POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Energy Engineering

GISEle: an innovative GIS-based approach for electric networks routing

Relatore: Prof. Marco Merlo
Correlatrice: PhD candidate Silvia Corigliano

Tesi di Laurea Magistrale di:
Darlain Irenée EDEME Matricola 878237
Tommaso CARNOVALI Matricola 878712

Anno Accademico 2018-2019

Extended Abstract

Nomenclature

Symbol	*Description*
T	Expected lifetime
t	Year of life
C_t	Capital expenditure
$O\&M_t$	Operation and maintenance expenditure
F_t	Fuel expenditure
E_t	Electrical energy generated
r	Discount rate
LCOE	Levelized Cost of Energy
n	Number of data points
m	Number of attributes
l	Number of iteration for convergence
m_a	Maximum number of neighbor for a point
m_a	Average number of neighbors

Introduction

Energy is one of the fundamental pillars for the development of human society. Nevertheless, still nowadays a relevant portion of world population cannot rely on modern energy sources, and 1 billion people live without electricity access. The United Nations defined the 7^{th} SDG, with which they committed to *"ensure access to affordable, reliable and modern energy fuel for all"*. Therefore, governments, international organizations, private industries, investment funds and NGOs cooperate towards improvement of existing energy systems were available and in creating new ones were access is not guaranteed yet.

Even in low developed countries, urban areas have often relatively high electrification rate, therefore electrification projects are mainly focused in rural areas where, especially in the African context, access to electric energy usually lacks at all. The available strategies to distribute electricity in these locations are different:

- National grid expansion: it is a national governments' duty, can ensure low specific energy cost but when dealing with extremely remote localities implies huge capital investments;

- Integrated micro-grids

- Off-grid systems: usually rely on locally available renewable resources which are still considerably expensive

A proper evaluation methodology to choose the best one is fundamental to support decision-makers in planning efficient and effective investments. The evaluation process should cover each phase of the planning work:

- **Territory analysis** in which area-specific morphological, social and technical characteristics are accounted;
- **Load estimation** of communities' energy needs;
- **Generation sizing**: identification of best technologies combination to be installed in order to fulfill the energy demand;
- **Electric network routing**: planning of electric grid optimal topology reaching the overall energy communities'

consumers.

The techno-economic feasibility of a project could be quantifies through the Levelized Cost of Energy (LCOE), defined as the average minimum price at which the electricity must be sold in order to break even over the project lifetime. It allows direct comparison of concurrent solutions.

$$LCOE = \frac{\sum_{t=1}^{T} \frac{C_t + OeM_t + F_t}{(1+r)^t}}{\sum_{t=1}^{T} \frac{E_t}{(1+r)^t}} \qquad (1)$$

The chosen final solution is generally the one guaranteeing the lowest LCOE, provided that eventual superimposed constraints on pollutants, carbon emissions and minimum renewable fraction are respected.

So far, several approaches and tools have been developed and applied. They often have one critical drawback: neglecting the importance that spatial information (like population distribution and target area morphology) cover in an evaluation which is expected to be as much comprehensive as possible. These data are essential for the electric networks design and their costs estimation which, when neglected, lead to systematical capital investments miscalculations. With the present thesis work the authors aim at building up a geospatial data-based procedure, overcoming these criticalities by autonomously performing electric network design processes and optimal energy strategy planning.

To rationalize investments and enhance their efficiency, discriminating between highly populated areas suitable to become energy communities, and too sparsely distributed households to be connected, it is a paramount step that has to be accomplished by means of a spatial clustering process. Thereafter, once the communities' social composition and geographical extension have been defined, the procedure should be able to perform the community-specific energy needs assessments, and design the electric network reaching all the single community consumers. The linking infrastructure between the community and the existing HV national grid cost must be investigated and compared to the standalone configuration net present value. Finally, proper power supply systems satisfying local energy needs have to be techno-economically sized, and a final economic comparison between the different possible strategies has to be done through LCOE minimization. Stated the primary role that geo-referencing plays in such a methodology, GIS result to be the most appropriate data structures and developing environment on which implement the proposed methodology. The procedure developed within this thesis project, called *GISEle*, tested and validated in the rural province of Namanjavira, Mozambique, pointing out the evolution and innovation with respect to the other tools available in the literature.

State of the Art - Algorithms and Models

Due to the wide variety of activities that *GISEle* performs, it should be capable of handling different problems ranging from terrain analysis to spatial clustering, operational research and power system sizing, each one with its own background of research activities.

GIS Geographic Information Systems (**GIS**) are designed to store, manage, and visualize spatial or geographic data, integrating a range of geographical features into a single analytical model in which data are geo-referenced to cartographic projection. Actually, every dataset equipped with spatial attribute can be modelled into GIS environments with different available formats: *rasters* are used to display continuous information across an area that cannot be easily divided into single features, like solar irradiance or ground elevation; *vectors*, instead, provide a way to represent discrete features like rivers, houses or roads. Combining data together, it becomes possible to characterize each point of the region in analysis. Sumic et al. [1] identified GIS as the proper computer platform to develop automated routing of underground residential distribution system; Monteiro et al. [2] developed a GIS spatial methodology for a simple point-to-point overhead economic corridor selection for new power lines. In doing that, they considered several aspects affecting the realization costs of an electric line: distance from the road network, ground slope, land-cover type, natural obstacles presence. Monteiro et al. [2] exploit these information to quantify the global capital expenditure needed to build a unit length of electric line across a specific land piece. *OnSSET* [3], one of the available tools for electrification strategy planning, instead, combines such data defining a *"penalty factor"* which is then applied to the standard line cost.

Data collection Data collected through satellites, can be easily gathered, even if with a usually rough resolution. Instead, data requiring on-site measurements or surveys which are generally highly time and resources-consuming are not always available, especially when dealing to developing countries. For this reason, the majority of tools devoted to electrification planning consider sun and wind within the available energy sources (GHI and wind speed data are shared for free by NASA or other governmental agencies), whereas they discard hydropower.

Direct punctual measurements of hydro power potential cannot be made on wide regions without devoting huge amount of money and time, however the *US Soil Conservation Service*[4] provides a methodology based on empirical correlations to estimate surface water runoff: the *CN-method.* It has been developed to predict ground response to single rainfall events, analyzing land cover, soil type, rainfall and evapotranspiration rate, but several studies exploit it also to estimate monthly flow rate and the hydropower potential [5], [6], [7].

Clustering As anticipated the best densely populated areas identification process is done through clustering techniques. Cluster analysis refers to a statistical process with which a population of data is grouped in sub-groups, or clusters, such that each element of each cluster has significantly more affinity with the other members of its sub-group compared with the other clusters' ones. Nowadays clustering indicates a relatively vast portfolio of models, relying on different approaches to the data classification problem leading to many structured algorithms. The reason there are several approaches is related to the fact that each problem has its peculiarity and each set of data needs to be addressed in the most suitable way. For this reason, it is important to note that when applying clustering to a population of data, having substantial knowledge of the characteristics of that dataset is an unavoidable precondition. Cluster analysis models can be mainly grouped as proposed by:

- **Density-based** Clustering Algorithms

The main problem with data mining algorithms in general, and particularly with clustering algorithms, is their scalability. Highly complex problems generally imply large-sized data frames, and unfortunately, each algorithm can easily handle specific order of magnitude of data.

Grid routing Tracking the optimal topology of the electric networks, providing an estimate on its capital cost too, is the main goal of this work's proposed approach. Identifying the cheapest way to reach the energy users within a unique energy system and actually drawing the corresponding network is an optimization problem belonging to the mathematical branch of operational research, devoted to find the optimal or near-optimal solutions to complex decision-making problems. Specifically, such problem is part of *Shortest Path Problems* defined as the search of the path connecting a set of defined nodes in a connected and weighted graph, such that the sum of its constituent edges' weight is minimized. The associated literature is rich since it is, for instance, the base of satellite navigator's logic.

A *graph* $G = (V, E)$ consists in a set of vertexes V and of edges E connecting couples of nodes and representing their relationship. An example is depicted in figure 1.

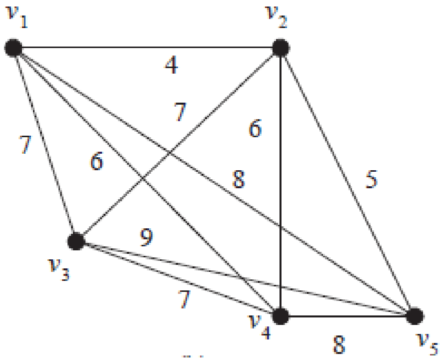


Figure 1: Example of weighted, not oriented graph

A *tree* is a graph structure in which ∀ couple of vertexes $u \neq v \in G$ there is exactly one single path from u to v. A **Minimum Spanning Tree (MST)** of a graph (figure 2) is a subset of the edges that connects all the vertices together, without any cycles and with the minimum possible total edge weight. The cost of the tree can be defined as the sum of its edges weight.

$$w(T) = \sum_{e \in T} w(T) \qquad (2)$$

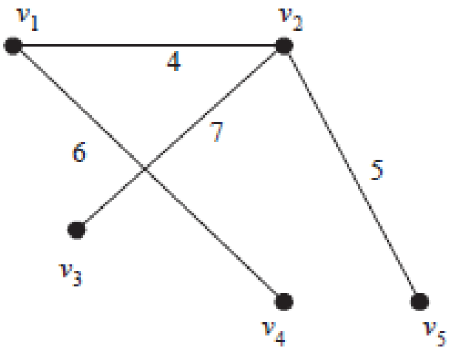


Figure 2: Minimum Spanning Tree of the graph in figure 1

This can be an initial rough way to sketch the electric network, and two main algorithms are available to solve the problem: *Prim* and *Kruskal*. Nevertheless, the solution returned is still rough. MST produces grids in which the connections between terminals are straight, based on aerial-distance, ignoring any obstacle that might be in the middle. Laying an electric line across a virgin jungle or a swampland, unless it is the shortest path, may result much more difficult and expensive than follows the road around it. *Dijkstra* provides a partial solution to this issue: its algorithm, finding the shortest path between a source nodes and all the others, allow to determine the *shortest path* connecting two terminals within a dense weighted graph. Considering a path π_{uv} as the sequence of edges (and consequently of nodes) connecting a source u to a target node v, the *shortest path* is defined as the path π^*_{uv} that has a cost w_{uv} less than or equal to that of any other path π_{uv} between the same vertexes, so that:

$$w(\pi^*_{uv}) = min_{\pi_{uv} \subseteq G} w(\pi_{uv}) \qquad (3)$$

The **distance** d_{uv} between two nodes in G, is defined as the cost of of minimum path which connect them, or $+\infty$ if no path exists between them:

$$d_{uv} = \begin{cases} w(\pi^*_{uv}), & \text{if path exists} \\ +\infty, & \text{if not} \end{cases} \tag{4}$$

Dijkstra algorithm initially makes a very high estimate of the minimum distance between each pair of points in the graph $D_{uv} \geq d_{uv}$. Each time that a shorter path between u and v is found, the distance value is updated until the estimate becomes accurate, that is $D_{uv} = d_{uv}$.

The time required to calculate the distances from the source s and the target t of the graph is in the worst case O(m+n log n) as demonstrated by Demetrescu et al. [8].

Dijkstra's algorithm provides the correct optimal solution, but, as drawback, can be applied only to a single couple of terminal points.

When the number of target vertexes increases, as in the case of electric grid loads, the problem takes the name of *Steiner tree problem.* The objective is the same: "to find the least cost path", but with a higher mathematical complexity. This problems typology are classified as NP-completed and the solution can only be approximated. Despite its complexity, this is the best representation of the real problem. The nodes, divided in *terminals* and *non-terminals* represent respectively the nodes which must be included in the solution, like cluster's consumers, and the ones constituting the territory framework with its own characteristics. An example of Steiner tree is shown in figure 3. The several algorithms developed to address the problem require a consistent amount of memory, therefore, are suitable only for small graphs involving a limited amount of terminals.

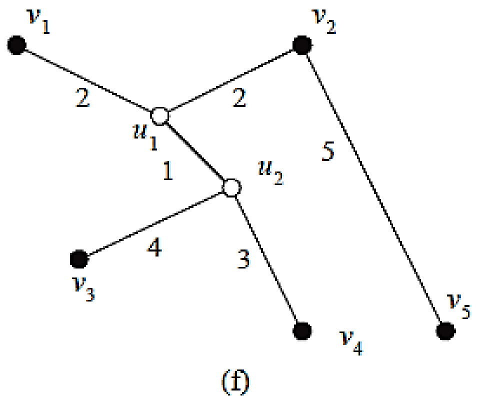


Figure 3: Steiner tree with black terminal nodes and white non-terminal ones

Tools

Currently, several tools are available on the market helping a broad-spectrum of stakeholders planning energy electrification strategies [9]. Each of them addresses one or multiple issues in energy planning in rural areas of the developing world, through multidimensional perspectives. They have been developed by different entities (Massachusetts Institute of Technology, Royal Institute of Technology in Stockholm, Economic Community of West African States, Politecnico di Milano and National Renewable Energy Laboratory among the many), with diametrically opposed characteristics, for instance:

- addressed to public or private sector
- proprietary or open source license
- top-down or bottom-up approach
- regional-specific or global dimension

GISEle is the result of an accurate analysis through which we defined which were the important gaps missing in this tools portfolio.

Proposed approach

The procedure proposed with this thesis work, herein called *GISEle*, is rooted on the analysis of spatial data and runs through all the passages leading to the identification of the optimal techno-economic solution to bring the energy where it lacks. In addition to the technical configuration, exploiting the potentialities of GIS environment, GISEle is able to consider also solution's spatial characteristics:

- the spatial distribution of consumers and generation plants;
- the detailed topology of the electric grid which would connect them together and, optionally, to the existing national grid.

Those aspects allow GISEle to supply accurate evaluations about the final costs of different possible solutions. The overall *GISEle's* structure is presented in figure [**?**].

Data gathering The procedure starts with the collection of all the data needed for the analysis from internet platforms. The hydro-power potential is the only lacking dataset which is not directly available, but since *GISEle* includes it within the available energy sources, a specific methodology has been developed to estimate it through indirect measurement. Performing a terrain hydrological analysis combined with the CN-method, it provides both monthly mean water flow rate and available head, and the hydro-power potential is estimated with:

$$P = Q * \alpha * \rho * g * \Delta h * \eta_{power-plant} \quad (5)$$

Because of the complexity of phenomenons involved, the obtained values of water flow are imperfect, therefore the model has been validated comparing the results with punctual real measurements data in order to define a corrective factor α.

Once all the necessary datasets have been defined, they can be combined together in a discrete representation of the target region: a python routine, executable in the python environment of QGIS, has been written at this purpose. Each point of the resulting matrix will embed all the geographical characteristics relative to its location. The information gathered and embedded in each node, constitutes a set of the spatial aspects and ground characteristics which impact costs of an electric line realization. Due to the high cost-variability country by country, the concept of a "penalty factor" applied to a base-cost of electric line has been preferred rather than providing an absolute value to each single voice impacting on it. The following contributes have been considered: *distance from road, slope, land cover, lakes, river flow rates, protected areas.* The *"penalty factor"* is calculated as:

$$Penalty\ factor = 1 + \sum_{i}^{penalty\ aspects} penalty_i \quad (6)$$

A penalty factor map of the Namanajavira province is provided in figure 6.

Clustering The main objective of the clustering process is to move beyond the basic approach with which each cell is considered by itself, neglecting the strategic value of closely located highly populated areas: the algorithm will proceed by identifying valuable groups of cells instead of only one cell at a time.

The population clustering, moreover, is a fundamental step of the proposed procedure to reduce the computational burden of our model. The routing procedure applied to the whole region of interest, indeed, would require a big amount of time because of the high number of points to be considered; furthermore it would create an electric line connecting each populated point, independently from its location, included the most isolated ones for whom a stand alone solution would be more appropriate. Clustering, solves these issues, by classifying those single isolated points as outliers, and grouping all the others in different clusters which will

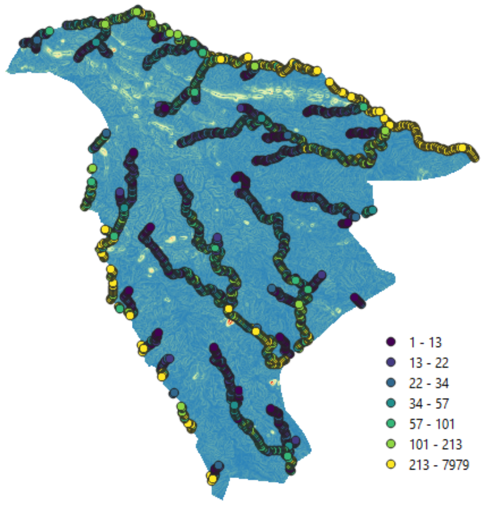


Figure 5: Hydro power potential map in Namanjavira

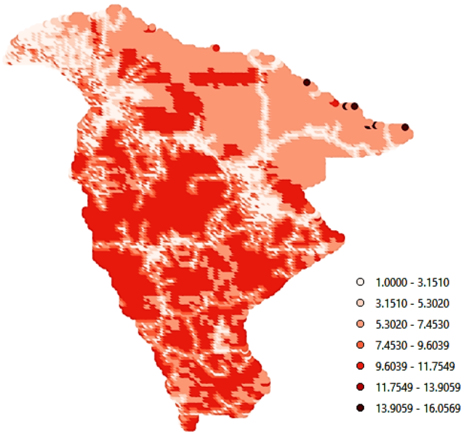


Figure 6: Penalty factor distribution of Namanjavira administrative place

be considered as unique, separated energy communities. The following routing algorithm will than be applied to each cluster, drawing its specific electric grid topology. In this way, the amount of data involved in each routing process, ad so the memory required to manage them, will be limited. Clustering algorithms are fast: generally they outweigh routing algorithms in terms of computational time. Therefore their application speeds up the global procedure.

