield approach, is performed in a second step, not to overload the optimal substation siting problem. Moreover, those algorithms do not choose the substation location on the basis of a predetermined set of solutions, as other techniques usually do but are able to locate them on the basis of geometric criteria.

The goal of the paper is to compare different clustering techniques, to show strengths and weaknesses when dealing with the complex problem of secondary substations planning. The novelty of the proposed work is to explore and adapt different algorithms to the specified problem, without focusing just on the widely spread K-means technique, and compare the results applied on two different real case studies. The analysed literature works based on clustering techniques focuses in fact mainly on K-means algorithm and apply different post processing techniques to improve its results. None of the cited works consider comparing K-means with other techniques and only seldom they include complete analysis related to technical constraints. In [22], the authors subdivide clusters which overcome a capacity limit without considering maximum cluster extension, included instead in [23]. Among the wide variety of clustering techniques, three have been selected: K-means, hierarchical agglomerative clustering and a graph partitioning algorithm [30].

K-means is the literature benchmark and used in the present paper as the reference model; a variant is proposed including a post processing technique. Agglomerative clustering is interesting because, differently from K-means, it allows providing as input a distance threshold without specifying a priori the number of clusters. The graph partitioning algorithm has interesting potentialities, allowing to design at the same time substations area and route the LV grid. Finally, all the techniques do not create outliers, meaning that all the points belong to one cluster, a necessary condition to provide total electrification in the region. Algorithms are investigated in terms of their suitability to the problem, analyzing their ability to respect some important indicators, specifically the maximum area supplied by substations, a proxy for voltage drops, and the power associated to each transformer which should be comprised within maximum and minimum levels.

The rest of the paper is organized as follows. Section 2 shortly describes the Gisele procedure, the wider project where the present work aims to be inserted. Sections 3–5, report the proposed clustering methods and the authors' adaptation to the problem. Section 6 describes the two analyzed case studies and Section 7 summarizes the main results obtained.

2. Gisele Project

As stated in the introduction, the present work was developed as part of an open source-tool for optimal rural electrification planning, being developed by the authors during the last few years. The tool, called Gisele (GIS for rural Electrification), is written in Python language and is available, in its first open release, on GitHub (see web page www.e4g.polimi.it (accessed on 20 December 2020)). Figure 1 shows the flowchart of the procedure that, on the basis of Geographic Information System data, aims to identify and design the optimal electrification strategy of rural areas choosing among grid extension and off-grid technologies. The procedure starts with a population clustering algorithm, to identify densely populated areas in the region of interest. For each cluster it then performs a preliminary design of the distribution grid connecting populated points, the sizing of a microgrid supplying the cluster's energy needs and finally it performs a least cost optimization to identify which clusters to connect to the utility power grid already in place and which to electrify with isolated microgrids. In the flow chart, the additional step provided by the work presented in this paper is highlighted in red. Siting of secondary substations will allow the procedure to simplify the subsequent algorithms and to design more accurately the medium voltage grid.

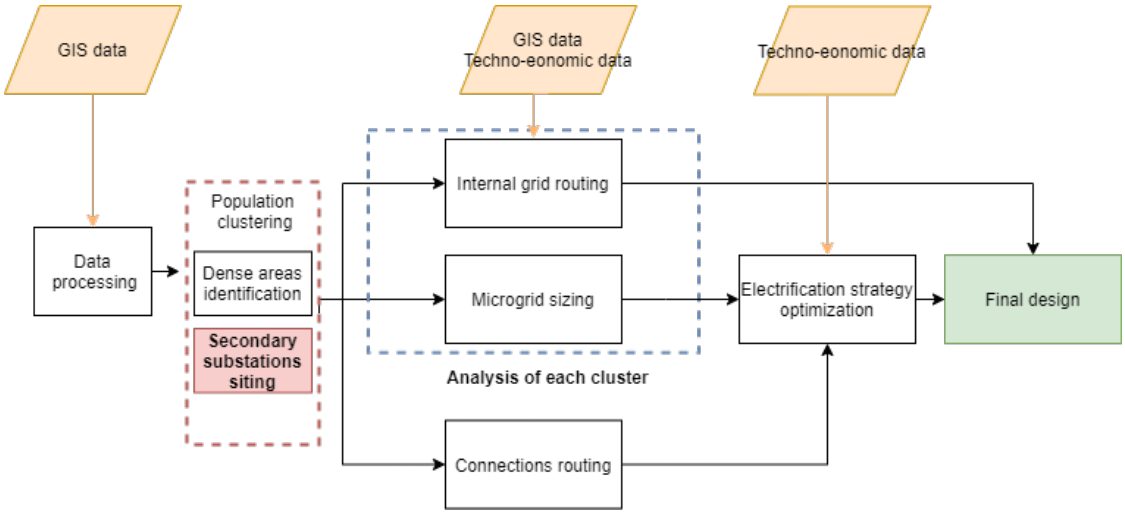


Figure 1. Flow chart of Gisele procedure.

3. K-Means Based Clustering

The first algorithm chosen to size and locate substations is the K-means algorithm, one of the simplest unsupervised learning algorithms applied to the clustering problem. K-means clustering is a method belonging to the family of center based algorithms that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [31]. This results in a partitioning of the data space into Voronoi cells that are clusters with homogeneous convex shapes. The algorithm starts with a random initial partition and keeps reassigning the observations to clusters based on the similarity between the observation and the cluster centres until a convergence criterion is met. The objective function is the minimization of the sum of the distances of the points with their cluster centroids:

$$Obj = \sum_{i=1}^{k} \sum_{j=1}^{n} \left| c_i - x_j \right|^2 \quad (1)$$

being c_i the location of the centroid of cluster i and x_j the location of the point j belonging to cluster i. k is the number of clusters and n the number of points belonging to cluster i.

The method is relatively scalable and efficient in processing large data sets and its average time complexity increases linearly with the number of observations being O(k n T), where n is the number of observations and T the number of iterations. Within the present work, each centroid found by the algorithm represents the position of a secondary substation, which supplies with a LV grid all the points belonging to the same cluster. The only input parameter required by the algorithm is the number of clusters k. There is hence no control over the size of the clusters neither terms of geographical extension nor in terms of number of points included within each cluster.

In order to solve some of the criticalities intrinsic in the algorithm, two improvements have been implemented.

1. Weighted k-means clustering: the distance of the point with respect to the centroids of the clusters is weighted according to the power associated (w_i); the objective function becomes hence the following

$$Obj = \sum_{i=1}^{k} \sum_{j=1}^{n} w_i \left| c_i - x_j \right|^2 \quad (2)$$

   This process allows having more homogeneous clusters in terms of peak power load demand since centroids will be biased towards high power nodes.
2. Clusters post processing: the algorithm is iteratively run, subdividing wide clusters until all of them respect the distance constraint limits (i.e., distance threshold set equal to the LV grid maximum feeders' length). The procedure is reported in the flowchart of Figure 2.

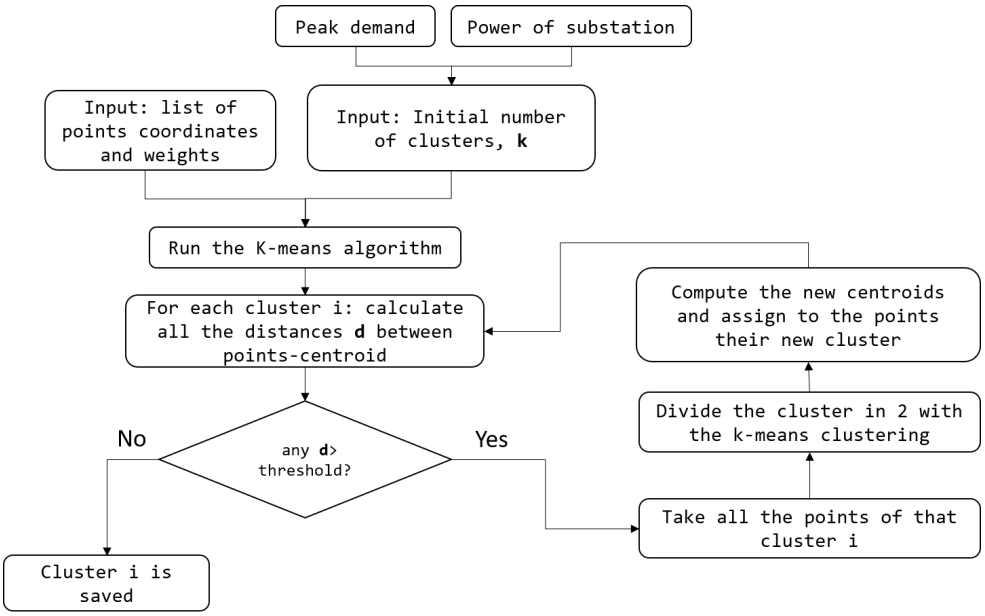


Figure 2. Flow chart of kmeans post processing.

Input Data

As stated above, in addition to the coordinates and weights of observations (i.e., the populated points in the region of interest and the power associated), the only input data required by the algorithm is the number of clusters k. Different approaches are used in the literature to identify a priori the optimal number of clusters k, e.g., elbow method, information criterion approach, silhouette, etc. [32,33]. Those are mainly mathematical approaches, looking at similarities among data. Our problem is more a geographical problem, where the issue is subdivide an area into a reasonable number (reasonable from an electrical standpoint) of homogeneous subareas [34]. Within the present work, the required number of clusters, i.e., the number of distribution transformers to be installed, is estimated by subdividing the total peak power load by the expected size of distribution transformers. The reasoning is to try to create an acceptable number of clusters in terms of peak power and then, with the procedure described above, subdivide the ones which do not respect distance constraints. When applied in practice, this procedure would require sensitivity by the user that could act on the input parameters, weights and number of cluster, to bias the solution towards smaller or bigger sized transformers.

4. Agglomerative Clustering

The second algorithm analysed within the present work is agglomerative clustering, which belongs to the family of hierarchical clustering algorithms [35]. It is a bottom-up approach that, starting from several clusters, one for each observation, i.e., populated point, aggregates the closest clusters until reaching a maximum number of clusters or a certain distance threshold, while minimizing clusters dissimilarities. It is necessary to measure dissimilarities among observations in order to decide which clusters to combine first, choosing a metric and a linkage criterion. In the present work, the Euclidean distance metric, i.e., the linear distance among two points, is used to compute pairwise distances among observations. The selected linkage criterion, i.e., the measure of the distance between to clusters, is the complete-linkage-clustering. This method computes the distance between two clusters (C_1 and C_2) as the distance among the two points (x, y) (each one belonging to one clusters) which are farthest away from each other:

$$D(C_1, C_2) = max(d(x, y)) \quad x \in (C_1), y \in (C_2)$$

The procedure is stopped when the selected distance threshold, a parameter set by the user, is reached. The found clusters represent the area of afference of each secondary substation, which can then be sited by computing the centroid of each cluster. The possibility of imposing a distance threshold on the created clusters avoids the risk of creating too extended clusters, incurring high voltage drops. On the other hand, there is no limit on the peak power of each cluster and the computational time is greater than K-means with an order of magnitude of $O(n^2 logn)$ or in the best cases $O(n^2)$. Those characteristics make the algorithm particularly suited for siting substations in sparse populated rural areas and less appropriate for dense urban areas.

Input Data

The algorithms requires in input the coordinates of the populated points and a parameter: either the number of clusters or a distance threshold. In the present work, the latter is used. A higher distance threshold allows creating a smaller number of clusters with bigger dimension. Since the maximum distance between two points in a cluster represents a rough estimation of the maximum cluster diameter, the distance threshold is set to be twice the maximum acceptable radius of the low voltage grid. Imposing a constraint on this value does not bound the low voltage feeder lengths but it works as a good proxy for the maximum distance between cluster centroids (i.e., substations) and points.

5. Clustering and Substation Placement Based on LUKES

Alongside the aforementioned clustering methods, we propose a method based on the classic algorithm introduced by Lukes [36] for finding the optimal weighted clustering of connected acyclic graphs, i.e., trees. We will first discuss the procedure and then recap the necessary input data.

5.1. The LUKES Partitioning Algorithm

Given a tree $\tau = (V, E)$ with associated node weights $w_i \in \mathbb{N}, i \in V$ and edge weights $v_{ij} \in \mathbb{R}, (i, j) \in E$, the algorithm finds the optimal partition T^*. A partition of a graph T, much like a partition of a set, is defined as collection of k disjoint clusters of nodes $C_y = \{i_n\}$, $y \in 1...k$, such that $\cup_{y=1}^{k} C_y = V$.

T^* is optimal in the following sense: under the constraint that the weight in each cluster is less or equal to a weight limit W, the partition T^* is such that the sum of the weights of the intracluster edges (edges from the original tree τ that connect nodes in the same cluster) is maximal. This property is equivalent to the weight of the edges cut by the partition being minimal.

$$T^* = \underset{T}{\operatorname{argmax}} \sum_{C_y \in T} \sum_{\substack{i,j \in C_y \\ (i,j) \in E}} v_{ij} \tag{3}$$

$$s.t. \sum_{i \in C_y} w_i \leq W \qquad \forall C_y \in T$$

We will refer to this algorithm with the name LUKES, and its complexity, as shown in the original article, is $O(W^2 n)$. With respect to the other clustering methods presented in this article, LUKES does not leverage a distance metric in order to establish clusters of points, but needs an underlying tree that connects the nodes to operate. This characteristic is both an advantage and a disadvantage. The downside is that, when using a dataset consisting only of load endpoints and their associated power requirements, preliminary work has to be done in order to establish an underlying tree for the LUKES algorithm to work on. On the other side, this state of affairs offers a few advantages:

- The edge weights of the tree can be adapted to mirror arbitrary, non homogeneous metrics, representing e.g., obstacles or difficult terrain;

- The edges of the underlying tree can be forced to follow preferential paths. A typical example are streets. This makes the method suitable also for clustering tasks in urban areas where the street plan is available, and would influence the actual routing.

5.2. Dataset Preprocessing

As previously mentioned, it is necessary to preprocess the dataset in order to obtain a starting tree usable by the LUKES algorithm to obtain the clusters. In the present study, the data available is in the form of a discretized grid of power request density, and thus the preprocessing will consist in the calculation of the minimum spanning tree of the graph where the nodes represents the data points, and the edges exist only between a node and its k nearest neighbors, where k is chosen high enough to grant a connected minimum spanning tree [37]. The weight of the edges, which for our purposes must inversely mirror the cost of connecting the two nodes, is calculated to be to the opposite of the physical length of the connection it represents, plus an offset to keep the values strictly positive. In this way, the minimization of the weight of the edges cut by the partitioning maximizes the avoided cost.