Accordingly, the implemented clustering algorithm is a density-based, with a particular adaptation based on an important input data property: points do not represent a singular element like in general clustering problems but have a fundamental property which is the population. Thus, DBSCAN algorithm has been selected: a particular adaptation of this algorithm in which the point weighting criterion is precisely its population attribute was implemented. Therefore, the minPts input parameter changes its meaning from "minimum number of points" to be found in a neighborhood to define it as core neighborhood, to "minimum number of people". Firstly, this allows the algorithm to create clusters having the exact communities' extension shape. Additionally, the ability of this algorithm of identifying scarcely populated areas allows to neglect vast zones with few people, prioritizing highly populated ones.

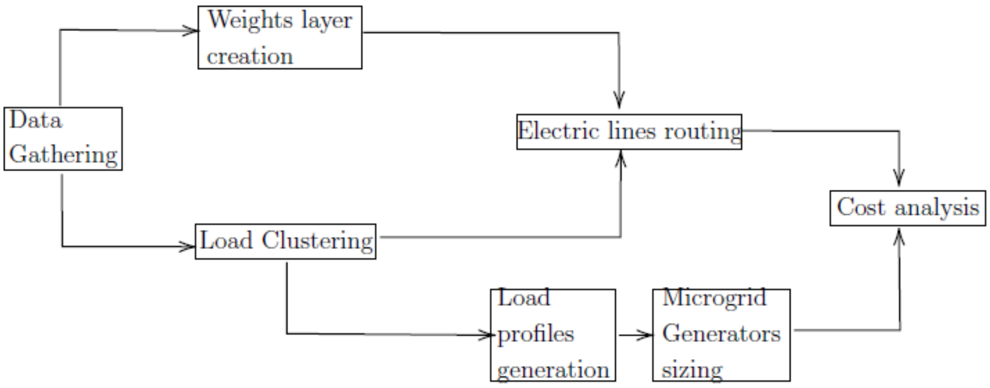


Figure 4: GISEle's procedure structure

DBSCAN's input parameters, eps and minPts, need to be defined. To do so, a sensitivity investigation is performed analyzing each couple of parameters' combinations in terms of:
- number of resulting clusters;
- % of clustered people over the target area total population;
- % of clustered area over the total target area.

Clusters represent independent energy communities for which the optimal electrification strategy needs to be investigated. Once defined the procedure continues with the core step: the electric grid routing.

Grid routing The regular graph where to apply the available *Minimum path algorithm* is built allowing each node to be connected only with its eight neighboring nodes. The cost of the single edge connecting two nodes i and j is defined as:

$$C_{ij} = L_{ij} \cdot \frac{UC}{1000} \cdot \frac{p_i + p_j}{2} \qquad (7)$$

Which considers the distance between the extremes L_{ij}, the cost of a length unit of electric line UC and a mean *"penalty factor"* between those related to the extremes i and j. Every populated point within it (with a value of population greater than 0) becomes part of the set of terminal nodes of a Steiner tree problem whose solution will lead to the definition of the electric line topology with minimum cost. The python package Networkx provides a function for an approximate solution but, due to computational burden, it is suitable only for small graphs with few terminal points. This limits its application to the smallest clusters. To overcome this limit providing a solution even for big clusters the authors developed an original approach combining the potentialities of MST and Dijkstra algorithms. An initial, rough model of the electric connections is initially traced by applying the MST to the overall set of cluster's populated point. An analysis of the MST solution is then performed, making a distinction between *short lines* connecting neighboring points and *long lines.* Short lines are already well defined since they connect neighboring points and their cost can be estimated by equation 7. Therefore they become part of the final tree. With regard to long lines l_{ij}, they need to be fragmented into sequence of adjacent elementary edges. To do so, the terminals i and j are respectively set as source and target for the Dijkstra algorithm which returns the *least cost path* connecting them. During each iteration the algorithm is designed to continuously store the newly created connections updating the corresponding edges cost to a value of zero: this makes it capable of recognizing already built paths which, if crossed, would lead to no additional costs. Once all *long lines* have been fragmented and re-structured the algorithm stops: the final solution is a tree, connecting all the cluster's populated nodes. This tree will only be made of elementary edges, each one with an associated cost, which algebraically summed give the total network cost. Finally, as the final goal is discerning between standalone and grid-connected clusters, the Dijkstra algorithm adopted to manage long lines is applied in order to draw the infrastructure linking each cluster to its closest national grid's substation. Figure 7 present an example of cluster grid equipped with HV-line connection.

Load assessment The energy needs assessment of clusters is based on the definition of reference profiles: in the initialization phase, previously computed reference load profiles are required, i.e. daily load profiles of reference users categories. This approach assumes that geographically close communities share socio-economic analogies in terms of basic energy needs, so in first approxi-

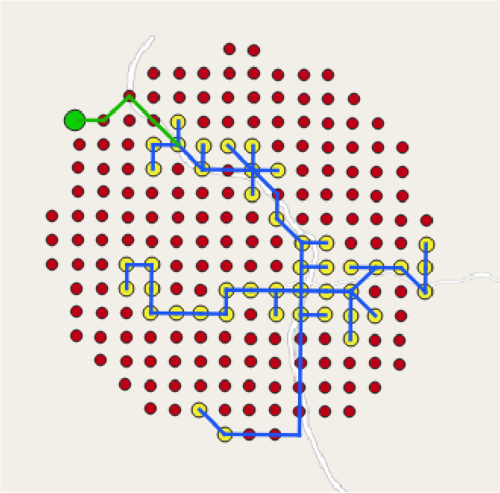


Figure 7: Example of electric grid (blue line) connecting the populated points (yellow) within a cluster (red points). The green line represents the further connection to the closest HV substation (green point).

mation, their load curves will approximately have similar shapes.

A previously computed reference load profile is taken as input,and associated to a specific proxy, the most suitable being the number of households in the community. The reason behind this choice relies on assumption that, in developing countries' rural areas, communities are basically household-centric, meaning that their core is made of rural households.

Considering that specifically developed and highly consolidated power generation sizing tools are already available on the market, this task will at present be outsourced. In particular the utilized tool has been the National Renewable Energy Laboratory's software HOMER (*Hybrid Optimization of Multiple Energy Resources*) Energy Pro. From the proposed power supply system configurations, for each cluster, 3 are selected for the final analysis: 1 - the cheapest one

2 - the cheapest solution providing 100% renewable energy

3 - the one still exploiting fossil fuels, but with the highest fraction of renewable penetration

In Namanjavira, the primary connection route where the majority of people is concentrated, runs along a ridge. In the nearby, therefore, the estimated hydropower potential is very limited. This fact, combined with the high rainfall seasonality, reflects into a low availability level of water resources. Despite commonly being the most cost effective renewable resources in stand alone rural power supply system, in this particular case hydro-driven alternative is included only in high renewable-fraction solution.

techno-economic evaluations The final output delivered by *GISEle* is a cluster specific comparison between the two main possible electrification strategies: isolated micro-grid and grid-connected energy systems. Isolated micro-grids are not connected to the HV national grid, therefore they do not have to sustain the costs related to the siting of the power line linking the cluster's internal grid to the nearest HV substation but require a dedicated power supply system to feed the demand.

$$LCOE_{mg} = \frac{C_{grid}}{total\ energy} + LCOE_{gen} \quad (8)$$

Where: mg = micro-grid; gen = power generation. The alternative to connect the cluster network directly to the national HV transmission line gives access to electric energy at lower cost, but requires realizing the further electric line linking the cluster to the closest substation.

$$LCOE_{HV} = \frac{(C_{grid} + C_{con})}{total\ energy} + LCOE_{NG} \quad (9)$$

(NG: National energy price)

Finally, the two alternatives are compared and the optimal strategy is defined.

Application and results

The proposed procedure has been implemented, tested and validated on a real-life case study in the Namanjavira province, Mozambique, selected since one of the authors is currently working on an electrification project in this area. A collaboration has been set up with the implementing agency, and the data gathering process was somehow more reliable. Furthermore, being the uncertainty in terms of computational burden of the final model relatively high, concentrating on one smaller area gave to the authors the possibility of testing all algorithms' outputs, despite having a 200m spatial resolution, quite high compared to other tools that generally don't go below 1km. Mozambique's rural areas have the particularity of having a population mainly distributed over the few road corridors that cross the country (a detail is shown in 8), the routing algorithm was expected to identify. Thereafter the two DBSCAN's defining parameters eps and minPop were set respectively as 480 meters and 1140 people based on three complementary sensitivity analyses on: outputs' number of clusters, percentage of clustered people and area. This configuration leads to 13 clusters, covering the 13% of the total area with an approximate value of clustered people of 60% (figure 9). The final configuration depicted in figure 9 is the result of a further manual processing in which adjacent clusters has been joined together into a single one. Unselected areas (i.e. percentage of the population not candidated for an electrification process) are supposed to be managed by small equipment like Solar Home Systems. Overlapping these clusters to aerial images from Bing it can be seen how the algorithm is able to identify zones with high number of houses (10). Figure 11 shows the internal population density distribution in cluster n.2: it is intuitive to see the differences between densely populated and border areas. An ulterior validation is the good superposition level of many clusters over the existing formally recognized communities. For instance, clusters n. 0, 2, 5, 6 and 12 (figure 10) shows how, despite being not formalized yet, many remote rural areas have experimented relatively high population growth rates and have now population comparable to other recognized ones.

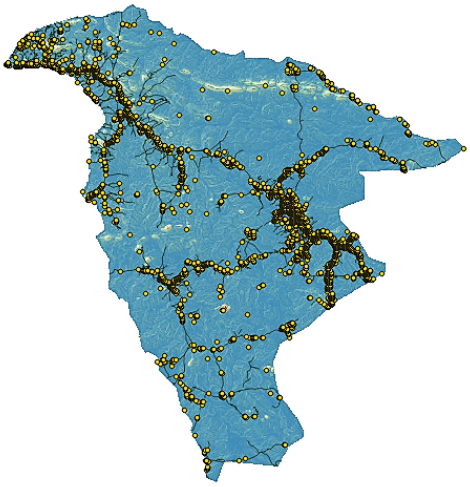


Figure 8: Populated points distribution along road network

Initially, the grid routing algorithm has been applied on the whole Namajavira, without accomplish any clustering process including in the final grid all the points with more than 10 people. The output depicted in figure 12 shows a wide grid: long lines has been traced throughout the territory in order to connect even the remotest points. Such long lines allow to really appreciate the ability of the routing algorithm to find not the shortest, but the least cost path connecting two terminals (figure 13). Applying the algorithm only to limited areas within the clusters' borders, instead, the final topology is completely different (figure 14). Single long connection in remote areas are no more clearly identifiable, since all the too isolated populated nodes has been discarded by the clustering process which classified them as outliers and could be considered as candi-

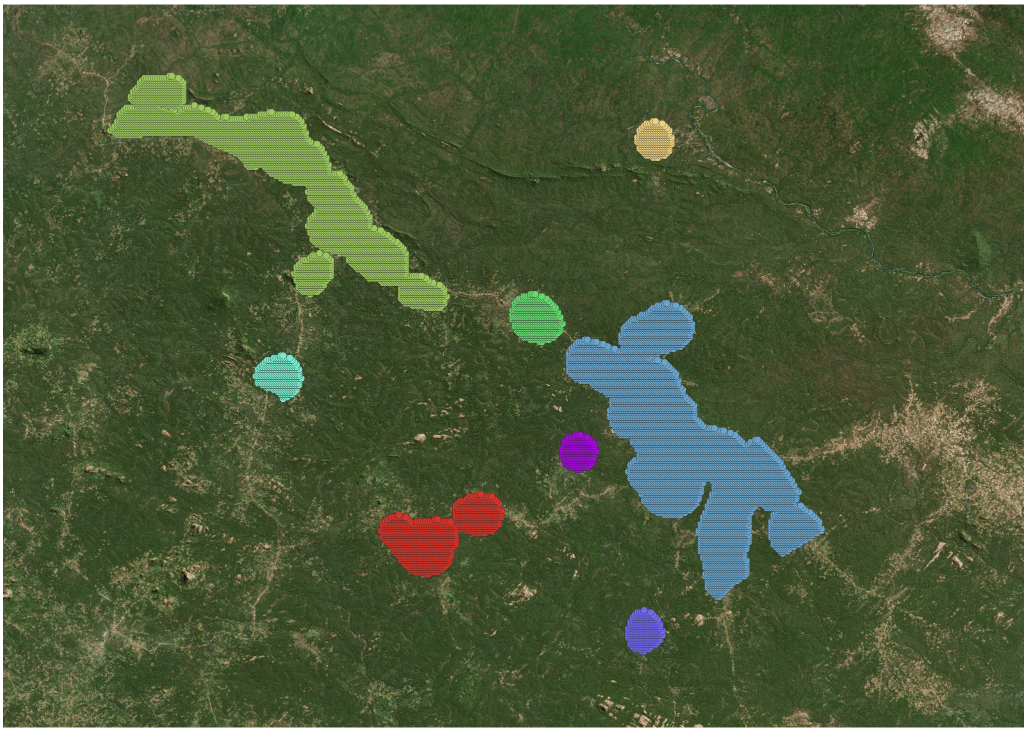


Figure 9: Clustering process output

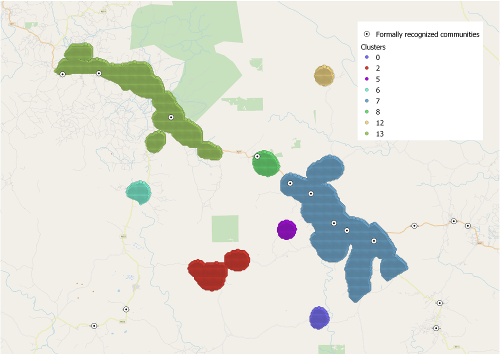


Figure 10: Clusters over formally recognized villages

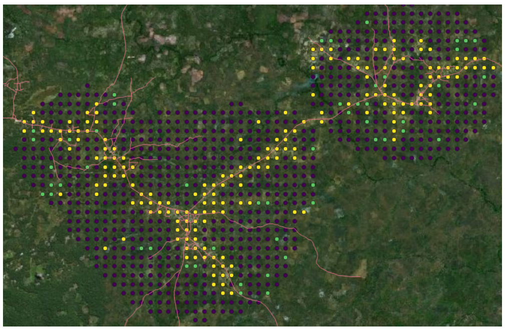


Figure 11: Cluster 2 internal population density distribution

date for stand-alone energy systems. The entire southern part of the region and the main portion of northern one, remains unelectrified. Furthermore the limited area of a cluster reduces the global number of points that the grid algorithm must manage so allowing to produce more detailed solutions. In this analysis all the clusters' nodes with population value greater than 5 (the average number of people in an household) have been connected to the final grid.

A summary of the techno-economic analysis performed for each cluster is produced ending with a comparison between LCOE values of the two covered strategies: *HV line connection* and *Isolated Micro-grid* depicted in table 1. As expected, the high energy demand of big clusters always justify the connection to the national grid, independently from its distance. With regard to small clusters, instead, the HV connection results cost effective only when really close to a substation, or if low-cost routing corridors can be

$/kWh						
	SOL_1	SOL_2	SOL_3	HV	SOL_3 vs HV	Suggestion
0	1,06596101	1,65856101	1,23156101	2,20641017	-44,18%	SOL_3
3	1,082754086	1,558654086	1,193554086	0,949602549	25,69%	HV
4	1,095637782	1,694837782	1,254937782	0,717694925	74,86%	HV
5	1,40069968	2,04879968	1,62879968	1,474287348	10,48%	HV/SOL_1
6	1,440921303	2,160321303	1,620321303	2,069372963	-21,70%	SOL_3
7	0,843332913	1,628532913	0,844532913	0,513102535	64,59%	HV
8	1,496454942	2,323454942	1,607954942	1,228991736	30,84%	HV
12	1,312093482	1,865993482	1,476993482	0,950283549	55,43%	HV
13	0,819433373	1,629933373	0,874033373	0,484072761	80,56%	HV

Table 1: LCOE of each proposed solution for each cluster

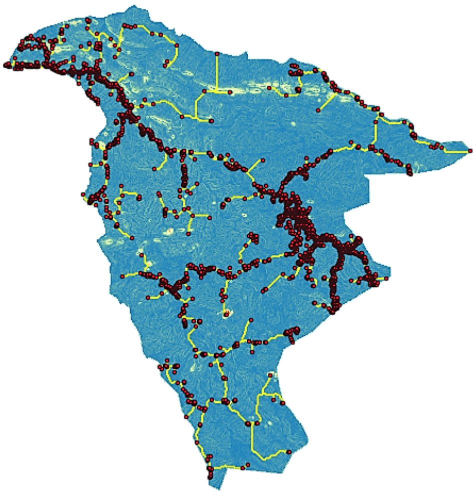


Figure 12: Electric network covering the whole Namanjavira

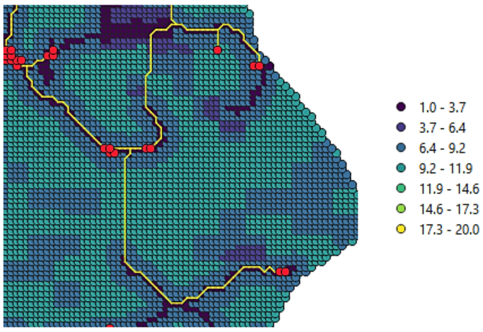


Figure 13: Long lines connections over penalty factor layer

exploited for link realization. The internal grid cost analysis on the overall LCOE reveals it impacts from 50 to 200%, underlying the fundamental importance of an appropriate electric network modelling.