5.3. Heuristic Node Reallocation

Equipped with the preprocessed dataset in the form of a tree, the LUKES algorithm can be executed on it. We now introduce a set $\{W_1 ... W_n\}$ of real-valued powers corresponding to the commercially available substations. The global cluster weight limit W to use in problem (3) is chosen as the power of the distribution substation of maximum size, $W = W_n$. This ensures that no cluster will include a total power demand higher than the largest available substation size. The LUKES algorithm has no lower bound on cluster size, but only the aforementioned global upper bound; furthermore, it does not consider natively the intermediate sizes. Therefore, it is possible that the clustering that solves the optimization problem (3) leads to clusters i whose weight (power) sum $\sum_{i \in C_y} w_i$ is very small, or just above one of the values $\{W_1 ... W_n\}$. This would force to install, to serve those clusters, oversized substations. Therefore, starting from the solution obtained, we want to find a *reallocation* of some nodes among the clusters that minimizes a virtual cost of the substations and the lines, calculated using a normalized cost associated with each substation power and the opposite of the edge weights, as further explained in Section 5.5. In order to tackle this problem, a reallocation procedure is identified. The nodes included in each cluster are considered embedded in euclidean space by using their geographical coordinates. In this reference system, the convex hull of the points is calculated, and a set of *boundary nodes* $\{\beta_{y,\iota}\}$ is selected as those nodes that respect both the following conditions:

- Lie on the convex hull;
- Have neighbors (in the original graph) in another cluster.

It is trivial to show that any connected subgraph of a tree is a tree, therefore each of the subgraphs τ_y induced by the cluster C_y is a tree. Furthermore, it is always possible by removing one edge (i, j) to subdivide a τ_y in two parts: $(\tau_{y,(i,j)_1}, \tau_{y,(i,j)_2})$. We proceed by enumerating such subdivisions of each τ_y for which the following conditions are respected:

- The first part includes at least one boundary node;
- the end-user nodes contained in the first part sum to a power under a certain global threshold W_r.

In symbols:

$$\tau_{y,(i,j)_1} \cap \{\beta_{y,\iota}\} \neq \emptyset \tag{4}$$

$$\sum_{i \in \tau_{y,(i,j)_1}} w_i \leq W_r \tag{5}$$

By doing so, we obtain an enumeration of the reallocations, i.e., couples $(\beta_{y,\iota}, \tau_{y,e,(i,j)_1})$ that represent nodes that it is possible to reallocate to the neighboring cluster connected to $\beta_{y,\iota}$. These reallocations, in general, are not all compatible with one another. Two

reallocations may involve one or more of the same nodes, and therefore not be compatible. According to this relationship, the reallocation may be arranged in a graph whose nodes are the reallocations themselves, and an edge exists if the two corresponding reallocations are compatible. We name this arrangement the *compatibility graph* G_τ. Notice that the compatibility graph is not uniquely defined for the starting tree τ, but it depends on the parameters that we chose in the previous steps, such as the available transformer catalog, W_r, et cetera. The compatibility graph allows us to find groups of possible reallocations that correspond to *cliques* (complete subgraphs) in G_τ. Since in such a clique all the reallocations are connected by an edge with one another, they can all be executed at the same time. Even for modest W_r, enumerating all cliques in the compatibility graph would quickly become computationally heavy. Luckily, the reallocations are largely redundant, in that reallocations involving similar total node powers between the same two clusters would have roughly the same effect on the resulting cluster sizes. We therefore find a few, representative reallocations for each ordered pair of clusters. The reallocations are divided in bins with width equal to the power of the smallest reallocation, and for each bin, we select only one: the one that has maximum degree in G_τ. All other reallocations are cancelled from the compatibility graph, obtaining a reduced compatibility graph G_τ^R whose size is dramatically smaller than the original. The cliques of G_τ^R are then simply enumerated, and the one that allows minimizing the sum of opposite edge weights (connection costs) and substation normalized cost is chosen as the optimal one and applied to the clusters, obtaining the final clustering solution.

5.4. Substation Placement

Once the final clustering solution is obtained, we proceed to place the substations. For each cluster C_y, the substation is placed on an edge of the corresponding graph τ_y. The criterion used to choose the edge is *edge betweenness centrality*, a centrality measure based on the number of shortest paths between couples of nodes containing that edge. This measure was chosen because edges with high centrality will tend to have a topologically baricentric position in the cluster, with a favorable distribution of distances from the substation and thus minimal total line cost given a certain level of prescribed network performance.

5.5. Input Data

Geographical coordinates with the associated power request are, as with the other methods, the input data that is elaborated into a starting tree as explained in Section 5.2. The parameters needed by the algorithm are weights to be associated with the graph edges and a catalog of available substation sizes with associated normalized costs. As explained above, the final solution is guided by evaluating the sum of the weights of the edges and the normalized cost of the substations, so a brief discussion is needed about these parameters. Since the cost of the lines cannot be fully determined before performing the sizing of the grid, the weight of the lines is considered to be, as stated in Section 5.2, equal to the opposite of their geometrical length in meters, plus an offset to keep all values strictly positive. In rural case studies like the ones considered in this paper, particularly Omereque, since the method does not intrinsically consider limitations on the geographical span of a single feeder, the catalog of substation powers is intentionally limited to substations that will not accomodate an area too extended. This implicitly represents an active contstraint of the optimization problem, and therefore there is no need to carefully normalize the costs associated to the substation; they can be simply scaled to be very low with respect to the line costs to incentivize the use of the bigger sizes inside this limit, using the smaller ones when needed to cover leftover parts of the area. If only one substation power is used, obviously, the cost will not have any influence on the solution.

6. Case Study

Two different case studies have been analysed to test the siting substations procedure. The first one is related to a densely populated area in Namanjavira, a rural administrative post in Zambezia, Mozambique. The second case study is Omereque, a municipality of Cochabamba, Bolivia, inhabited by few people scattered on the mountains. The reasons behind the choice of the case studies are manifold. First of all, the authors have contacts with NGOs working locally to provide support for the fulfilmment of SDG7, i.e., access to electricity for all. This allowed to gather reliable data related to dispersed and non electrified areas, which is normally difficult or subject to many uncertainties. The study presented in this paper, moreover, can be considered a step of wider studies where the issues of optimal electrification in the communities will be adressed more comprehensively. Finally, the two communities are interesting because of their diverse nature. They are related to areas of approximately the same surface but starkly different number of inhabitants (see different population densities in Table 1), inhabitants have different energy needs (i.e., power per capita) and distribution grids regulations are different. Even though these two cases are not enough to have a perspective on the whole rural electrification problem, they allow to test the algorithms in different conditions and to point out pros and cons of each approach.

Table 1. Parameters of the case study.

	Namanjavira	Omereque
N people	18,920	700
Area (km^2)	425	465
Population density (pp/km^2)	4	5
People/household		
Average load/household (kW)	0.4	1.5
LV feeders maximum length (m)	1000	600
Transformer nominal power (kW)	50	20

6.1. Mozambique-Namanjavira

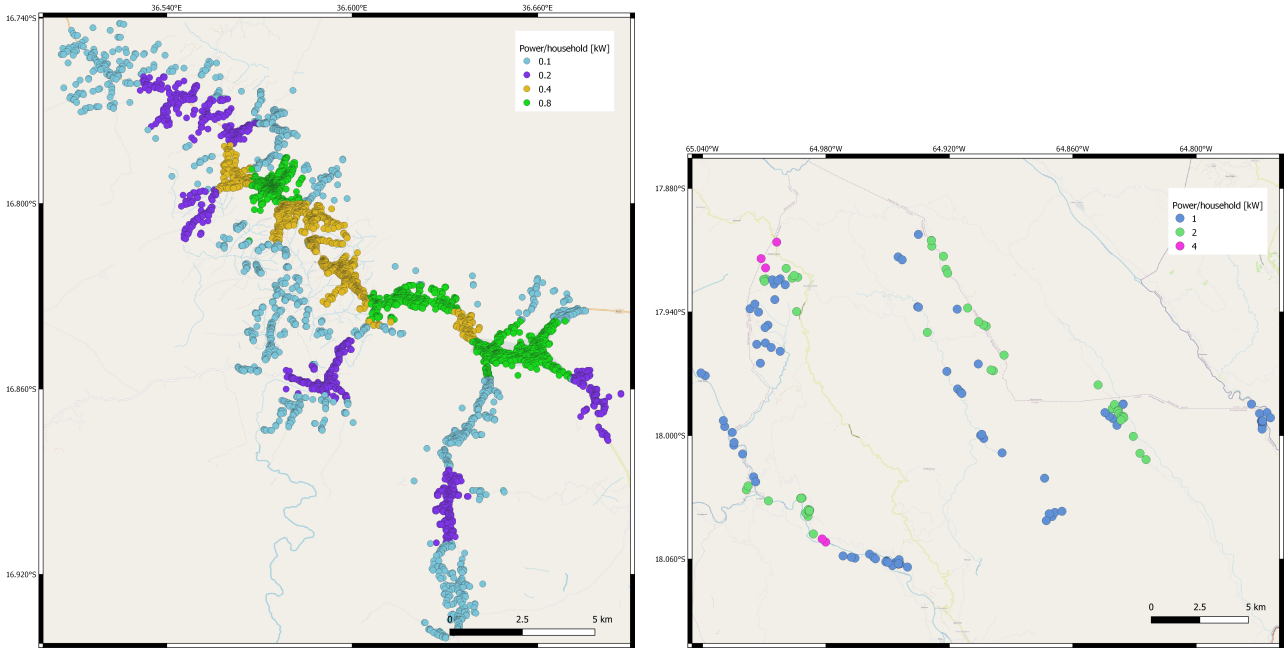
Mozambique has an electrification rate of 35%, dropping at 22% when looking at the rural population, with a total of 20 milion people still without access to electricity. Zambezia province is one of the most critical in terms of development and energy. It hosts one fifth of the country population, 93% of whom lives in rural areas with low electrification access. Data collected for the present work are related to the Namanjavira administrative post in the region and gathered thanks to a collaboration with the NGO COSV.

Population data of Namanjavira was collected from the High Resolution Settlement Layer, developed by Columbia University [38] with a resolution of 30 m, that allows distinguishing between single households. Tested with on-field data coming from some villages in the area they resulted accurate in detecting populated spots. In the area are present approximately 4730 households for a total of more than 18,000 people. The power per capita associated to each populated point was derived by [7] together with a geographical analysis where more power is associated to points closer to main roads and village centers. No information related to the maximum length for low voltage feeder was provided. It was hence set at 1000 m as a standard value.

6.2. Bolivia-Omereque

Bolivia has an electrification rate of 93%, one of the lowest in South America, dropping to 79% in the rural areas of the country.The municipality of Omereque is composed by 11 communities without access to electricity, for a total of 137 households. The NGO Luces Nuevas, operating in the Bolivian territory, and studying strategies for granting electricity

access in the region, provided the data related to the inhabited and non electrified houses. An average number of 5 people per household has been assumed, with an average power per households varying according to the distance from roads and established settlements. The average peak power per household has been assumed equal to 1.5 kW, equivalent to Tier 4 of the Multi-Tier Framework [26]. Only a high value of electricity consumption would in fact justify the connection to the national grid in that remote area. Distance constraint for the low voltage grid is 600 m according to Bolivia regulations. Table 1 summarizes the principal characteristics of the studied areas.



(**a**) Namanjavira (**b**) Omereque

Figure 3. Population distribution of the case studies.

Figure 3 shows the population distribution and the power associated to the households of the two case studies. Table 2 summarizes the parameters used for the various methods to perform the substation placement. For M4, the input parameter is the catalog of the available substation sizes, together with a normalized cost used in the heuristic reallocation procedure described in Section 5.3. This parameter is discussed in more detail in Section 5.5.

Table 2. Simulations summary.

Model	Algorithm	Input Parameters	Value Omereque	Value Namanjavira
M1	Weighted Kmeans	N clusters	10	37
M2	Weighted Kmeans with post processing	N clusters distance constraint	10, 600 m	37, 1000 m
M3	Agglomerative	distance constraint	1200 m	2000 m
M4	Lukes	Substation power(cost)	20(-)/100(3)-50(2.5)-20(2) kW	50 kW

7. Results and Discussion

The four methods have been run on the two different case studies, for a total of 8 simulations, summarized in Table 3. We compare the methods according to five metrics: number of clusters, maximum cluster radius, maximum and minimum cluster power and computational time. The cluster radius is computed as the maximum distance between each of the cluster points and the centroid. This is a proxy for the maximum length of low

voltage feeders, that should be limited in order to avoid high voltage drops. The LUKES-based algorithm is the only one that provides a first design of the low voltage grid, making it possible to better estimate low voltage feeders' length while the others only estimate the influence area of each transformer. Nevertheless, in order to make all algorithms comparable, the linear distance between populated points and substations is considered for all the models. Computational time has been inserted among the indicators due to the fact that the procedure is intended to be part of wider tools for rural electrification planning (e.g., [7]). It should constitute a first step for pre-feasibility studies within the planning of wide geographical areas and as such simplicity and computational speed are relevant advantages which could lead to the choice of one approach with respect to others.

Table 3. Comparison of the methods.

	Namanjavira				**Omereque**			
	M1	**M2**	**M3**	**M4**	**M1**	**M2**	**M3**	**M4**
N clusters	37	100	58	49	10	47	42	12
Max cluster radius (m)	2000	999	1316	3051	3614	585	754	5041
Max cluster power (kW)	236	198	305	50.0	32	20	24	20
Min cluster power (kW)	4	0.1	1	1.5	10	1	1	11
Computational time (s)	2.46	6	0.34	86,000	0.11	0.97	0.02	20

In Figures 4 and 5, the distribution of the results for cluster radius and power are displayed in a boxplot. The following maps (Figures 6–9) show graphically the clustering results. The model results of M1 have not been reported on a map because they are related to the simple K-means procedure, improved by M2. As for Lukes model, two results related to Omereque case study are reported. In the first one, three possible transformers sizes are provided to the algorithm, in the second one transformer maximum size is fixed to 20 kW.