Conclusion

In a global framework where guaranteeing electricity access to all is set to be one of the main challenges that the international community is going to face in the next future, *GISEle* places itself among the available approaches to be adopted when allocating resources and capitals at this purpose.

Spatial dimension of the target areas has a significant impact on rural infrastructures' total cost, therefore, when ignored, the final results can lead to wrong strategy implementations. When discerning between different electrification strategies, not considering electric networks as a fundamental component of energy supply systems is, from the authors' perspective, an error which needs to be avoided, since it leads to wrong evaluation of the LCOEs.

GISEle's goal is to create a methodology capable of having a more holistic approach in rationalizing investments and resources deployment in tackling energy access in the developing world. This procedure is suitable for all kind of stakeholders: international organizations, private sector and governments. Its final framework is composed by several steps. Starting from GIS data, spatial char-

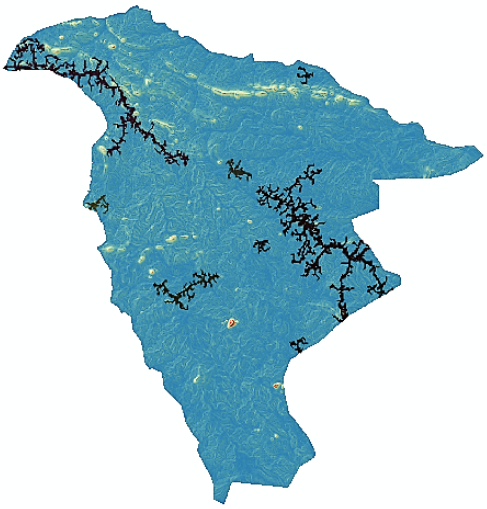


Figure 14: Clusters electric grid

acteristics of the target area are efficiently analyzed in order to extract valuable insights and design the optimal electric network topology. Finally, discrimination between standalone and grid-connected configurations is performed for each single energy community and a comprehensive optimal electrification strategy is suggested.

The fundamental algorithms on which the proposed methodology is based, are specific adaptations introduced in order to optimally fit the addressed problem.

Within GIS data management environment, the total area is segmented into a regular grid of quadratic cell, each one having geo-specific attributes concurring in defining its suitability for infrastructural building. A remarkable achievement is the consistency of a totally open-source data-based procedure. Resources, loads and terrain morphology publicly available datasets have sufficiently high resolutions to obtain valuable results: their significance has been evaluated on field and the outcome is that for the principal ones the reliability is sufficiently high.

The obtained results from GISEle deployment in Namanjavira province are promising and create pre-conditions for the application of the same approach not only for the optimal electric grid routing but also to identify suitable locations for on-grid power plants. Lastly it is worth to emphasize how the Mozambique *Renewable Energy Atlas*, has been an extremely both time and resources-consuming process, which with the help of *GISEle* could be done in a much more efficient way, having as added benefit, much more structured results as effective costs of the energy access, and concrete infrastructure design.

Bibliography

[1] Z Sumic, T Pistorese, and S.s Venkata. Automated underground residential distribution design–Part 1: Conceptual design. *IEEE Transactions on Power Delivery - IEEE TRANS POWER DELIVERY*, 8:637–643, 1993.

[2] Cláudio Monteiro, Ignacio J. Ramírez-Rosado, Vladimiro Miranda, Pedro J. Zorzano-Santamaría, Eduardo García-Garrido, and L. Alfredo Fernández-Jiménez. GIS spatial analysis applied to electric line routing optimization. *IEEE Transactions on Power Delivery*, 20(2 I):934–942, 2005.

[3] http://www.onsset.org/#.

[4] Soil Conservation Service. https://www.nrcs.usda.gov/wps/portal/nrcs/site/.

[5] Gah-Muti Salvanus Yevalla, Dadjeu Nguemeu Seidou, Bang Vu Ngoc, Tran Van Hoi, and Tabod Charles Tabod. Hydrological Studies for the Assessment of Run-of-River Hydropower Potential and Generation over the Wouri-Nkam River using GIS and Remote Sensing Techniques. *Aquademia: Water, Environment and Technology*, 2(1):1–7, 2018.

[6] Henry A. Adornado and Masao Yoshida. GIS-BASED WATERSHED ANALYSIS AND SURFACE RUN-OFF ESTIMATION USING CURVE NUMBER (CN) VALUE. *Journal of Environmental Hydrology*, 18(Paper 9):1–14, 2010.

[7] I. Sahu and A. D. Prasad. ASSESSMENT OF HYDRO POTENTIAL USING INTEGRATED TOOL IN QGIS. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-5(November):115–119, nov 2018.

[8] Camil Demetrescu, Irene Finocchi, and Giuseppe Italiano. *Algoritmi e Strutture Dati*. McGraw-Hill, 2004.

[9] Magda Moner-Girona, Daniel Puig, Yacob Mulugetta, Ioannis Kougias, Jafaru AbdulRahman, and Sándor Szabó. Next generation interactive tool as a backbone for universal access to electricity, 2018.

Contents

Abstract

In a global context where about 1 billion people still relies on traditional energy sources, the 7th United Nations' SDG of *"ensuring access to affordable, reliable and modern energy fuel for all"*, is set to be one of the primary goal of our generation. The attempts made so far to collaborate to the purpose have been often hampered by administrative, social and economic barriers, as well as a poor planning of the strategies to follow, which lead them to low-effectiveness results. The correct identification of the best electrification strategy is a fundamental step to increase energy access in rural areas of developing countries. The available tools developed to support governments and organizations in this task often have the drawbacks to focus only on few specific aspects or technologies, and especially often neglect the importance that spatial information like population distribution and the electric network topology cover in a comprehensive evaluation. *GISEle*, whose development process is depicted in this work, is a new Geographic Information Systems (GIS)-based procedure which addresses these issues covering all the different steps required for a comprehensive electrification strategy evaluation. Spatial information are exploited to identify populated areas present in the analyzed geographical region, and to design the electrical network connecting the clusters' consumers in a single energy community pursuing the minimization of costs. To this end, an innovative minimum-path algorithm has been developed by the authors. It is able to overcome the limits of existing approaches applied to operational research in graph theory, allowing to define the *Least cost Tree* also for graphs systems with many terminal points. As final result, together with grid topology, the complete procedure provides also a suggestion about the optimal energy supply strategy to follow between stand-alone micro and mini-grid and existing national grid extension. The performance of *GISEle* has been tested in the rural province of Namanjavira, Mozambique, highlighting the potentialities and the improvement that a spatial approach can add to the electrification strategy definition. This makes it a supporting instrument suitable both for governments, private sector and international organizations acting in this field.

Sommario

In un contesto globale in cui 1 miliardo di persone fa ancora affidamento su fonti energetiche tradizionali, il settimo SDG sancito dalle Nazioni Unite che si propone di *"assicurare a tutti l'accesso a fonti energetiche sostenibili, affidabili e moderne"*, è destinato ad essere una delle principali sfide della nostra generazione.
I tentativi di cooperazione finora fatti per raggiungere questo scopo si sono spesso scontrati con difficoltà amministrative, sociali ed economiche, così come con una scarsa pianificazione delle strategie da seguire, che li ha condotti a risultati poco efficaci. La corretta indentificazione della migliore strategia di elettrificazione è un passaggio fondamentale per incrementare l'accesso all'energia nelle aree rurali di paesi in via di sviluppo. I *tool* disponibili sviluppati per supportare governi e organizzazioni in questo compito, hanno spesso come carenza il focalizzarsi solo su un limitato numero di aspetti o tecnologie, e, nello specifico, spesso trascurano l'importante ruolo che le informazioni spaziali quali la distribuzione della popolazione e la topologia delle reti elettriche ricoprono in una valutazione che ambisce ad essere esaustiva. *GISEle*, il cui processo di sviluppo è descritto in questo lavoro, è un nuova procedura basata su GIS che affronta queste criticità. Le informazioni spaziali sono sfruttate per identificare aree sufficientemente popolate all'interno della regione in esame, e al fine di tracciare l'infrastruttura elettrica che connette tutti i consumatori di ciascun *cluster* all'interno di una singola comunità energetica, perseguendo una minimizzazione dei costi. A questo scopo, un innovativo algoritmo di percorso minimo è stato sviluppato dagli autori. Esso è in grado di valicare i limiti degli approcci esistenti, tipicamente applicati nella ricerca operativa della teoria dei grafi, permettendo di definire l'*albero di costo minimo* anche per sistemi di grafi con un elevato numero di nodi. Come risultato finale, assieme alla topologia della rete, la procedura complessiva fornisce anche un'indicazione riguardo la strategia di approvvigionamento energetico ottimale da seguire, tra *micro* e *mini-grid* indipendenti oppure estensione della rete nazionale esistente. Le performance di *GISEle* sono state testate sulla provincia di Namanjavira, Mozambico che ne hanno evidenziato le potenzialità ed il valore aggiunto che, un approccio spaziale può apportare al processo di definizione di strategie di elettrificazione ottime. Ciò rende questo nuova metodologia uno strumento di supporto valido sia per governi che per il settore privato e organizzazioni internazionalli che operano in questo settore.

List of Figures

List of Tables

List of Acronyms

Acronym	***Description***
AMC	Antecedent Moisture Conditions
ARENE	Autoridade Reguladora de Energia
ARERA	Autoritá di Regolazione per Energia Reti e Ambiente
B1S	Batched 1-Steiner
BIC	Bayesian Information Criterion
CN	Curve Number
CNELEC	Conselho Nacional de Electricidade
CRS	Coordinate Reference System
CSI	Consortium for Spatial Information
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DEM	Digital Elevation Model
DER	Distributed Energy Resources
DG	Distributed Generation
DNI	Direct Normal Irradiation
ECOWAS	Economic Community of West African States
EDM	Energia De Moçambique
EDP	Energias de Portugal
EE	Energy Efficiency
ET	Evapo-Transpiration
GCS	Geographic Coordinate System
GHI	Global Horizontal Irradiance
GIS	Geographic Information System
GRDC	Global Runoff Data Center
HOMER	Hybrid Optimization of Multiple Energy Resources
HV	High Voltage

Acronym	***Description***
I1S	Iterated 1-Steiner approach
KMB	Kou, Markowsky and Berman
LCOE	Levelized Cost Of Energy
LV	Low Voltage
MFD	Multiple Flow Direction
MIREME	Ministério dos Recursos Minerais e Energia
MST	Minimum Spanning Tree
MV	Medium Voltage
NG	National Grid
NGC	Non-Geographic Cost
NGO	Non-Governamental Organization
NRCS	Natural Resources Conservation Service
O&M	Operation and Maintenance
ONSSET	Open Source Spatial Electrification Tool
PCS	Projected Coordinate System
PET	Potential Evapo-Transpiration
PV	Photovoltaic
REM	Reference Electrification Model
RE	Renewable Energy
REP	Renewable Energy Projects
RNM	Reference Network Model
SCS	Soil Conservation Service
SMT	Steiner Minimum Tree
TCC	Terrain Cross Cost
UNEP	United Nations Environment Program
WB	World Bank
WMO	World Meteorological Organization
WRDC	World Radiation Data Center
ZEL	Zelikovsky

Introduction

Energy is one of the fundamental pillars for the development of human society. History teaches that new discoveries in energy field have always brought to periods of wealth and strong development to those nations which could benefit from them; easy access to energy sources has sustained their growth, whereas it's lack has determined their downfall. Still nowadays, a relevant portion of world population does not rely on modern energy sources and the objective to guarantee access to everyone is one of the main challenges that the international community is going to face in the next decades. With this aim, in 2015, 193 Member States of the United Nations defined the seventh of 17 Sustainable Development Goals (SDG) in which they commit to *"ensure access to affordable, reliable and modern energy fuel for all"* by 2030 [1], marking a new level of political recognition of the problem [1].

The pursuit of this objective is driven by ethical and solidarity reasons: the reduction of inequalities, the improvement of medical and educational services, the empowerment of people which often lead the activities of Non-Governmental Organizations (NGO); but also it leads to important economic implications.
Investing capitals in energy and on its reliable and affordable distribution, leads to an increase of population's living standards, the development of industrial and manufacturing activities, and the born of new services. This living environment attracts investments from foreign countries and multinational corporations which find in such development the opportunity to create new markets, involving also National Governments, International Institutions like the International Monetary Fund (IMF) or the World Bank, investments funds and big private companies among the actors of this process.
Despite the strategies adopted in cooperation plans by the international community in the second half of XX century have been classified as failures by the community itself due to the adoption of wrong paradigms, "progress has been made worldwide. The number of people without electricity access fell below 1 billion for the first time in 2017. Meanwhile, updated data shows that the number of people without access

[1] SDGs are the big objectives fixed by the United Nations to achieve a better and more sustainable future for all. They address the global challenges related to poverty, inequality, climate, environmental degradation, prosperity, peace and justice; energy is at the heart of many of these[2].

to clean cooking facilities has been gradually declining" [1]. And after a period of stagnation, now countries are giving more aid [3]. Since 2012, more than 100 million people per year have gained electricity access, an acceleration from the rate of 62 million people per year seen between 2000 and 2012 [2]. However, many more rely on poor quality electricity services.

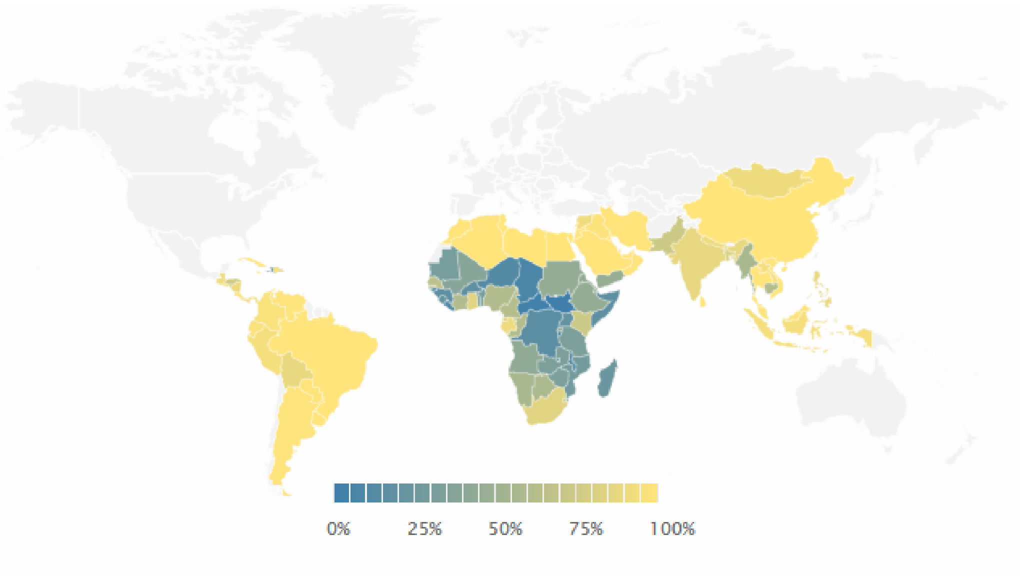


Figure 1: Proportion of population with access to electricity, 2017 [source: IEA][4]

The available strategies to spread the electrical energy in areas where it lacks are different:

- **National grid expansion**: It requires high capital investments, but is able to provide energy at cheaper cost with respect to that produced by isolated systems. This solution can be performed only by national governments.
- **Integrated microgrids**: independent system, aggregating loads and distributed energy resources, that acts as a single controllable entity with respect to the grid. A micro-grid is provided with a grid connection and is able to connect and disconnect from the grid to enable it to operate in both grid-connected or island-mode.
- **Off grid systems**: lacking of grid connection, this solution is usually implemented to electrify remote areas. Stand alone systems can range from home based systems, relying on a single source, to mini grids, integrating more loads and more than one source of energy.
  The cost of energy provided by this alternative could be much higher than that provided by the grid, due to the impossibility to exploit scale economy, but often results the only available solution.

From 2000 to 2016 nearly all of those who gained access to electricity worldwide

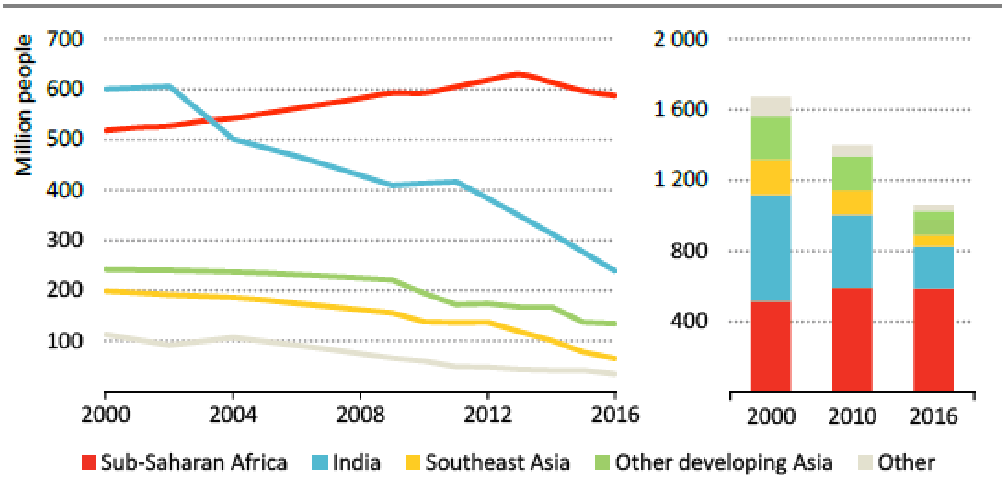


Figure 2: Population without access to electricity by region [source: IEA][2]

Note: Other includes Middle East, North Africa and Latin America.

exploit new grid connections, mostly with power generation from fossil fuels. Over the last five years, however, renewables have started to gain ground, as have off-grid and mini-grid systems, and this shift is expected to accelerate [2].