A first striking observation is that, given the low demand and high extension of the areas to serve, particularly Omereque, there is a compromise to be struck between cluster radius and substation power. From plots of Figures 4 and 5 it emerges very clearly that limiting the geographical span of the clusters, as done explicitly in M2, has a repercussion in terms of a very low, difficult to realize substation power distribution. M1, that is the classical weighted K-means procedure, does not perform well for neither of the metrics, since clusters radius and substations power values are spread over a wide range of values, not respecting given constraints. The post processing procedure implemented by the authors represents an effective solution to overcome the radius problem but with little control over power and optimality of the subdivision. Computational time increases but it is still reasonable even in Mozambique case, with many observations. The agglomerative clustering method (M3) outperforms K-means over different aspects: clusters are more homogeneous and their distance can be controlled without interfering with the algorithm, moreover computational time resulted to be lower. Finally, the algorithm is able to respect distance constraints with a much lower number of clusters, especially in Namanjavira case. Considering that each installed transformer has investment and operational costs, this method could allow relevant savings for the investors. From the box plots it can be seen, however, than there is no control over cluster power, that on average has very low values but peaks up to 300 kW in Namanjavira case. Conversely, the LUKES-based method makes explicit use of available substation sizes but has no explicit check on cluster radius. If offered a portfolio of substation sizes ranging up to 100 kW, the cluster radius becomes rapidly unmanageable on the cases analyzed, with projected maximum user distances from the substation of several kilometers (see Figure 9a). The algorithm chooses in this case the maximum transformer size (i.e., 100 kW) for each of the clusters. If using a power-based method like this, therefore, one has to force a low maximum substation size available in order to limit the radius indirectly. A forced size of 20 kW, at the very low end of the spectrum of realizable substation powers, was found to induce clusters with

maximum radius of a few kilometers on the least dense areas (see Figure 9b), which is nevertheless above the prescribed limit. Namanjavira, on the other hand, is more densely populated and the algorithm suggests clusters of acceptable radius (see Figure 8) when the transformer size is set at 50 kW. LUKES proved to be the slowest of the methods and it had a very high run time in the Namanjavira case, the biggest one. Since the LUKES-based clustering explicitly takes into account electroduct pathways and has highest chances of inducing clusters of acceptable radius in densely populated areas, further research will be conducted on this method applied to concentrated underlying demand where preferred pathways cannot be ignored, i.e., urban areas. From a technical perspective, the methods proposed provide an improvement with respect to the most commonly used approach for substations siting, that is the subdivision of the area into geometric cells of uniform size [4], and underlines the importance of studying different methods and solutions. In particular, between the methods investigated, M3 resulted to be the most suited to the distribution grid routing. It results computationally effective and capable to properly manage boundaries correlated to the maximum length of the feeders, respecting the technical viability of the solution. Finally, since clusters' power is not bounded, the nominal power of the substations could be adapted to each singular sub-area characteristics (i.e., very small transformers are proposed for very scattered areas whilst bigger transformers are selected for densely populated areas).

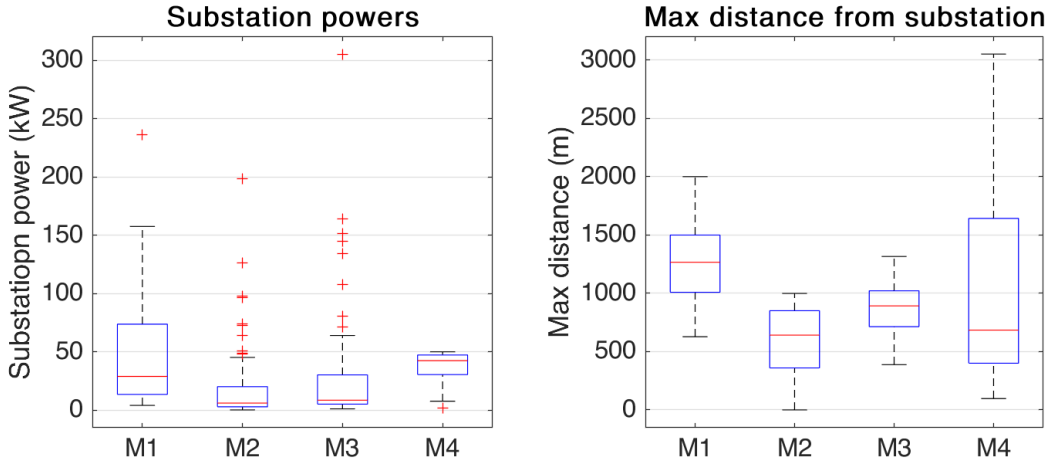


Figure 4. Namanjavira-Power and max distance distribution.

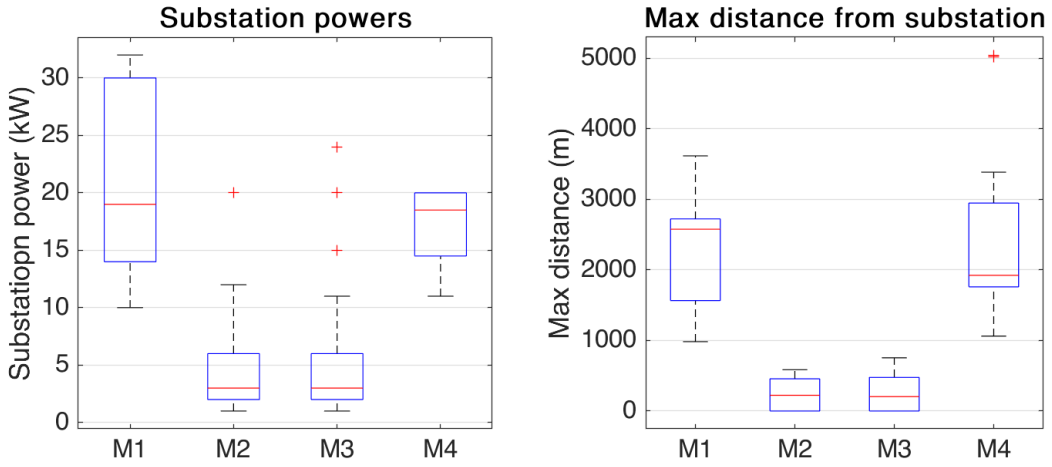
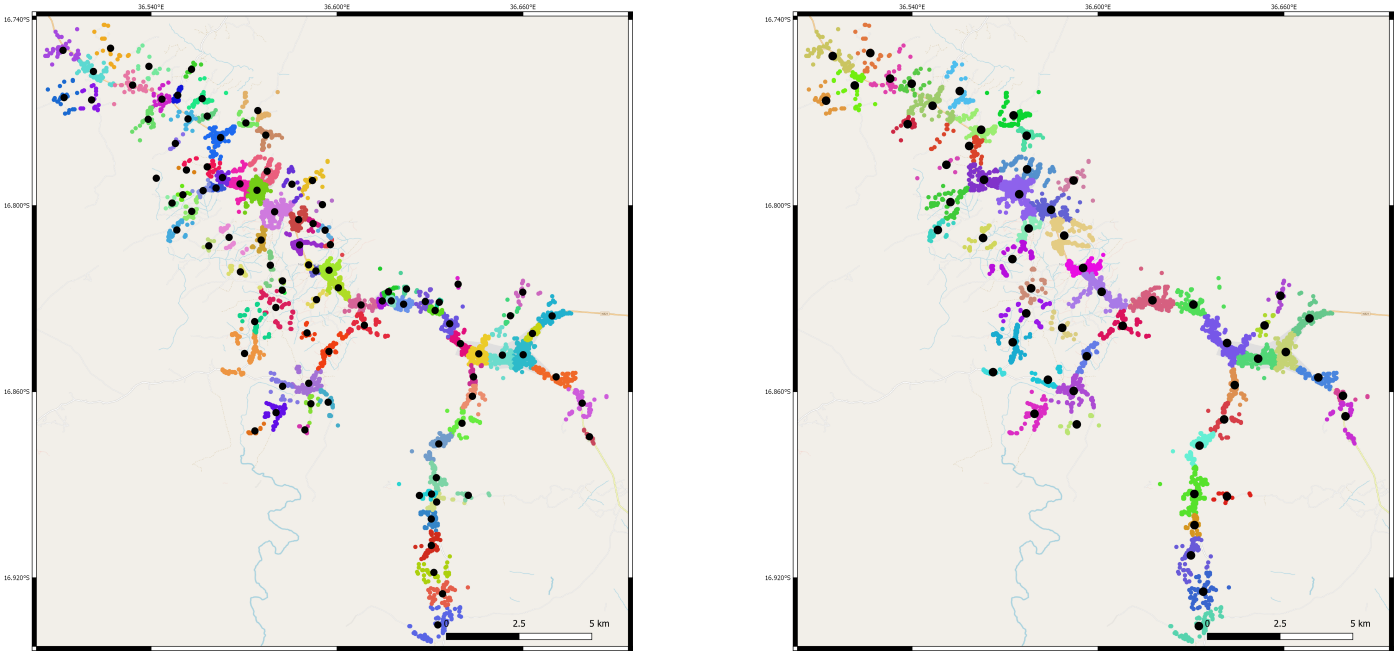
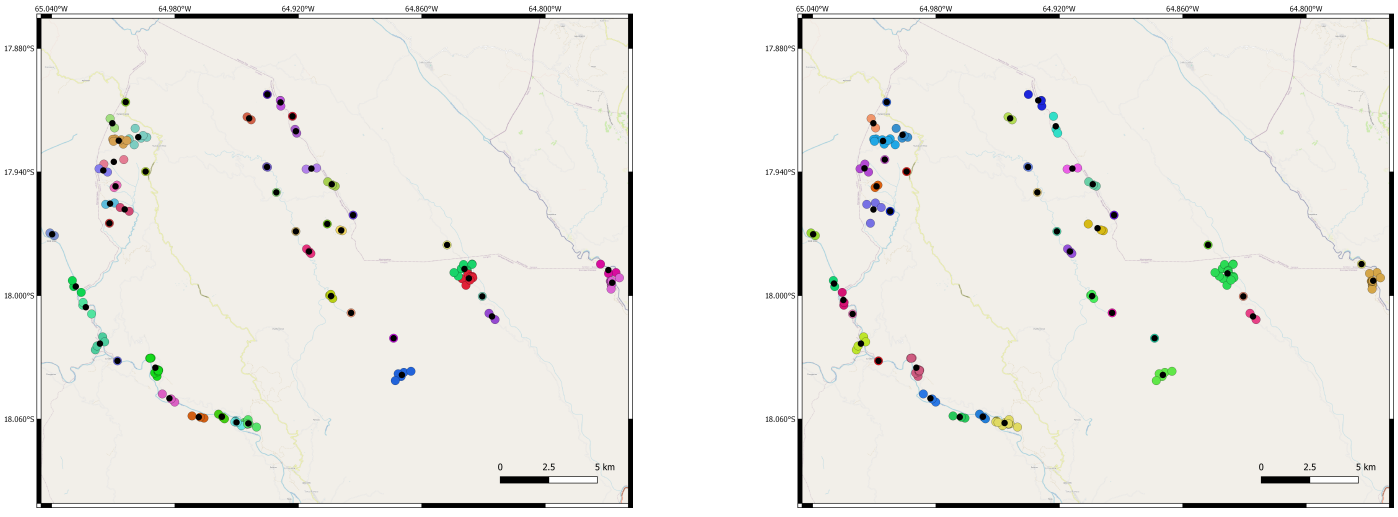


Figure 5. Omereque-Power and max distance distribution.



(**a**) M2: Kmeans with post processing (**b**) M3: Agglomerative

Figure 6. Results of clustering procedures in Namanjavira.



(**a**) M2: Kmeans with post processing (**b**) M3: Agglomerative

Figure 7. Results of clustering procedures in Omereque.

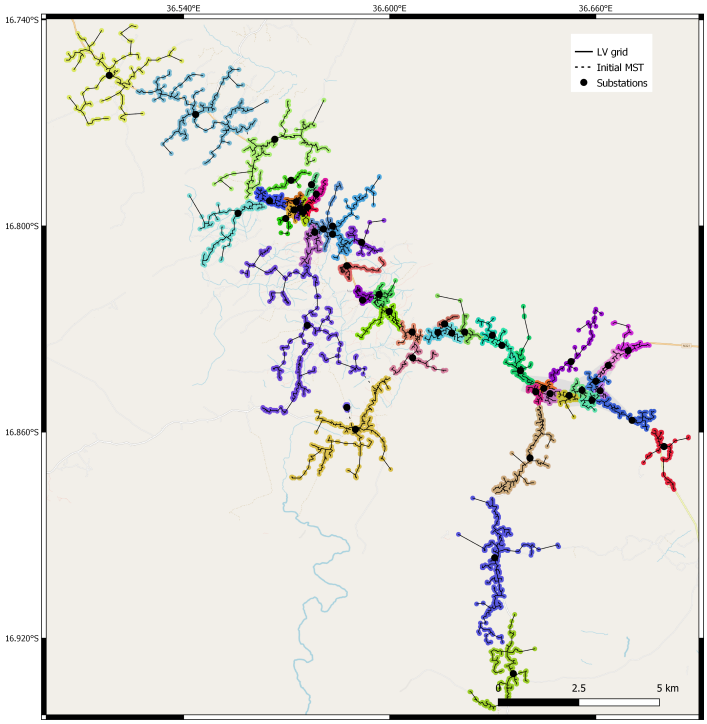
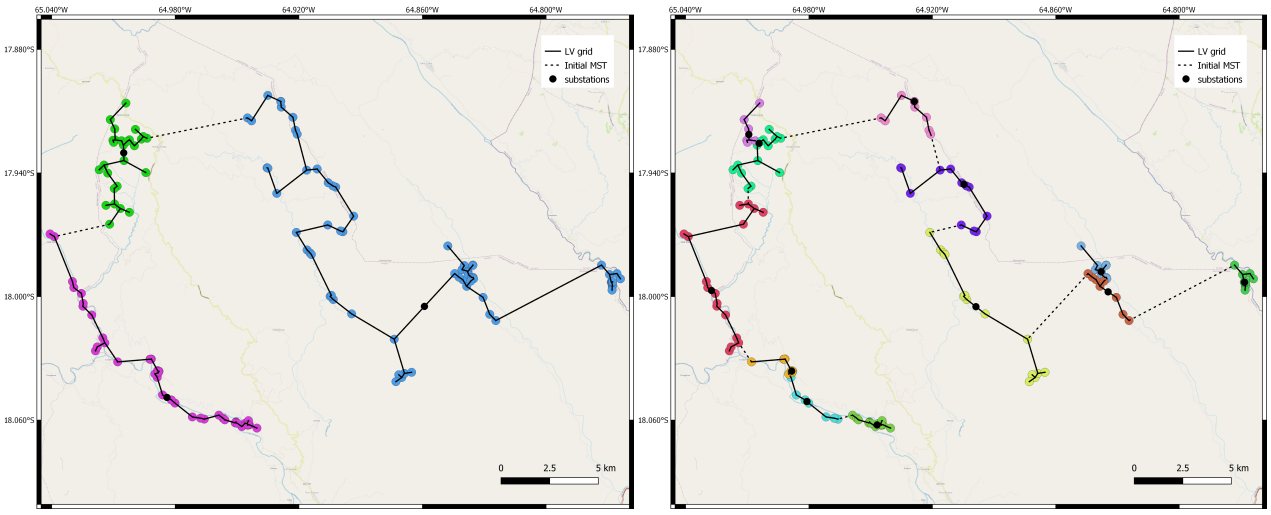


Figure 8. Results of Lukes clustering in Namanjavira.