As shown, different choices require different capital expenditure and involve different players. Therefore making the right choice in the planning of the electrification strategy becomes of priority importance.

A good coordination is essential between donors in order to do not vanish the efforts with nonsense overlap of multiple projects, and technology makes available several tools to support different actors in each phase of their planning work:

1. TERRITORY ANALYSIS:
   This initial stage is devoted to study the characteristics of the area of interest. Identify the presence of the national grid and where it is planned to expand; which places are more suitable to be connected in the future, and which other are enough isolated to make convenient the realization of a microgrid or the diffusion of stand-alone home-based systems. This kind of analysis is crucial both for national governments and NGOs:

   - Governments can wisely plan their investments in grid expansion.
   - NGOs, instead, need to chose suitable areas for off-grid projects without incur the risk to superpose their activity to that of the government.

   Literature already proposed tools designed to handle the problem: Onsset, an open tool developed by the division of Energy Systems Analysis at KTH in Stockholm[5], performs this analysis exploiting georeferenced data: land characteristics, population distribution and electrification level are analyzed to identify the least-cost electrification option(s) between those presented above. Details are depicted in chapter 2.

2. LOAD ESTIMATION:
   Energy habits can be determined by means of detailed survey collected on site or making reasonable estimates about social composition and economical activities. By these information, the energy provider can model a daily profile of energy demand, and then size the energy generation system to fulfil it.

3. GENERATION SIZING:
   This is the last step of the design process. Also in this case literature propose already consolidated tools: Homer or Calliope guide the donor with great detail in the identification of the best combination of energy technologies to be installed. They are able to consider almost all the technical, economical and environmental factors which influence the generation performances: availability of renewable resources along the year, efficiency of machinery and the impact of climate condition on them, pollutant emissions and the influence of fuel cost. A multitude of different combinations are simulated, optimized and finally sorted pursuing the minimization of the energy cost.

The Levelized Cost Of Energy (LCOE) is defined as the average minimum price at which electricity must be sold in order to break-even over the lifetime of the project; or equally, the present value of the price of the produced electrical energy (usually expressed in units of currency per kilowatt-hour or megawatt-day), considering the economic life of the plant and the costs incurred in the construction, operation and maintenance, and the fuel costs [6].
It allows for the direct comparison of the costs of electricity generation projects, even with unequal economic life, capital costs, risks and returns, capacity factor, efficiencies, and fuel costs [7].

$$LCOE = \frac{Sum\ of\ costs\ over\ lifetime}{Sum\ of\ the\ electrical\ energy\ produced\ over\ lifetime} = \frac{\sum_{t=1}^{n} \frac{C_t+OeM_t+F_t}{(1+r)^t}}{\sum_{t=1}^{n} \frac{E_t}{(1+r)^t}} \tag{1}$$

C_t	:	Capital expenditures in the year t
OeM	:	Operations and maintenance expenditures in the year t
F_t	:	Fuel expenditures in the year t
E_t	:	Electrical energy generated in the year t
r	:	Discount rate
n	:	Expected lifetime of system or power station

This is the main and most explanatory indicator of a project economic feasibility. If people are not willing to pay at least a fee equal to LCOE to buy electricity,

the project would not be sustainable on the long run. In view of this, once the constraints about pollutant emissions and renewable-energy fraction have been set, donors usually choose the technical solutions which guarantee the lowest LCOE.
Nevertheless, if determined considering only costs linked to generation devices, the numerator of the LCOE lacks of a fundamental voice of cost: **the electric lines**. The electric grid connects all loads in a unique energy system and are thus essential for the realization of a micro-grid. The resulting LCOE will be, therefore, underestimated: the wider the geographical area under study the higher the LCOE error.
Usually the electric lines cost is not of easy and fast estimation that is the reason why, despite essential, it is often neglected.
The design of a new electric power line, is indeed a time consuming and costly activity that requires massive and detailed spatial information and project engineers experienced in the terrain. This process could be easier if a geospatial approach to the problem is adopted. Onsset, exploits the potential of GIS (*Geographic Information System*) to analyze ground characteristics and thus estimate the expenditure needed to connect an area to the national grid. Its approach, however, is limited to an high level analysis:

- Grid cost estimation is limited to connection between target area and national transmission and distribution line. No internal grid is modelled.
- Only direct, aerial distances are considered. Analysis in terms of Operational Research for the minimization of connection cost are not performed. As a matter of facts, there isn't any design of the local grid, just a rough estimation of the costs.

The position of both loads and generation plants must be defined in advance in order to calculate the length of connections. The scattered the loads, the higher will be the cost to connect each other in a unique grid, and an analogue reasoning can be made in relation to the distance between energy source and users. This second aspect is particularly relevant for those technologies which are strictly conditioned by the presence of primary resources like wind turbines and hydro power plants. Sometimes, especially in case of low energy needs, expenditure for connections can overcome the one related to generation. Automated line design has been an active research field in the last decades, but most of the efforts have been made in the detailed sizing of all the elements of the new power line and only a few references include approaches based on realistic geographic information systems (GIS).
The aim of the authors with this thesis work is to define a procedure and to code a tool able to overcome such shortcoming in the design of electric energy system. The *GISEle* procedure, whose name reflects the purpose to apply GIS to rural electrification, strives to unify all the steps depicted above for the design of electric solution devoted to bring the energy where it lacks, with a particular focus on

the modelling of electric lines across the territory. The automation of the routing process, indeed will reduce both the time consumed and the gap between planning and erection, allowing the study of multiple routing solution in a short time. It will exploit potentialities of single tools already developed for specific purposes, with the added value of a deep-rooting on GIS environment, which allows to exploit the own advantages of geo-referenced data.
GISEle is designed in order to:

- Collect morphological, social and energy data from the target area;
- Identify load centers by means of clustering logic; split them in different groups which will constitute single energy systems and excluding those items which are isolated;
- Model the least cost electric grid interconnecting all the entities constituting the energy system, with an adequate estimation of costs and length;
- Trace the optimal path for the electric connection of the cluster with the national grid pursuing the logic of minimum cost;
- Build the load profile of the community and design the energy generation system necessary to supply it;
- Present the resulting electric grid topology on a visual geo-referenced map.

This tool aims to produce a more accurate estimation of both cost of energy of an isolated system (which now include the component related to electric connection), and the one related to national grid connection, allowing a more grounded comparison between on-grid and off-grid alternatives. This will help government planners and off-grid electricity system entrepreneurs to make better decisions about how to plan and implement electrification efforts. Considerations about the shape of the grid, and the positioning of generators, previously relegated to on-site observations, could now be made in advance.
The discussion about the work accomplished will be structured as follow:

Chapter 1 ad 2 starts with a literature review which constitutes the foundation on which the proposed methodology has been built, divided in two macro areas:

1. Exposition of mathematical instrument and referenced scientific methodologies applied to accomplish the tasks of our work.
2. Deep analysis of proprietary software and open source tools representing the state of art of electrification planning strategy in developing countries. Software that has been so far proposed (most of them are still in a preliminary stage) by universities and research centers worldwide are also presented. For each one strength and weaknesses are presented and critically discussed.

Chapter 3 is devoted to the detailed description of the development process of *GISEle*. It represents the core of the authors' work. Each phase, from the data gathering to the final costs analysis, will be explained in detail, justifying the choices and highlighting the problems. Particular attention will be devoted to load's clustering and the electric grid modeling algorithms which constitute the added value of the project

In Chapter 4 a real life study case, sited in Mozambique, is presented. The administrative place of Namanjavira, in northern part of the country, has been chosen to test the procedure due to the availability of data, gained thanks to the collaboration between Politecnico di Milano and COSV, an Italian NGO active in that region, where one of the authors is employed. Thanks to the real life study case the procedure has been tested and validated, pointing out the evolution with respect to the other tools already proposed i literature.

Eventually Chapter 5 discusses the results of the work, the objective accomplished and the further direction of improvement along which the project is going to develop in the next future.

Chapter 1

State of the Art - Algorithms and Models review

The approach investigated with this thesis work aims to properly take into account every step involved in the assessment of the optimal electrification strategy and performed through spatial data analysis.
Due to the wide variety of activities that performs, therefore, it is necessary to handle different problems ranging from terrain analysis to clustering, operational research and power system sizing, each one with its own background of research activity. In this section, the theory behind these fields that constitutes the foundation of the approach proposed, and the available instruments to perform a suited work are presented.

1.1 GIS

Stated the primary role that geo-referencing plays in GISele, GIS results to be the most suitable environment in which let our project grow.
Technically, a Geographic Information System (GIS) is *"a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data"[8]*. Just like in 90's transparent paper or a glossy-surfaced table lighted from below were used to combine different information in a single view, nowadays GIS technologies integrate a range of geographical information into a single analytical model in which data are geo-referenced to cartographic projection. All the data are thus equipped with additional information, stating the exact position on the Earth's surface to which they refer.

The way in which the information about position are interpreted by the system, and data located on a map, depends on the defined *Coordinate Reference System (CRS)*. The most common way to identify a position on Earth is by means of coordinates: latitude and longitude, expressed in degrees, identifying a position on a spherical surface. This is called Geographic Coordinates System (GCS). A Projected Coordinate System (PCS), instead, locates the data on a flat surface. Analysis in a spherical reference system results quite complex and a flat reference system is usually preferred, especially for visual purpose. Nevertheless, in representing the sphere on a flat surface, distortion occurs, similar to what would happen if you cut open a tennis ball and tried to flatten it out. There are different strategies and algorithm to perform this conversion, each one identified by its unique EPSG code. Some example of the wide variety in projection strategies are sown in figure 1.1. For a map to make sense, it is fundamental that, at the beginning of a work, all the data have the same CRS.

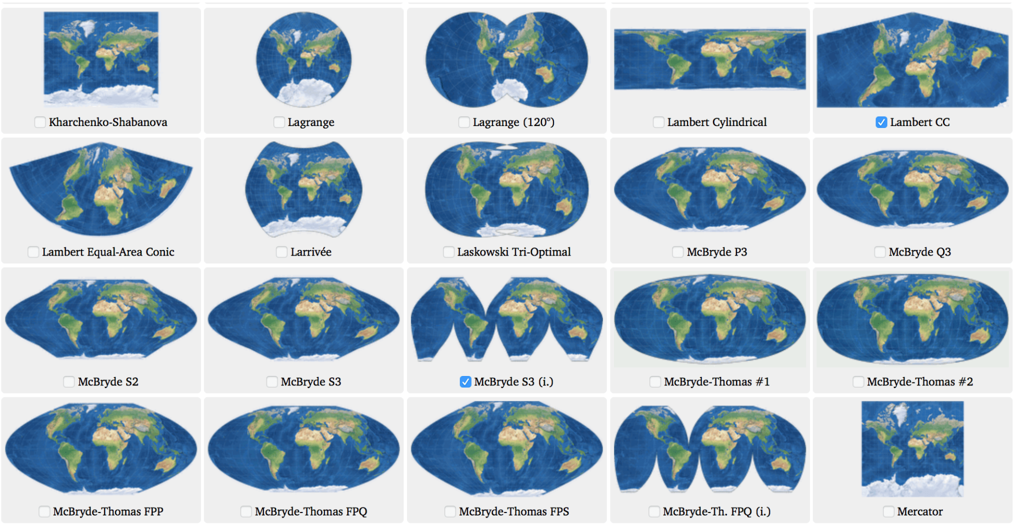


Figure 1.1: Example of different projections strategies of the Earth map on flat surface

GIS works with two different types of data: *raster* or *vector*.

- In raster format the area under study is split in a regular grid of cells, each containing a value that represents the conditions for the area covered by that cell. Raster data are used in a GIS application when we want to display information that is continuous across an area and cannot be easily divided into single features[9].
  This is generally the case of all the data gathered by means of satellite measurements, such as terrain altitude, mean temperature, cloud cover, precipitation or even population density. Fig.1.2 is a raster map representing the amount of global precipitation on Earth. It seems to be continuous because the data resolution is 1 km^2, but zooming a specific section the raster nature of the map becomes clear, as shown in 1.3

- Vector data, instead, provide a way to represent real world features: single, discrete features like rivers, houses, roads, electric lines, national borders can

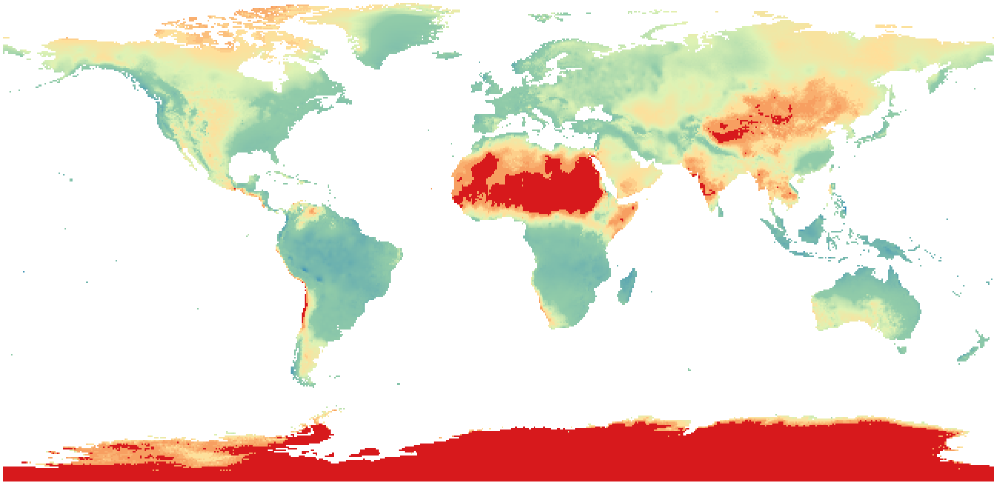


Figure 1.2: Global mean precipitation map on January.
Source: WorldClim

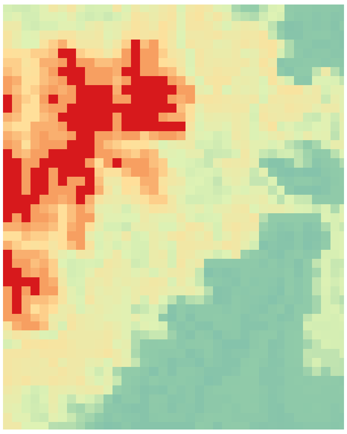


Figure 1.3: Detail of raster map

be represented in a GIS environment as vector data.
A vector feature has its shape represented using one of three possible geometries depending on its nature: *Point, Line, or Polygon.*

- *Points* are used to describe dimensionless features (or those features for which it is not so important to represent their geometric extension). This could be the case of fountains in a town, substations along an electric line, or bridges along a river path. A point describes a position in space using an X, Y and optionally Z coordinates.
- Where the geometry consists of two or more vertices and the first and last vertex are not equal, a *line* feature is formed. All the linear features are represented by lines: rivers, roads, electric lines.
- If the first and last vertexes coincide, instead, a *polygon* is created. Polygons describe enclosed areas, like national borders, city districts, or natu-

ral reserves; or the extension of a specific feature. If the level of detail is enough high, buildings could be defined as polygons instead of points.

The attributes of a feature describe its properties or characteristics. For example a road line may have attributes that describe whether it is surfaced with gravel or tar, how many lanes it has, whether it is a one way street, and so on and so forth. Groups of multiple features of the same type can be treated as single element, becoming *Multipoint*, *Multiline* or *Multipolygon*. The road network is a Multiline feature formed by the set of all the polyline roads that compose it.