(**a**) Substation sizes: 100 kW, 50 kW, 20 kW (**b**) Substation sizes: 20kW

Figure 9. Results of Lukes clustering procedure in Omereque.

Preliminary Cost Analysis

As final remarks, a preliminary cost analysis has been performed, to compare the effectiveness of the algorithms in minimizing the total investment cost, given by the sum of substations cost and power lines. For each of the algorithms a minimum spanning tree connecting the nodes of the clusters has been computed. The length of the MST is a proxy for the total LV feeders' length and allows having an estimation of the investment cost for power lines. The number of clusters, and hence of secondary substations allows computing the total cost for this equipment. The LV feeders' cost is assumed to be 6 k$/km in Bolivia and 15 k$/km in Mozambique, according to the data from NGOs and the MV/LV substations' cost is a function of their nominal power as reported in Table 4. The cost

analysis for substations is just based on the tranformers' CAPEX, supposing installation on poles. This solution is considered relevant to the rural areas electrification problem under investigation, given the limited unit power of those transformers and the need to minimize the economic costs. The results are reported in Table 5.

Table 4. Transformers' costs.

P nom (kVA)	Namanj. Cost (k$)	Omereque Cost (k$)
20	3.75	2.5
50	5.7	3.8
63	6.75	4.5
100	10.2	6.8
160	10.8	7.2
200	11.55	7.7
250	12.15	8.1
315	13.95	9.3

Table 5. Preliminary cost analysis.

	Namanjavira				**Omereque**			
	M1	M2	M3	M4	M1	M2	M3	M4
LV feeders length (km)	281	270	271	262	63	23	24	65
N substations (m)	37	100	58	49	10	47	42	12
LV Feeders cost (k$)	4215	4050	4065	3930	378	138	144	394
Substations cost (k$)	240	473	309	283	32	119	108	35
TOT cost (k$)	4455	4523	4374	4213	410	257	252	429

In Omereque the minimum cost is provided by M3, while in Namanjavira by M4. Such a result is coherent with the one reported in Table 3: M4 performs properly for dense populated areas (where the feeder maximum length boundary is not active), whilst it is not capable to manage scattered ones. Omereque has a population so sparse that the initial LUKES tree partitioning, even with a cluster limit size W as low as 20 kW, finds geographically extended clusters and, therefore, very long lines. Once again, this showcases how it is difficult to strike a balance between a cluster size high enough to justify a substation and an acceptable geographical span in the limit case of very sparse population. Clearly the relative importance between these aspects depends on the actual costs of the components, that may be influenced not only by the country standards but also by the actual path of feeders, that may be required to follow roads or even be buried underground. Considering both technical and cost analysis, M3 performs much better than M1 and slightly better than M2, whilst M4 results not technical viable, consequently, M3 results as the algorithm that could better be applied for the problem.

Finally, the costs for the MV distribution grid as well as operational cost should also be taken into consideration from a proper cost analysis, for instance losses along lines and in the transformers as well as maintenance and replacements could influence the optimal solution. Those considerations could be taken into account in a second step, when the actual grid routing will be performed, though losing the global optimality of the solution, an intrinsic limit of the proposed procedure.

8. Conclusions

In this paper we explored four methods for addressing the problem of greenfield planning of the electrification of rural areas. The methods have been applied on two different real case studies and compared according to different metrics. The K-means algorithm, the most commonly found in the literature, did not prove to be effective in finding the optimal site for secondary substations, being the solution dependent on the arbitrarily chosen number of transformers. The agglomerative clustering algorithm with

distance threshold limit proved to be very interesting for the case in analysis both from a technical and from a cost perspective, minimizing the number of total clusters while respecting radius limits. While it seems effective for rural electrification planning, it would be less applicable for urban planning, where high power density requires different approaches. LUKES, a clustering method based on total cluster power, struggled to find an acceptable solution in terms of cluster radius in the very sparsely populated Omereque, while it showed better result in the denser Namanjavira. LUKES seems to be more appropriate to the design of low voltage grids in urban areas, where its potential of following preferential paths such as roads could be exploited. Further research work will be carried out in this sense.

The solutions proposed show direct applicability potential as tools for the sketch design of a network in areas to be electrified, either as a first iteration to speed up the final planning or as a method for a rough evaluation of costs and line lengths involved. The most important output is the secondary substation siting, which in the authors' opinion can be used directly as a stand-alone starting point for routing and sizing a LV network with other methods ([39]).

To conclude, the comparison of different clustering methods has been valuable for identifying limits and strengths of each of them. Although it has been found that it is very challenging to conjugate an acceptable geographical span of the area served by a single substation with a substation power high enough to justify the installation, the methods constitute a significant improvement to the classic K-means both from a technical and economic perspective. The short computational time and simplicity of the algorithms, especially agglomerative clustering, make them suited to be included in wider projects.

Author Contributions: Conceptualization, S.C. and M.M. ; methodology, F.R. and S.C.; validation, C.O.D. and F.R. and S.C.; writing—original draft preparation, C.O.D., S.C. and F.R.; writing—review and editing, S.C. and F.R.; supervision, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: Silvia Corigliano is partially funded in her research activities by Enel Foundation.

Institutional Review Board Statement: Research did not involve humans.

Informed Consent Statement: Research did not involve humans.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: Some of the data useful to perform analysis related to Mozambique study case were provided by COSV. Data related to Bolivia study case were provided by the ONG Luces Nuevas. Part of Federico Rosato's activity was carried out in the frame of the Swiss Centre for Competence in Energy Research on the Future Swiss Electrical Infrastructure (SCCER-FURIES) with the financial support of the Swiss Innovation Agency (Innosuisse—SCCER program).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IEA. SDG7: Data and Projections, IEA, Paris. 2020. Available online: https://www.iea.org/reports/sdg7-data-and-projections (accessed on 20 December 2020).
2. World Bank, Sustainable Energy for All (SE4ALL) Database from the SE4ALL Global Tracking Framework. Available online: https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS (accessed on 20 December 2020).
3. Berizzi, A.; Delfanti, M.; Falabretti, D.; Mandelli, S.; Merlo, M. Electrification Processes in Developing Countries: Grid Expansion, Microgrids, and Regulatory Framework. *Proc. IEEE* **2019**, *107*, 1981–1994. [CrossRef]
4. Willis, H.L. *Power Distribution Planning Reference Book*; CRC Press: Hoboken, NJ, USA, 2004.
5. Georgilakis, P.S.; Hatziargyriou, N.D. A review of power distribution planning in the modern power systems era: Models, methods and future research. *Electr. Power Syst. Res.* **2015**, *31*, 89–100. [CrossRef]
6. Jordehi, A.R. Optimisation of electric distribution systems: A review. *Renew. Sustain. Energy Rev.* **2015**, *51*, 1088–1100. [CrossRef]
7. Corigliano, S.; Carnovali, T.; Edeme, D.; Merlo, M. Holistic geospatial data-based procedure for electric network design and least-cost energy strategy. *Energy Sustain. Dev.* **2020**, *58*, 1–15. [CrossRef]