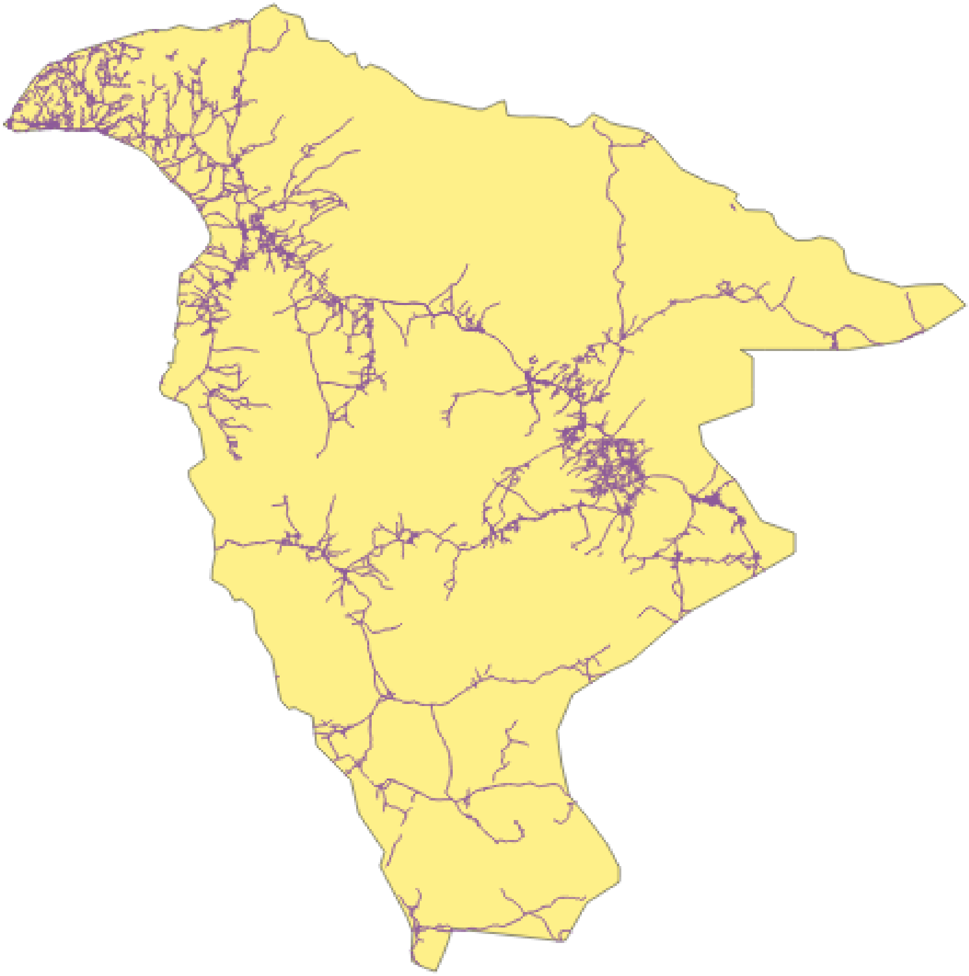


Figure 1.4: Road network for the district of Namanjavira, Mozambique

In the dawning of this technology data were few, they could be gathered only by means of local observation and the accessibility to public users was limited by the need for considerable investments to be made in hardware ad software. Almost until the end of 90's its access was limited to governmental institutions which use geospatial data for territory analysis: mapping natural resources of the country or

the population census. Nevertheless, over the past few decades, thanks also to the big amount of remote data made available by satellite observation, GIS have become an increasingly familiar aspect of urban planning and design practice, finding application in a wide variety of fields. Anthropology, urban planning, agriculture up to energy applications.
Sumic et al.[10] in 1993 identified GIS as the proper computer platform to develop automated routing of underground residential distribution system; and Monteiro et al.[11] developed a GIS spatial methodology for a simple point-to-point overhead economic corridor selection for new power lines. Besides electric lines routing, GIS reveals also a good instrument to identify suitable areas for the realization of energy generation plants, which must submit to strict constraints in terms of environmental impact. Cevallos Sierra et al.[12] in 2018 used GIS data to identify suitable areas for the development of non-conventional renewable energy projects (REP) within the borders of Ecuador's Republic, in order to estimate the maximum energy these technologies could contribute to the national electric energy system.

Satellites turn around the Earth gathering huge amount of data, without concerns about borders restriction; those values which can be measured by remote sensing are thus easily available for pretty much every country of the world with almost the same resolution. The fast evolution in machine learning, thanks to the high investments made in such field in the last year, increases the number of information which can be deduced through the analysis of satellite imagines. As a result, nowadays advanced algorithms are able to automatically identify roads, and distinguish houses. Nevertheless, lot of measurements still need to be collected on-site by means of direct observations which require, in some cases, lot of time and an high capital expenditure, especially if applied to wide areas. Such activities are in the hands of single countries and, inevitably, differences in both quality and availability of data become important between developed and wealthy countries and poorer ones. International agencies and NGOs contributes in reducing this gap by sharing the results of their humanitarian activities in underdeveloped countries. Countries which benefits more are those ones where the activity of such organisms is more intense. Eventually, an important role in mapping process is also reserved to volunteers: internet platforms like Openstreetmap allow everyone to contribute by digitizing features all over the world so that global databases continuously increase in number and details. Spatial data are usually collected within datasets (whose file extension depends on the nature of the data: whether they're raster or vector file) and can be managed, analyzed and presented only within a GIS software. ArcGIS[13], created by Esri, and its open-source version QGIS[14] are the most diffuse ones.

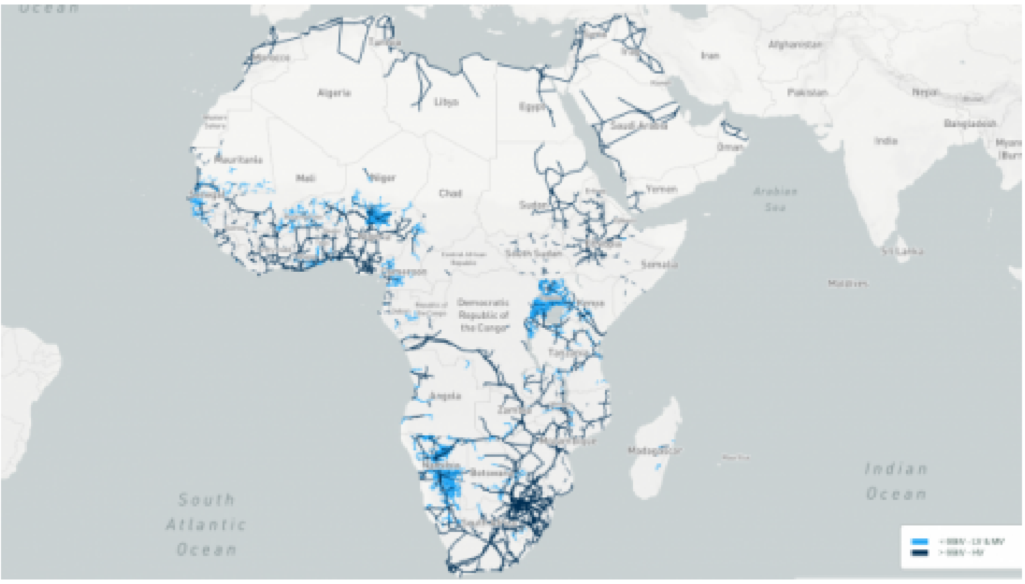


Figure 1.5: African transmission and distribution lines

1.2 Curve Number method

In the assessment of energy resources available within the area interested by a project, the estimation of the hydro-power potential is the most difficult challenge. Whereas global data of Global Horizontal Irradiance (GHI)[1] or wind speed can be easily collected by means of satellite measurements made available for free by a multitude of national research institutes like NASA or ESA directly in geo-referenced format, a database collecting measured values of water flow rates in each point of a river does not exist. Direct measurements of rivers' runoff through satellite analysis cannot be made yet, therefore huge amount of in-site measurements should be made. Sometimes, NGO's or governmental bureaus of environment and water management share results of measurement campaigns related to specific projects, managed along a meaningful time interval : the Global Runoff Data Center (GRDC), an international data centre operating under the auspices of the World Meteorological Organization (WMO), collects many of this data and make them available *"to help earth scientists analyse global climate trends and assess environmental impacts and risks"* [15]. Indeed, location analysis methodology for hydropower development has formerly depended upon onsite surveys and manual work which are costly and time consuming. Moreover, these values are punctual, mainly related to principal rivers and, as shown in figure 1.6 concentrated in developed countries. So, they are not really useful in the estimation of hydropower potential, including also mini and micro power plants, in rural areas of underdeveloped world.

[1] Global Horizontal Irradiance (GHI) is the total solar radiation incident on a horizontal surface. It is the sum of Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance, and ground-reflected radiation.

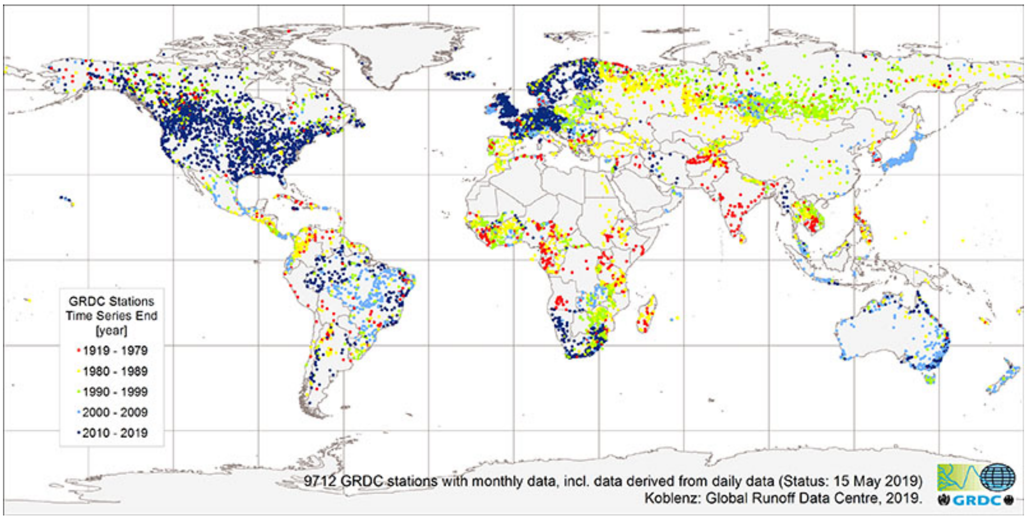

GRDC Stations
Time Series End
[year]
1919 - 1979
1980 - 1989
1990 - 1999
2000 - 2009
2010 - 2019
9712 GRDC stations with monthly data, incl. data derived from daily data (Status: 15 May 2019)
Koblenz: Global Runoff Data Centre, 2019.
GRDC

Figure 1.6: GRDC stations with monthly data (Status: 15 May 2019)

A computer-based model able to accomplish this task exploiting large amount of indirect environmental measurements is needed.

The **Curve Number Method** has been developed by the *USDA-Soil Conservation Service (SCS)* then became Natural Resources Conservation Service (NRCS)[16], for predicting direct runoff or infiltration from rainfall excess in water resources management, storm water modelling and runoff estimation. The runoff Curve Number (CN) is the result of an empirical analysis of runoff from small catchments and hillslope plots monitored by the USDA. Despite notably developed for the prediction of direct runoff after a single rainfall event, due to its simplicity and efficiency it has been also adopted in several studies to determine runoff in ungauged catchments [Indrajeet Sahu, A D Prasad, 2018 [17]; Adornado et al., 2010 [18]; Yevalla et al., 2018 [19]]. Moreover, many watershed models such as AGNPS, EPIC, SWAT and WMS use this method to determine runoff [20], and it is also the procedure adopted in Onsset hydropower estimation.
In hydrology, CN is used to determine value of rainfall which infiltrates into soil or an aquifer and, consequently, how much rainfall becomes surface runoff: a high CN means high runoff and low infiltration (urban areas), instead the lower the curve number, the more permeable the soil is (sand). Combining this parameter with a detailed morphological analysis of the target region it is possible to forecast how surface water moves through the territory and how much water funnels in each point.

CN is a function of land-use/land-cover and hydrological soil group. To each combination of this two features corresponds an empirical value of CN, obtainable by means of tables which can be more or less detailed in relation to the level of detail of available information about land-cover composition as shown by 1.1 and 1.2. All of the available curve number tables are based on the original ones provided by the USDA in its technical release number 55 (TR-55)[21].

Cover Description		**Curve number for hydrologic soil group**			
		A	**B**	**C**	**D**
Open space (lawns, parks, golf courses, cemeteries, etc.)	Poor condition (grass cover < 50%)	68	79	86	89
	Fair condition (grass cover 50 to 75%)	49	69	79	84
	Good condition (grass cover > 75%)	39	61	74	80
Impervious areas	Paved parking lots, roofs, driveways, etc. (excluding right of way)	98	98	98	98
Streets and roads	Paved; curbs and storm sewers (excluding right-of-way)	98	98	98	98
	Paved; open ditches (including right-of-way)	83	89	92	93
	Gravel (including right of way)	76	85	89	91
	Dirt (including right-of-way)	63	77	85	88
Western desert urban area	natural desert landscaping (pervious are only)	63	77	85	83
	Artificial desert landscaping (impervious weed barrier, desert shrub with 1 to 2 inch sand or gravel munch and basin borders)	96	96	96	96
Urban districts	Commercial and business (85% imp.)	89	92	94	95
	Industrial (72% imp.)	81	88	91	93
Residential districts by average lot size	$\frac{1}{8}$ acre or less (town houses) (65% imp.)	77	85	90	92
	$\frac{1}{4}$ acre (38% imp.)	61	75	83	87
	$\frac{1}{3}$ acre (30% imp.)	57	72	81	86
	$\frac{1}{2}$ acre (25% imp.)	54	70	80	85
	1 acre (20% imp.)	51	68	79	84
	2 acre (12% imp.)	46	65	77	82

Table 1.1: Example of highly detailed CN table for fully developed urban areas

CN values spreads from a minimum of 30 to a maximum of 100. An hydrologic group is a group of soils having similar run-off potential under similar storm and cover conditions. Soil properties that influence run-off potential are those that influence the minimum rate of infiltration for a bare soil after prolonged wetting and when not frozen. NRCS[16] has divided soils into the following four hydrologic soil groups:

- A. (Low runoff potential): The soils have a high infiltration rate even when thoroughly wetted.They chiefly consist of deep, well drained to excessively drained sands or gravels. They have a high rate of water transmission.

- B. The soils have a moderate infiltration rate when thoroughly wetted. They chiefly are moderately deep to deep, moderately well drained to well drained soils that have moderately fine to moderately coarse textures. They have a moderate rate of water transmission.

- C. The soils have a slow infiltration rate when thoroughly wetted. They chiefly have a layer that impedes downward movement of water or have moderately fine to fine texture. They have a slow rate of water transmission.

- D. (High runoff potential): They chiefly consist of clay soils that have a high swelling potential,a permanent high water table,a clay pan or clay layer at or

Cover Description	**Curve numbers hydrologic Soil Group**			
	A	**B**	**C**	**D**
Agriculture	72	81	88	91
Double Crop	62	71	88	91
Plantation	45	53	67	72
Commercial	89	92	94	95
Industrial	81	88	91	93
Urban	89	92	94	95
Village	72	82	87	91
Land with scrub	36	60	73	79
Land without scrub	45	66	77	83
Scrub forest	33	47	64	67
Canal	100	100	100	100
River	97	97	97	97
Reservoir	100	100	100	100
Prosophis	61	70	74	78
Quarry	71	87	89	91

Table 1.2: Example of simplified CN table

near the surface, and shallow soils over nearly impervious material. They have a very slow rate of water transmission.

Together with land-cover and hydrologic soil type, the ability of the ground to absorb water or not (expressed through CN) is also function of soil's starting conditions when the rainfall event begins: the so called *Antecedent Moisture Conditions (AMC)*.
If the rain comes after a sunny and dry period, usually the terrain will be able to absorb much more water before runoff, therefore CN should be higher with respect to normal condition. At the opposite, terrain could be already wet at rainfall arrival, so its absorption capacity will be reduced, and CN number higher. Those one obtained by means of TR-55 tables are values of CN referred to normal conditions, and expressed as CN_{II}. AMC, classified as DRY, NORMAL or MOIST, are determined by means of the ratio between two parameters: precipitation (P) and potential evapotranspiration (PET). They represent respectively the amount of rain which falls inside a unit area and the ability of the atmosphere to remove water through Evapo-Transpiration (ET) processes[2] [22]. They're both measured in millimeters of waters and referred to the same time interval.

- DRY CONDITIONS (I) are measured for $\frac{P}{PET} < 0.8$

$$CN = CN_I = \frac{4.2 * CN_{II}}{10 - 0.058 * CN_{II}} \tag{1.1}$$

- NORMAL CONDITIONS when $0.8 \leq \frac{P}{PET} < 0.9$

$$CN = CN_{II} \tag{1.2}$$

- MOIST CONDITIONS (III) when $\frac{P}{PET} \geq 0.9$

$$CN = CN_{III} = \frac{23 * CN_{II}}{10 + 0.13 * CN_{II}} \tag{1.3}$$

Once determined the value of CN, the method continues defining the set of equations which lead to the desired estimation of runoff over the unit area.

$$Q = \begin{cases} 0 & \text{for } P \leq I_a \\ \frac{(P-I_a)^2}{(P-I_a+S)} & \text{for } P > I_a \end{cases} \tag{1.4}$$

Where:

[2] The FAO introduced the definition of PET as the ET of a reference crop under optimal conditions, having the characteristics of well watered grass with an assumed height of 12 centimeters, a fixed surface resistance of 70 seconds per meter and an albedo of 0.23

- Q → Runoff
- P → Rainfall
- S → *Potential maximum soil moisture retention after runoff begins*
- I_a → *Initial Abstraction.*
  It represents the amount of water before runoff which is lost due to infiltration in the underground or because intercepted by vegetation

The value of S is linked to CN through the formula:

$$S = \frac{25400}{CN} - 254 \qquad\qquad S = \frac{1000}{CN} - 10 \tag{1.5}$$

The difference is only in unit of measurement: the left one in millimeters [mm], the right one in inches [in].
Originally, from the study of many small, experimental watersheds, I_a was set equal to the 20% of S: $I_a = 0.2 * S$.
However, more recent analysis performed by Hawkins et al.[23] with hundreds of rainfall-runoff data from numerous U.S. watersheds states that the ratio of I_a to S varies from storm to storm and watershed to watershed and a ratio equal to 0.05 seems to be more appropriate.
Together with the change of the ratio $\frac{I_a}{S}$, also the value of S changes. The equations 1.5 stated above are suitable only to determine $S_{0.2}$, the value of potential maximum retention in which the 20% of the storage is assumed to be initial abstraction. The new parameter $S_{0.05}$ can be determined from $S_{0.2}$ by means of a simple correlation obtained from model fitting results:

$$S_{0.05} = 1.33 * (S_{0.20})^{1.15} \tag{1.6}$$

And the final runoff equation becomes:

$$Q = \begin{cases} 0 & \text{for } P \leq 0.05S \\ \frac{(P-0.05S_{0.05})^2}{(P+0.95S_{0.05})} & \text{for } P > 0.05S \end{cases} \tag{1.7}$$

1.3 Clustering

With cluster analysis, we identify a statistical process with which a population of data is grouped in sub-groups, or clusters, such that each element of each cluster has significantly more affinity with the other members of its sub-group compared with the other clusters' ones. It is well known in statistics and computer science fields as it is one of the main pillars of many fundamental areas of study like data mining, machine and deep learning, image and pattern recognition.

The term was originated in anthropology by Driver and Kroeber in 1932[24], introduced to psychology by Joseph Zubin in 1938[25] and Robert Tryon in 1939[26] and famously used by Cattell in the beginning of 1943 for trait theory classification in personality psychology[27].

Nowadays clustering indicates a relatively vast portfolio of models, relying on different approaches to the data classification problem leading to many structured and algorithms with which each one of us interacts every single day. The reason there are several approaches is related to the fact that each problem has its peculiarity and each set of data needs to be addressed in the most suitable way. For this reason, it is important to note that when applying clustering to a population of data, having substantial knowledge of the characteristics of that dataset is an unavoidable precondition[28]. Some interesting real-life examples of clustering are for instance, finding if an animal species is genetically closely related to a second or third species, selecting which is the best ad to propose to a social network specific user, identifying steps in a trip from spatial-temporal GPS data, or creating algorithms able to recognize objects in images. Cluster analysis models can be mainly grouped as proposed by [29]

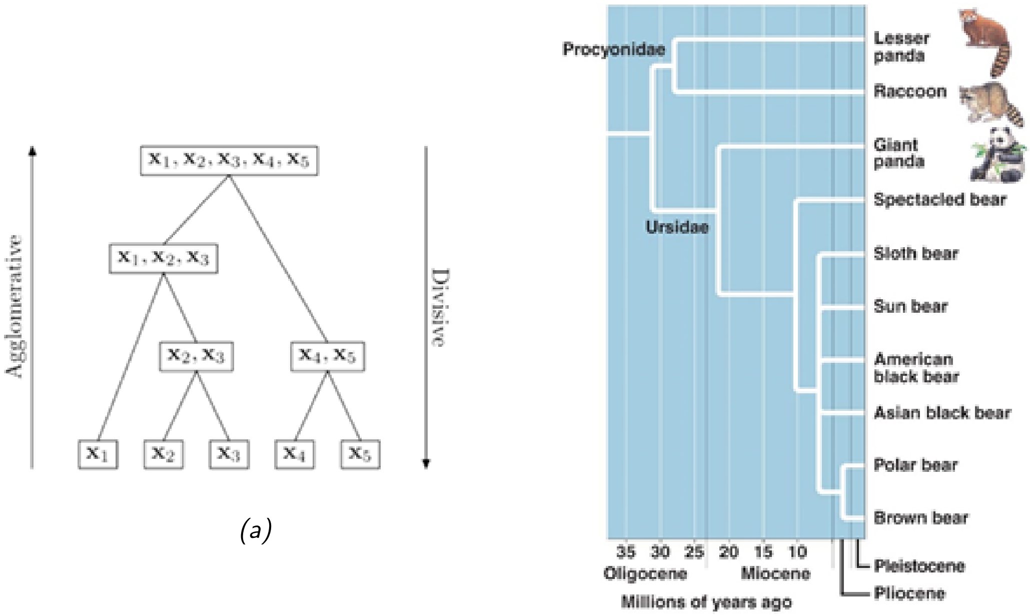
- Hierarchical Clustering Techniques or connectivity models
- Center-based Clustering Algorithms or centroids models
- Density-based Clustering Algorithms
- Graph-based Clustering Algorithms
- Grid-based Clustering Algorithms
- Model-based Clustering Algorithms or distribution models

We will briefly discuss each class by describing one of their representative algorithms.

Hierarchical Clustering

Hierarchical clustering divides datasets into a sequence of nested partitions: its related algorithms are distinguished into agglomerative hierarchical algorithms and divisive hierarchical algorithms (see fig 1.7a. The first category, use a bottom-up approach, starting with the most disperse configuration in which every object is a cluster by itself. Then iteratively it merges pairs of objects following a proximity criterion, ending (hypothetically) with a single cluster comprehensive of all data. The second one is reciprocal, so, through a top-down approach, the entire objects' population starts as one unique cluster, which gradually divides itself into sub-clusters until the desired number of clusters is reached. The last sentence reveals one of the key criticalities of clustering algorithms, which is being parametric, meaning that

some exogenous parameters need to be provided, and in this parameters' definition relies the already mentioned paramount understanding of what the studied dataset is. In this case for both algorithms' models, the operator needs to a priori know in how many clusters the initial population needs to be segmented.
The process of this model is generally represented by a so call n-tree diagram. Between the listed examples, the best suitable one for hierarchical clustering is finding the phylogenetic tree of animal evolution using DNA sequencing data[30]. In this case, the proximity measure is the "edit distance" between each sequence. It's intuitive to see why this clustering technique is the best suitable one considering the underlying structure of the considered data. The result of this technique, used for example to demonstrate the closest relative to the lesser panda, between the giant panda and the raccoon is represented in the figure 1.7b.



(a)

(b)

Figure 1.7: Hierarchical clustering scheme

Disadvantages of this clustering technique involve:

- Clustering order, as data that have been initially allocated in a cluster different from the most appropriate one, cannot be recovered and regrouped;

- Proximity measure strong dependence, as even little changes in the way proximity between objects is computed can lead to huge differences in the final result.

- Computationally inefficient, as each object is continuously compared to all the other ones;

Center-based Clustering Algorithms

This class, which is maybe the most commonly used among the overall categories, performs its tasks by minimizing its own objective function. Each resulting cluster is represented by a center point, commonly called centroid. We will analyze the k-means algorithm, the most illustrious member of this class, highlighting strengths and weaknesses. It gives its name to its fundamental parameter: k is a natural number, simply defining the desired final number of clusters to be found among the studied dataset. To reach a result, the algorithm pursues the goal of finding the k centroids' locations minimizing the total squared distances between objects and centroids. Centroids are not necessarily members of the starting population. A cluster number attribute is assigned to each object, corresponding to its closest centroid.
The basic version of k-means works as follow:

1. The user defines D and k and respectively the data set and the number of required clusters;

2. (C1, C2,..., Ck) initial clusters are randomly defined placing k random centroids in the data space;

3. Assign each data point to the closest centroid;

4. When all data have been assigned, compute each cluster's new centroid;

5. Repeat steps 2 and 3 until no changes in cluster membership occurs;

Between the previously proposed clustering applications, customer segmentation[31] is certainly the one in which center-based clustering is mostly used with consequent great popularity. The basic concept is very simple: similar consumers buy similar products. With this approach, the e-commerce giant, which owns a huge amount of data concerning consumers' habits in terms of expenditure, researches, and favorite products. Applying the k-means algorithm, it is possible to divide the whole population of consumers into k clusters, each one containing consumers with closely related spending habits and tastes. The consequence is obviously a more efficient advertising, and the ability to predict how much a product which has never been bought or even searched by a consumer, would have appeal on her/him, depending on data of similar consumers' tastes.
The main weakness of k-means is related to its parameters: sometimes it is not possible to know how many clusters best represent the starting population because for the vast majority of data this information is implicit in the data itself. Secondly, when applied to data with some sort of spatial dimensions, k-means cannot find not-convex clusters. The geometric representation of k-means output is known as

Voronoi diagram, partitioning the whole space into k subsets in which, each subset's point is closer to its corresponding center than to any other subsets' point. From the other side, the popularity of this algorithm is also due to its extremely high efficiency compared to the others. Because of that, some methods have been developed to overcome its main issue[32]. Nowadays, three methods exist to find an approximation of the best value of k:

- The "Elbow method"
- The "Average silhouette method"
- The "Gap statistic method"

With each of them, it is possible to identify an approximative suitable k value to be used as input for the algorithm. Nonetheless, the forced convexity of the results persists, just as much as the Voronoi division of the space. Most of the time, due to its simplicity and efficiency, this algorithm is used as the first investigation strategy for large datasets in order to obtain some insights before proceeding with more time-consuming algorithms

Density-based Clustering Algorithms

This relatively new class of algorithms is based on the concept of density, which can, depending on the considered problem, assume different meanings. The most popular among them is certainly DBSCAN (Density-Based Spatial Clustering of Applications with Noise), proposed in 1996[33] and developed with the goal of discovering arbitrarily shaped clusters in large spatial databases. Its name also contains additional information, which is the definition of "noise" as the counterpart of its fundamental goal, finding densely populated subsets. DBSCAN defines the concept of density combining a reference neighborhood and reference numerosity. These two parameters, usually referred as *eps* and *minPts*, can respectively be seen as the radius of a circle and the number of objects that need to fall inside that circle to consider it a dense area. This dense area will be represented by its center point, that consequently will be defined as *"core point"*. From all core points the algorithm will step by step evaluate the whole starting population and the points which are not identified as core will fall under the categories of *"border points"* and *"noise points"* (or "*outliers*"): the first one is composed by points which are "reachable" from at least one core point, which means that they fall into one of the eps-radius circles surrounding core points, and the second ones, as already anticipated, are points outside all the previous eps-radius circles.
The basic version of DBSCAN works as follow:

1. The user defines D, *eps* and *minPts*, respectively the data set, radius of the

recursively studied neighborhood and the minimum number of points required to be found in the eps-neighborhood to define it as dense;

2. from a random point, draw the corresponding eps-radius neighborhood, and:
    (a) If there are at least *minPts* points, classify it as *”core point”*;
    (b) If there are less than *minPts* points, check each of the neighborhood points
        i. If there are points meeting the condition in 2a classify them as *”core points”* and classify the starting point as *”border point”*;
        ii. If there are no points meeting the condition in 2a classify the starting point as *”noise”*;
    (c) Continue until all points have been classified;
3. Browse the *”core points”* and group them depending on whether they belong to each other’s neighborhood. Progressively number the resulting groups: the total will be the number of clusters.
4. Browse the *”border points”* and assign them a cluster attribute corresponding to their closest *”core point”* cluster number.
5. End when all points have been assigned to a cluster (*”noise point”* statuses do not change from 2(b)ii)

DBSCAN (density-based clustering in general) has several advantages with respect to other techniques. Firstly, its parameters and the notion itself of density is generally easily valuable by a user which has enough knowledge about the considered data; differently from the final number of clusters which instead generally is exactly what we would like to have as the output of a clustering procedure. Secondly, especially when dealing with data with spatial dimensions, DBSCAN has the capacity to generating arbitrarily shaped clusters. Finally, the possibility of classifying data as outliers totally changes the perspective when it comes to prioritizing as bases of any kind of strategy.
Between the examples above, one is particularly indicated for DBSCAN application and is the identification of stops location[34] from a huge continuous GPS tracking points dataset. In this case, it is obvious that tracking a person, with for instance points continuously generated after a relatively small timestep will generate two types of data: long paths when traveling and dense areas with high number of objects when stopping for example in a city. Applying DBSCAN to such a structure the user will need to define *eps* and *minPts*: varying the first one will lead to results that may represent everything from buildings to countries, and the second one (which could be correlated to a time measure considering the constant timestep) could be used as time-threshold above which we can speak about an effective stop and not a simple transit. On the other hand their main weaknesses involve the fact

of not being totally deterministic, so for example if data are provided in two different orders, even if rarely, if two clusters are really close to each other it could happen that some border points may be falling in one or the other depending on the sorting input. This issue does not affect core and outlier points and has, depending on data types, a generally low impact on the final results. Moreover, while in a plane or in a tri-dimensional space the concept of distance between objects is easily computable, for high-dimensional data its notion could be hard to define. Finally, DBSCAN implicitly assumes that clusters have similar density, so when applied to clusters with high variability in terms of density it could be difficult to have meaningful results as the definition of *eps* and *minPts* is unique and cannot be cluster specific.

Graph-based Clustering Algorithms

Graph-based algorithms are a particular combination between the so called ”graph theory” and clustering techniques. Their functioning can be divided in two steps:

- Graph creation, in which a graph, or hypergraph, is created between all points composing datasets. Each point will be a ”node” and each line connecting two nodes will be an ”edge”: this part will not be deepened cause object of the next chapter.
- partitioning, where the real clustering process occurs. In it, a *relative closeness* evaluation is performed, and edges are gradually removed until final clusters are defined. This relative closeness can be defined through different approaches like a basic length threshold or other more sophisticated absolute or relative weighting criteria.

Grid-based Clustering Algorithms

This category is highly similar to the afore-described algorithms, with a major difference, from which it takes its name, is the fact that before starting its clustering process, a grid-based grouping is generally performed. It has generally these steps:

1. Partitioning the data space in the grid-like structure (cells);
2. Calculating cell-specific densities and sorting them accordingly;
3. Clustering cells finding each correspondent center;

This means that data are not treated as points anymore, but the algorithm only refers to the cells' density values. The result is a category of algorithms, among which the most popular ones are STING, OptiGrid, GRIDCLUS, GDILC, and WaveCluster, where the partitioning step is designed to better fit with the following pure clustering step, in which methodologies like the hierarchical and graph-based clustering techniques are applied. The great advantage, in this case, emerges when dealing with very large dataset: being the computational burden deriving for the dataset's dimension is accounted only in the initial partitioning step (which precisely because is a partitioning process is very fast), the clustering part burden only depends on the grid resolution adopted by the user. Generally, these algorithms are very fast processing big data with acceptable resolutions, but when increasing too much the latter could lead again to high computational burdens.

Model-based Clustering Algorithms

This category directly derives from statistics and relies on the assumption that the whole dataset follows some kind of probability distribution model, so a mathematical function capable of describing the likelihood of occurrence of the possible results of a random phenomenon. The common approach consists in using different models and trying to adjust the fit between models and data. In the latter, data are regarded as the output of mixture of probability distributions, each representing a different cluster, so the suitability of a cluster algorithm depends on the conformity of data with respect of the model.
In practice, each data's element k is modeled through a distribution model, like for instance the Gaussian one, defined by characteristic parameters like: a mean vector (μ_k), a covariance matrix (Σ_k)) and an element-specific probability of belonging to each different cluster. These parameters are generally estimated using the *Expectation-Maximization* algorithm by hierarchical model-based clustering. Finally, the most suitable model is selected through the *Bayesian Information Criterion (BIC)*: large BIC scores are associated with strong evidence of suitability of a model to the studied data. Model-based clustering techniques have the advantage of not being heuristic, so cannot be used for formal inference. As said, they assume that data are generated by a finite mixture of underlying probability distributions. The positive aspect is given by the fact that, differently from heuristic methods, the problems of determining the most suitable clustering method and number of clusters, with these algorithms the only task is the model selection among the probability framework.
Image recognition[35] is one of the machine learning fields in which this kind of algorithm is used. Training an Artificial Neural Network with a human-filtered number of datasets, as labeled photographs of different items: the photos' pixels will be the objects population, each one having specific attributes related for example to colors. Clustering their pixels based for example on the most abundant colors they contain allow to recognize the underlying patterns for each item category. Secondly, when an unlabeled image is submitted to the model, its pixels will also be clustered by the model, and the underlying trends will be compared to the stored one leading to a certain percentage of similarity and an output which would be the probability for the submitted photograph representing a horse or a house.

Computational complexity

The main problem with data mining algorithms in general, and particularly with clustering algorithms is their scalability. Highly complex problems generally imply large-sized data frames, and unfortunately, each algorithm can easily handle specific order of magnitude of data.

Category	Algorithm	Computational complexity	Required memory	Suitable for large dataset
Hierarchical	*CLINK*	$O(n^2)$	$O(n^2)$	No
Center-based	*K-means*	$O(l \cdot k \cdot m \cdot n)$	$O((n+k) \cdot m)$	Yes
Density-based	*DBSCAN*	$O(n \log n)$	$O(n)$	Yes
Graph-based	*ROCK*	$O(n^2 + n \cdot m_m \cdot m_a$ $+n^2 \log n)$	$O(min\{n^2,$ $n \cdot m_m \cdot m_a\})$	No
Grid-based	*STING*	$O(n)$	$O(n)$	Yes
Model-based	Gaussian Hierarchical	$O(k \cdot n)$	$O(n)$	Yes

	n	:	Number of data points
	m	:	Number of attributes
Where:	l	:	Number of iterations for convergence
	m_m	:	Maximum number of neighbors for a point
	m_a	:	Average number of neighbors

1.4 Graph Theory

The ability to track the topology of the electric network, making also an estimate on its capital cost, aims to be one of the key strength of the approach proposed with this work.

”Which is the cheapest way to connect together energy users and providers in a unique energy system? How would it look like?”

This is an optimization problem that belongs to the mathematical branch of Operational Research: a discipline that deals with the application of advanced analytical methods to help make better decisions in almost every field. Employing techniques from other mathematical sciences, operations research arrives at optimal or near-optimal solutions to complex decision-making problems.

More in detail, such problem is part of *Shortest Path Problems* defined as the search in a connected and weighted graph of the path which connects a set of defined nodes such that the sum of the weights of its constituent edges is minimized. The associated literature is rich since it is, for example, at the base of satellite navigator's logic; also telco companies exploits those algorithms to define the optimal combination of nodes in the internet web to send data packages from a server to another one. What follows is a presentation of the main algorithms and strategies adopted to the problem solution.