8. Hongwei, D.; Yixin, Y.; Chunhua, H.; Chengshan, W.; Shaoyun, G.; Jian, X.; Yi, Z.; Rui, X. Optimal planning of distribution substation locations and sizes - Model and algorithm. *Int. J. Electr. Power Energy Syst.* **1996**, *18*, 353–357. [CrossRef]
9. Sun, D.I.; Farris, D.R.; Cote, P.J.; Shoults, R.R.; Chen, M.S. Optimal distribution substation and primary feeder planning via the fixed charge network formulation. *IEEE Trans. Power Appar. Syst.* **1982**, *PAS-101*, 602–609. [CrossRef]
10. El-Kady, M.A. Computer-aided planning of distribution substation and primary feeders. *IEEE Trans. Power Appar. Syst.* **1984**, *PAS-103*, 1183–1189. [CrossRef]
11. Lin, W.M.; Tsay, M.T.; Wu, S.W. Application of geographic information system for substation and feeder planning. *Int. J. Electr. Power Energy Syst.* **1996**, *18*, 175–183. [CrossRef]
12. Crawford, D.M. A mathematical optimization technique for locating and sizing dxstribution substations, and deriving their optimal service areas. *IEEE Trans. Power Appar. Syst.* **1975**, *PAS-94*, 2–7.
13. Díaz-Dorado, E.; Miguez, E.; Cidrás, J. Design of large rural low-voltage networks using dynamic programming optimization. *IEEE Trans. Power Syst.* **2001**, *16*, 898–903. [CrossRef]
14. Cabrera-Celi, G.C.; Novoa-Guaman, E.G.; Vasquez-Miranda, P.F. Design of secondary circuits of distribution networks using clustering and shortest path algorithms. In Proceedings of the 2017 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America), São Paulo, Brazil, 20–22 September 2017; pp. 1–6.
15. Cossi, A.M.; Romero, R.; Mantovani, J.R.S. Planning and projects of secondary electric power distribution systems. *IEEE Trans. Power Syst.* **2009**, *24*, 1599–1608. [CrossRef]
16. Esmaeeli, M.; Kazemi, A.; Shayanfar, H.A.; Haghifam, M.R. Sizing and placement of distribution substations considering optimal loading of transformers. *Int. Trans. Electr. Energy Syst.* **2015**, *25*, 2897–2908.
17. El-Fouly, T.; Zeineldin, H.; El-Saadany, E.; Salama, M. A new optimization model for distribution substation siting, sizing, and timing. *Int. J. Electr. Power Energy Syst.* **2008**, *30*, 308–315. [CrossRef]
18. Mendoza, J.E.; López, M.E.; Pena, H.E.; Labra, D.A. Low voltage distribution optimization: Site, quantity and size of distribution transformers. *Electr. Power Syst. Res.* **2012**, *91*, 52–60. [CrossRef]
19. Najafi, S.; Gholizadeh, R. On optimal sizing, siting and timing of distribution substations. In Proceedings of the 18th Electric Power Distribution Conference, Kermanshah, Iran, 30 April–1 May 2013; pp. 1–6.
20. Hasan, I.J.; Gan, C.K.; Shamshiri, M.; Ab Ghani, M.R.; Omar, R.B. Optimum feeder routing and distribution substation placement and sizing using PSO and MST. *Indian J. Sci. Technol* **2014**, *7*, 1682–1689. [CrossRef]
21. Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
22. Yu, L.; Shi, D.; Guo, X.; Jiang, Z.; Xu, G.; Jian, G.; Lei, J.; Jing, C. An efficient substation placement and sizing strategy based on GIS using semi-supervised learning. *CSEE J. Power Energy Syst.* **2018**, *4*, 371–379. [CrossRef]
23. González-Sotres, L.; Domingo, C.M.; Sánchez-Miralles, Á.; Miró, M.A. Large-scale MV/LV transformer substation planning considering network costs and flexible area decomposition. *IEEE Trans. Power Deliv.* **2013**, *28*, 2245–2253.
24. Navarro, A.; Rudnick, H. Large-scale distribution planning—Part II: Macro-optimization with voronoi's diagram and tabu search. *IEEE Trans. Power Syst.* **2009**, *24*, 752–758. [CrossRef]
25. Wang, S.; Lu, Z.; Ge, S.; Wang, C. An improved substation locating and sizing method based on the weighted voronoi diagram and the transportation model. *J. Appl. Math.* **2014**, *9*. [CrossRef]
26. Bhatia, M.; Angelou, N. *Beyond Connections: Energy Access Redefined*; World Bank: Washington, DC, USA, 2015.
27. Riva, F.; Ahlborg, H.; Hartvigsson, E.; Pachauri, S.; Colombo, E. Electricity access and rural development: Review of complex socio-economic dynamics and causal diagrams for more appropriate energy modelling. *Energy Sustain. Dev.* **2018**, *43*, 203–223.
28. Riva, F.; Sanvito, F.D.; Tonini, F.T.; Colombo, E.; Colombelli, F. Modelling long-term electricity load demand for rural electrification planning. In Proceedings of the 2019 IEEE Milan PowerTech, Milan, Italy, 23–27 June 2019; pp. 1–6.
29. Ciller, P.; Lumbreras, S. Electricity for all: The contribution of large-scale planning tools to the energy-access problem. *Renew. Sustain. Energy Rev.* **2020**, *120*, 109624. [CrossRef]
30. Gan, G.; Ma, C.; Wu, J. *Data Clustering: Theory, Algorithms, and Applications*; SIAM: Philadelphia, PA, USA, 2020.
31. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; University of California Press: Oakland, CA, USA, 1967; Volume 1, pp. 281–297.
32. Kodinariya, T.M.; Makwana, P.R. Review on determining number of Cluster in K-Means Clustering. *Int. J.* **2013**, *1*, 90–95.
33. Chiang, M.M.T.; Mirkin, B. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *J. Classif.* **2010**, *27*, 3–40. [CrossRef]
34. Vahedi, S.; Banejad, M.; Assili, M. Optimal location, sizing and allocation of subtransmission substations using K-means algorithm. In Proceedings of the 2015 IEEE Power & Energy Society General Meeting, Denver, Colorado, 26–30 July 2015; pp. 1–5.
35. Vinothkumar, K.; Selvan, M. Hierarchical agglomerative clustering algorithm method for distributed generation planning. *Int. J. Electr. Power Energy Syst.* **2014**, *56*, 259–269. [CrossRef]
36. Lukes, J.A. Efficient Algorithm for the Partitioning of Trees. *IBM J. Res. Dev.* **1974**, *18*, 217–224. [CrossRef]
37. Graham, R.L.; Hell, P. On the history of the minimum spanning tree problem. *Ann. Hist. Comput.* **1985**, *7*, 43–57. [CrossRef]

38. Facebook Connectivity Lab and Center for International Earth Science Information Network–CIESIN–Columbia University, High Resolution Settlement Layer (HRSL)©2016 DigitalGlobe, 2016. Available online: https://ciesin.columbia.edu/data/hrsl/ (accessed on 21 December 2018).
39. Al-Jaafreh, M.A.; Mokryani, G. Planning and operation of LV distribution networks: A comprehensive review. *IET Energy Syst. Integr.* **2019**, *1*, 133–146. [CrossRef]