Def 1.4.1 *A Graph $G = (V, E)$ consists in*

- *A set V of "vertexes" ("nodes" or "terminals")*
- *A set E of edges: each one connects 2 nodes and represents the relationship between them.*

A graph can be Oriented or Not Oriented

In our case, nodes are representative of loads and generators, whereas, given that the cost of the electric connection between two nodes is usually proportional to their distance, the value of each edge belonging to the graph can be set as the metric distance between the connected nodes.

Def 1.4.2 ***Tree****: If T is a connected graph without any cycle, then T is called a tree"[36].*
Therefore, a tree is a graph structure in which:
$\forall$ *couple of vertices $u \neq v \in T$, there is exactly one single path from u to v.*

Def 1.4.3 ***Forest****: A forest is a graph all of whose connected components are trees. In particular, a forest with one component is a tree.*

Minimum Spanning Tree Problem

Def 1.4.4 *A* ***Minimum Spanning Tree (MST)*** *is a subset of the edges of a connected, edge-weighted undirected graph $G = (V, E)$ that connects all the vertices together, without any cycles and with the minimum possible total edge weight.*

The *cost* of the tree can be defined as the sum of its edges weight.

$$w(T) = \sum_{e \in T} w(e) \tag{1.8}$$

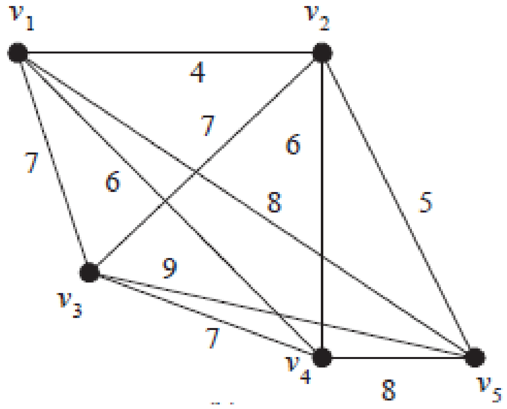


Figure 1.8: Example of weighted, not oriented graph

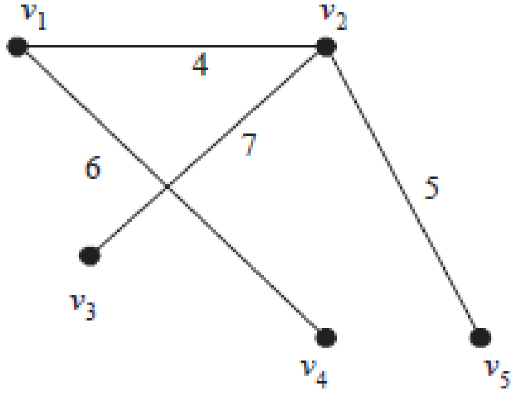


Figure 1.9: Minimum Spanning Tree connecting all the nodes of the graph shown in 1.8

Several algorithms have been developed to find the minimum spanning tree of a graph: main ones are *Kruskal algorithm* and *Prim algorithm* that are now shown. Both of them are named after their creators and exploit a *Greedy strategy* to find the solution.

Greedy strategy This logic, typically applied to optimization problem which requires consequent choices, suggests to make the locally optimal choice at each stage, without concerning about the future. Its limitation consists into the impossibility to change the choice made: if a candidate has been selected, it remains in the solution forever, and if it has been discarded, he will never be considered again. Problems that can be solved by greedy algorithms have two main properties[37]:

- **Optimal Substructure**: the optimal solution to a problem incorporates the optimal solution to subproblem(s).
- **Greedy Choiche Property**: locally optimal choices lead to globally optimal solution

The MST problem belongs to this category and the Greedy strategy is applied to choose which edges include in the solution and which discard.

Kruskal's Algorithm [38] Consider an undirected, weighted graph $G = (V, E)$ with n nodes and m edges, initially composed by n trees each one represented by a single node.
At each step Kruskal's algorithm finds an edge with the least possible weight that connects two trees in the the forest; the process continues until all the nodes are connected in a single tree.
The algorithm is the following:

1. Create a forest F, where each node in the graph is a separate tree;
2. Create a set S containing all the edges e in the graph sorted in increasing order of weight;
3. While S is not empty, consider the first, least costly, edge *e=(u,v)* of S:
    - If u and v belong to different trees → include the selected edge in MST and combine the two trees in a single one;
    - If u and v belong to the same tree → discard that edge;

When the algorithm ends, if the original graph was connected, the result is a minimum spanning tree; if not, the result is a minimum spanning forest.
Fig:1.10 shows an example of Kruskal's algorithm application.

The final MST produced by Kruskal's algorithm is corrected and the required final computational time is, in the worst case, O($m \cdot \log n$).

Prim's Algorithm [38]
The algorithm proposed by computer scientists Robert C. Prim in 1957 (but developed in 1930 by Czech mathematician *Vojtěch Jarník*) arrives at the same result following a different strategy. The starting conditions are the same: an undirected, weighted graph $G = (V, E)$ with n nodes and m edges, initially composed by n trees each one represented by a single node. Prim starts building the MST from an arbitrary vertex and, step by step, continues adding the cheapest possible connection which links the starting tree with another one. The logic can be defined as follows:

1. Create a forest F, where each node in the graph is a separate tree;
2. Select the arbitrary starting node/tree T;
3. While not all the nodes are part of T:

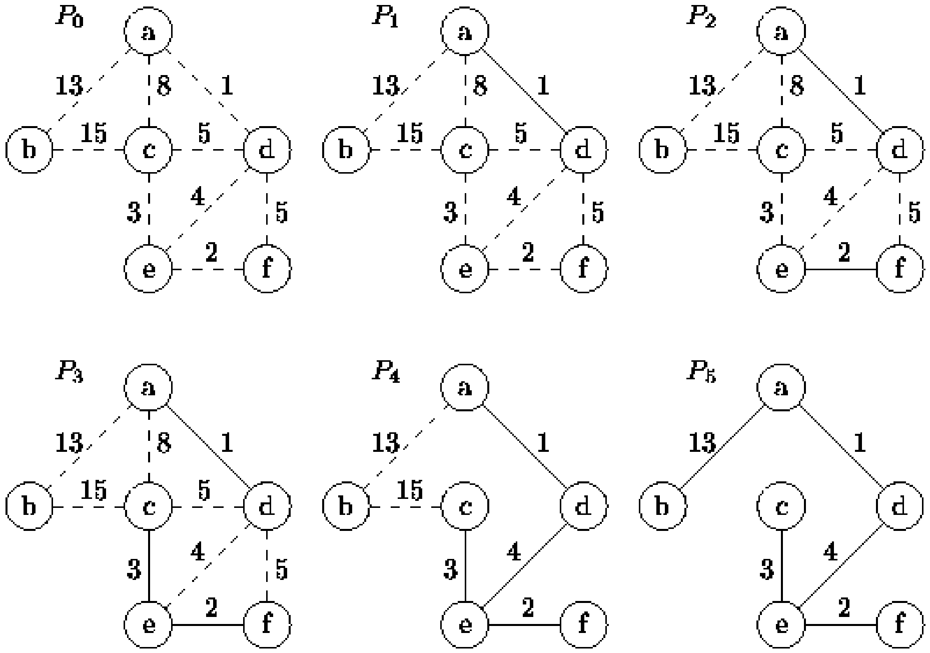


Figure 1.10: Kruskal's algorithm execution

- Identify all the edges $e = (u, v)$ such that u belongs to T and v not.
- Select the least costly one;
- Add it to MST and include v in T;

As stated before the Prim's algorithm leads to a correct solution of the problem, equal to that proposed by Kruskal. Fig1.11 demonstrates it highlighting the procedure in each step. Adopting appropriate data structure, it reaches the objective in time O$(m + n \cdot \log n)$ in the worst case.
Also Otakar Borŭvka in 1926 published its own method to solve MST problem. Nevertheless, it is not as strong as the other two since an accurate solution can be reached only in graphs for which all edge weights are distinct. For this reason we avoid to describe it in details.

Shortest Path

Returning to the original problem of connecting loads and DG, we can notice that the minimum spanning tree allows to reach the objective, but the obtained solution is not enough accurate yet. Indeed, applying the MST approach, the resulting electric grid shows the following characteristics:

- It is composed only by straight lines, each one connecting a couple of terminals;
- No intermediate substations which could lead to better and cheaper solutions are considered.

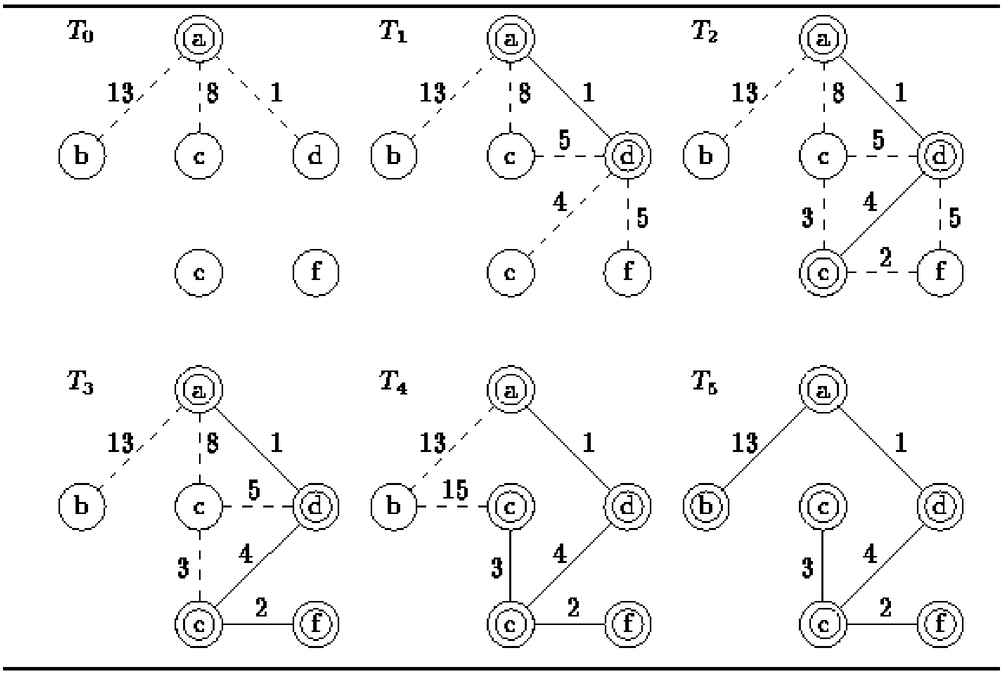


Figure 1.11: Prim's algorithm execution

This because MST approach builds the minimum weight tree that spans through **all** the terminals, therefore it allows to include in the graph only ”target nodes” (loads and generators), nothing else, because otherwise they would be included in the solution and connected. Figure 1.12 shows how a MST solution could be improved if intermediate, non target points are included in the analysis (Blue points represent target points, the red point represents an optional point).
Therefore, MST produces grids in which the connection between a couple of nodes is straight, based on aerial-distance, ignoring any obstacle that might be in the middle. If the objective is the modelling of electric transmission lines in ideal ground conditions, without big orographic obstacles like mountains or lakes, this could be a good approximation: indeed, usually, HV cables runs far from the ground and are not strictly affected by its characteristics. Nevertheless, just as cars cannot follow the same airplane routes, a good design of MV or LV lines cannot include such simplifications. *”The spatial nature of some of the aspect involved in power-line routing leads to a compromise between a straight line from one point to another and path deviation to avoid costly terrains, obstacles, or other intolerance criteria”[11].*
Laying an electric line across a virgin jungle or a swampland, unless it is the shortest path, may result much more difficult and expensive than follows the road around it. In order to find an optimal path, connecting a target point to another one, which takes care of what stands in the middle, more sophisticated algorithms are needed. In a dense graph, whose points and connections represent a discretization of the whole target region with embedded information about terrain characteristics, those

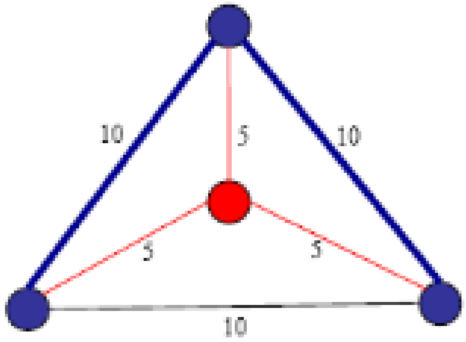


Figure 1.12: Example of improvements to MST solution

algorithms must be able to distinguish a limited group of *target points* (loads and DG) and find the best way to connect them. Below, they are presented.

Dijkstra Algorithm Given a graph G, for a given source node in the graph the algorithm developed by Edsger W. Dijkstra is able to find the shortest path between that node and every other. Therefore, if also a target node is defined, it will find the shortest path between source and target by stopping the algorithm once the shortest path to the destination node has been determined. The following definitions are fundamental premises to understand the algorithm.

Def 1.4.5 ***Sortest path:*** *Consider $G = (V, E)$ a connected and weighted graph with a cost function w: $E \rightarrow \mathbb{R}$. A minimum cost path between a pair of connected vertices u and v is a path π^*_{uv} that has a cost less than or equal to that of any other path π_{uv} between the same vertexes, so that:*

$$w(\pi^*_{uv}) = min_{\pi_{uv} \subseteq G}\ w(\pi_{uv}) \tag{1.9}$$

Lemma 1.4.1 ***Optimal substructure****: Consider $G = (V, E)$ a connected and weighted graph with a cost function w: $E \rightarrow \mathbb{R}$. Therefore, every sub-path of a minimum path in G is itself a minimum path in G.*

Def 1.4.6 ***Distance****: Consider $G = (V, E)$ a connected and weighted graph with a cost function w: $E \rightarrow \mathbb{R}$. The* **distance** *d_{uv} between two nodes in G, is defined as the cost of minimum path which connect them, or $+\infty$ if no path exists between them:*

$$d_{uv} = \begin{cases} w(\pi^*_{uv}), & \textit{if a path exists between u and v in G} \\ +\infty, & \textit{if not} \end{cases} \tag{1.10}$$

Lemma 1.4.2 ***Bellman's condition****: Consider $G = (V, E)$ a connected and weighted graph with a cost function w: $E \rightarrow \mathbb{R}$. Therefore, for each edge (u,v)*

$\in E$ and for every vertex $s \in V$, the distances between the vertexes satisfy the following inequality:

$$d_{su} + w(u, v) \geq d_{sv} \tag{1.11}$$

All the *Minimum Path* algorithms begin by making a very high estimate of the minimum distance between each pair of points in the graph $D_{uv} \geq d_{uv}$. Each time that a shorter path between u and v is found, the distance value is updated until the estimate becomes accurate, that is $D_{uv} = d_{uv}$.
This upgrade consists in considering a third vertex z, a path π_{zv} and applying the following *relaxing step*:

$$\textbf{If } (D_{uz} + w(\pi_{zv}) < D_{uv}) \textbf{ then } D_{uv} = D_{uz} + w(\pi_{zv}) \tag{1.12}$$

Given a graph $G = (V, E)$ with n nodes and m edges, in order to build the Minimum path tree T which includes all the vertexes reachable from a source node s the Dijkstra Algorithm acts as follow:

1. Define the Minimum Paths tree T, initially composed only by the source s.

2. Initialize the "distance array" in which the values of distances between the source and every other node of the graph are collected. At the beginning all the distances D_{sv} are set to $+\infty$, with the exception of distance D_{ss} that is obviously 0.

3. While not all the vertexes belongs to T:
Find the edge (u,v) where $u \in V(T)$ and $v \notin V(T)$ which minimizes $D_{su} + w(u, v)$; update the distance D_{sv} through the relaxing step $D_{sv} = D_{su} + w(u, v)$ and add the arc (u, v) to T

In (n-1) steps the tree T will be completed and all the nodes of the graph will be part of it. Figure 1.13 shows an example of the algorithm execution. It can be noticed how it results similar to Prim's algorithm: both start from a source s and follow a greedy strategy to choose the next edge to include in the graph. The time required to calculate the distances from the source s and all the other vertexes of the graph is $O(m + n \log n)$ as demonstrated by Demetrescu et al.[38].
In the same years Bellman, Ford and Moore[38] conceived another algorithm which accomplishes the same task. However, it performs an high number of useless "relaxing processes" which increase the computational time to $O(mn)$.
Until now, we are able to identify the optimal route across the graph to go from a source point to a target one: Dijkstra's algorithm calculates the minimum distance and, through the resulting *Minimum Path Tree* it is possible to extrapolate the sequence of steps to follow to build the path; but what if the points of the to connect are three, thirty or even some hundreds? This become a really complex problem, so complicated that the best solution can only be approximated.

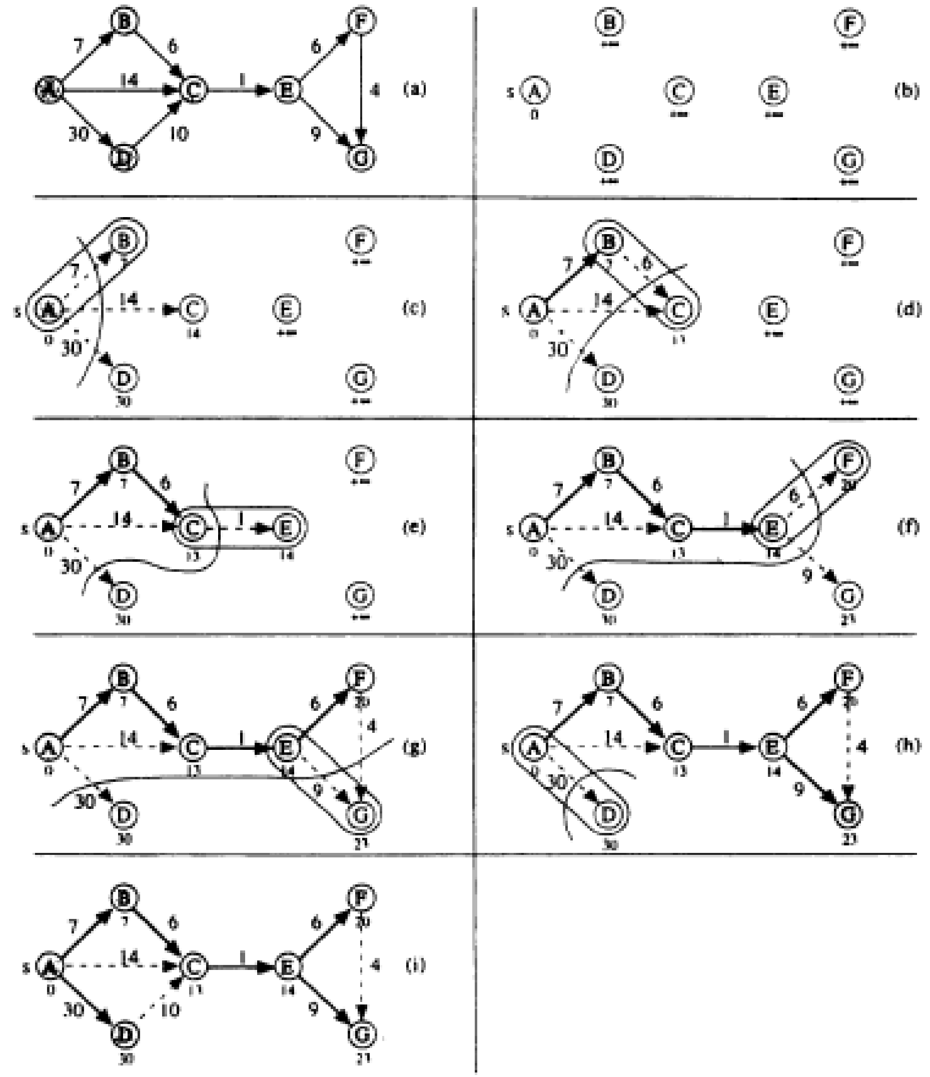


Figure 1.13: Execution of Dijkstra's algorithm with source s = A. Edges belonging to T are in bold, whereas the arc (u,v) with $u \in V(T))$ and $v \notin V(T)$ which minimizes $d_{su} + w(u,v)$ are circled

Steiner Tree Problem

Whereas a spanning tree spans all vertices of a given graph, a Steiner tree spans a given subset of vertices. The objective is the same pursued by Dijkstra, "to find the least cost path", but involving a greater number of target vertexes: instead of a single couple composed by *source and target*, now the targets are multiple and the lest cost way to connect all of them needs to be found. Figure 1.14 shows an example of a Steiner problem solution.

In the Steiner minimal tree problem, the vertices are divided into two groups: *terminals* and *non-terminals* nodes. The *terminals* are the given vertices which must be included in the solution, and so be part of the final grid, like loads and generators. As *non-terminals* nodes, are instead classified all the other nodes which compose the graph, representing the territory with its own characteristics. They're optional nodes, which become part of the solution only if the minimum tree which connects all the terminals pass from them. The cost of a *Steiner Tree* is defined as the overall weight of the edges which compose it.

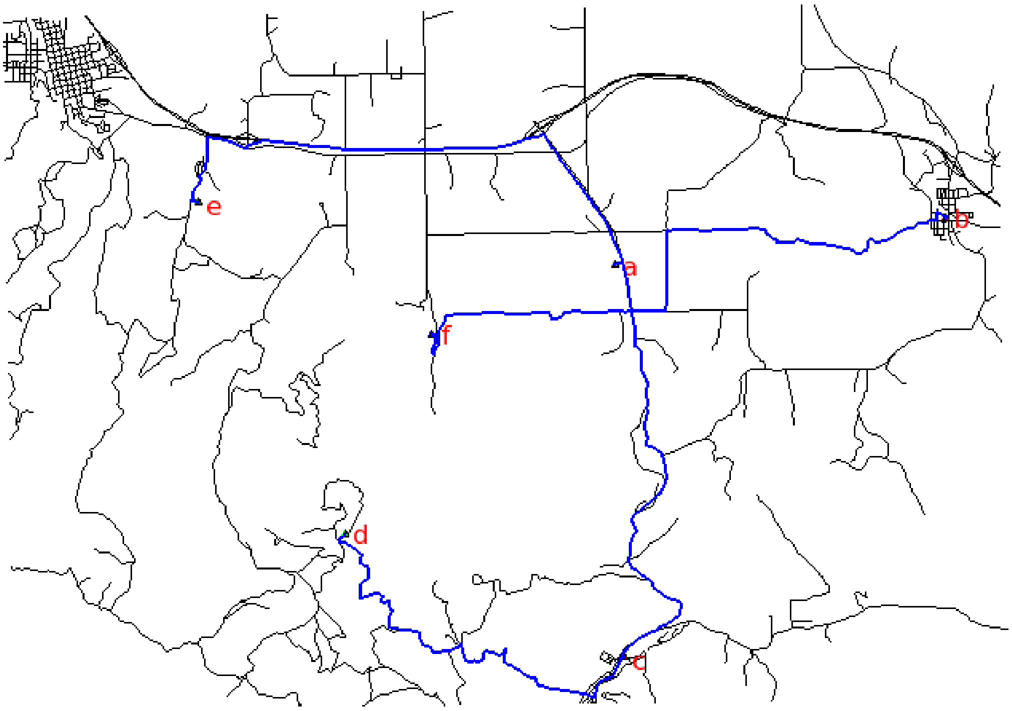


Figure 1.14: The minimum path which allow to connect together all the locations from a *to* f *following the road network is an example of Steiner problem solution*

Minimum Spanning Tree and *Minimum Path* problems illustrated before can be solved by means of polynomial algorithms like Kruskal, Prim and Dijkstra. They are called *polynomial* because require a polynomial amount of time to reach the solution and so are they considered to be manageable. *Steiner Tree Problem*, instead, belongs to a class of problems for which decisive polynomial algorithms have not been developed yet, and furthermore scientists assume that they do not exist at all. These problem family is classified as NP-complete. Find their exact solution would require a not manageable amount of time (months, years or even century), so that scientist prefer to solve them in an approximated way, provided that the execution time remains polynomial and the solution does not deviate too much from the exact one[38].

PROBLEM	:	Graph Steiner Minimal Tree
INSTANCE	:	A graph $G = (V, E, w)$ and a set $L \subset V$ of terminals
GOAL	:	Find a tree T with $L \subset V(T)$ so as to minimize $w(T)$

1^{st} **STRATEGY: Approximation by MST** Given an undirected, weighted graph $G = (V, E, w)$, made of V vertices, E edges with non-negative weight, L terminals which must be reached and V-L Steiner vertices which could be included in the solution in order to find the minimal cost tree:

1. Firstly we compute the shortest paths length between each couple of terminals L: the so called *metric closure* C_L.

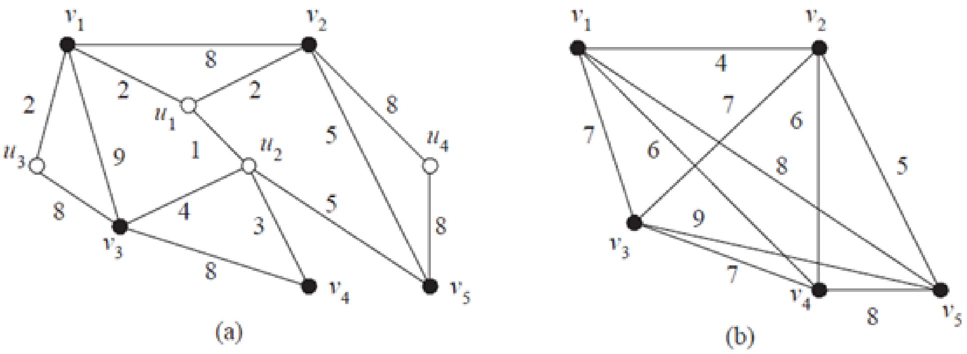


Figure 1.15: Step 1: Construction of the metric closure on terminals L

2. Find a MST T_L on the closure C_L, in which each edge corresponds to one shortest path on the original graph.

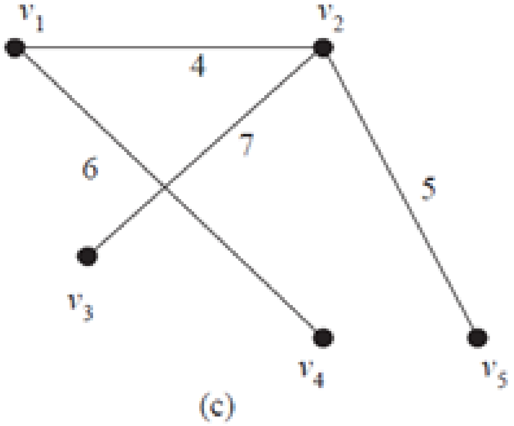


Figure 1.16: Step 2: MST on the metric closure

3. Finally the MST is transformed back to a Steiner tree by replacing each edge

with the corresponding shortest path.
At start, the shortest path between a first arbitrary couple of terminals is added to a new tree T, set initially empty.

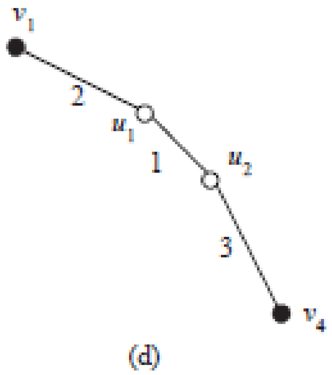


Figure 1.17: Step 3: First path added to the final tree

4. For any shortest path P between any other couple of terminals $(u, v) \in E(T_L)$:

- If less than 2 nodes of $P_{(u,v)}$ already belong to T: the whole path P is added to the tree T.
- If 2 or more nodes of $P_{(u,v)}$ already belong to T: only the sub-paths from the terminals to the vertices already in the tree are inserted.

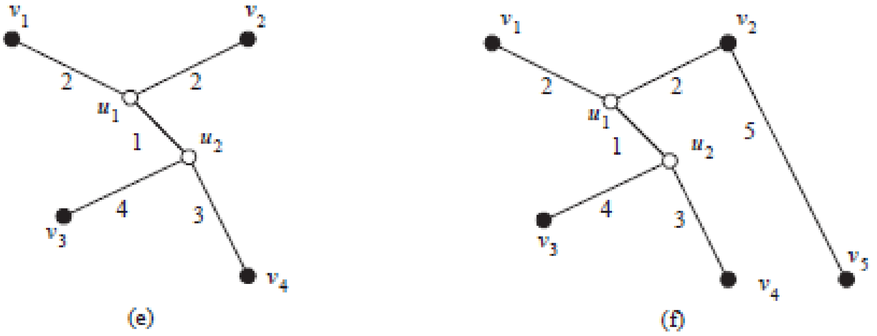


Figure 1.18: Step 4: addition of all the other paths

Note: 2 nodes of the path $[v_2, u_1, u_2, v_3]$ *already belong to T. Thus, only the subpaths* $[v_2, u_1]$ *and* $[u_2, v_3]$ *will be added to the tree T*

This procedure avoids any cycle, repetitions of segments already included in the final tree, and ensures that the terminals are included.

In a graph $G = (V, E, w)$ with L terminals, the time complexity of this algorithm is $O(V * L^2)$, dominated by the construction of the metric closure.

Def 1.4.7 *The* ***Steiner ratio*** *ρ is the ratio between the weight of the MST related to terminals L, and the weight of the corresponding Steiner Tree.*

$$\rho = \frac{w(MST(G_L))}{w(SMT(G_L))} \tag{1.13}$$

So, the higher, the better.

The MST approach to the Steiner Tree problem produce a result with a Steiner ratio of 2.[39]

2^{nd} **STRATEGY: The Iterated 1-Steiner (I1S) Approach** This heuristic method searches for an optimal solution by repeatedly including in the original terminals pointset, single additional Steiner points which produce a reduction in the overall weight of the final tree.
Given two pointsets A and B, we define the MST savings of B with respect to A as:

Def 1.4.8 ***Minimum Spanning Tree saving***

$$\Delta MST(A, B) = \mathit{cost}(MST(A)) - \mathit{cost}(MST(A \cup B)) \tag{1.14}$$

Setting as P the initial set of points composed by the terminal nodes, the Steiner candidate set H(P), as stated by the Hanan's theorem[3], is defined as the collection of the intersection points of all horizontal and vertical lines passing through points of P. For any pointset P, a 1-Steiner point with respect to P is a point $x \in H(P)$ that maximizes $\Delta MST(P, x)$.
Therefore, the process followed by I1S approach is the following:

1. Initialize the pointset P of terminal nodes
2. Define S as the set containing the chosen *1-Steiner* points, initially empty
3. Define $H(P \cup S)$ as the set of *Steiner candidates* determined by means of Hanan's theorem.
4. Find $x \in H(P)$ that maximizes $\Delta MST(P, x)$
5. Update S: $S \leftarrow S \cup x$

Steps from 3 to 5 repeats iteratively until there no longer exists any point x which brings a saving in tree's cost. The cost of $MST(P \cup S)$ will decrease with each added point.
Each MST computation can be performed in $O(n \log n)$ time, yielding an $O(n^3 \log n)$ time method to find a single 1-Steiner point. A linear number of Steiner points can

[3] In 1966 Hanan showed that for a pointset P there exists an SMT whose Steiner points S are all chosen from the Hanan grid, namely the intersections of all the horizontal and vertical lines passing through every point of P

therefore be found in $O(n^3)$ time, and trees with a bounded number of k Steiner points require $O(kn^2)$ time. This is a huge amount of time! Indeed, the I1S heuristic is provably optimal for 4 or less points.Nevertheless it produces more accurate results: *Steiner ratio* varies in relation to the number of terminals which compose the problem and the problem itself ($\frac{7}{6}$ and $\frac{13}{11}$ respectively for 5 and 9 terminal points), but it is at least 1.3 in general.
Results are good: for n=30 points, I1S is on average only about 0.3% away from optimal, therefore a variant of this method able to offer runtime improvements has been developed.

The Batched 1-Steiner Variant This upgrade of I1S approach amortizes the computational expense of finding 1-Steiner points by adding as many "independent" 1-Steiner points as possible in every round.

Def 1.4.9 *Two candidate Steiner points x and y are independent if:*

$$\Delta MST(P,x) + \Delta MST(P,y) \leq \Delta MST(P,x,y) \tag{1.15}$$

Each round of B1S greedily adds into S a maximal set of independent 1-Steiner points, and the total time required for each round is $O(n^2 \log n$. It is important to underline that both I1S and B1S, which select 1-Steiner candidate following the Hanan's theorem, have been conceived for metric geometries, but they can be generalized to arbitrary weighted graphs by combining the geometric I1S heuristic with other graph Steiner algorithms like KMB, ZEL, IKMB or IZEL as shown by Robins and Zelikovsky in [40].

Now that we have extensively discussed algorithms and strategies to move across a weighted graph in search of the optimal grid topology, it becomes necessary to talk about how to set the weight of each graph's edge.

1.5 Weighting strategies

Electric line power routing is an engineering task that optimizes the equipment installation and maintenance cost which are subjected to geographic, environmental, social and legal constraints. In the planning of the path to follow and areas that the line will cross, the existing constraints must be taken into account.
The weight of one edge of the graph which models the target geographic area should then reflect, besides the cost of the materials, the difficulties introduced by the multiple aspects of ground characteristics in building an electric line across that specific stretch of territory; these include slope, soil types, terrain costs, geographic restrictions, obstacle. Monteiro et al. [11] propose a distinction between *Nongeographic cost component (NGC)* and *Terrain cross cost (TCC).*

- NGC interests all the terms related to the equipment of the line which are independent on the geographic feature. These costs are directly associated with the line that connects two adjacent points and are uniformly distributed along it. Their absolute cost is only function of the line length so that a cost per kilometer is obtained.

- TCC, instead, are associated to the single node, which identifies the area surrounding it with its own characteristics. These are those costs referred to typical ground and environmental features which characterize a specific geographic place. For example, a point next to the coast line will embed an additional cost linked to the special insulation necessary to avoid corrosion.

TCC can be classified in several classes:

- **Accessibility:** represents the additional cost for equipment transportation, installation and maintenance. They are preponderant especially in rural and remote areas of developing countries where infrastructures are poor. These costs can be measured in terms of *distance from roads, from main cities or from docks.*

- **Specific geographic characteristics**: additional costs based on soil type, land use, vegetation coverage, urban areas, corrosive environment near the shores. A dense vegetation causes additional costs for cutting and pruning; different soil types, instead, lead to different cost of digging.

- **Terrain complexity**: geographic characteristics costs refer to flat terrain, but if the terrain orography becomes more complex further costs must be introduced. Non-flat terrains involve indeed more and higher towers. Such terrain complexity can be evaluated by calculating the average terrain slope.

- **Wind speed**: wind maps allow to define if in a specific location, the towers need a reinforcement to withstand stronger mechanical stress caused by an high wind speed.

- **Altitude**: altitude brings problems related to the icing and lightning risk which require more expensive security components.

- **Obstacles**: natural (rivers, channels) or artificial (roads, railways, telecommunication and power lines). Additional costs to cross them must be taken into account.

Define wisdom absolute values for each of these aspects is a complex task because it requires high level of experience and knowledge about technologies and in electric lines construction. They change a lot country by country, therefore also a deep knowledge of specific national markets becomes fundamental. In developing countries the cost of labour is usually low, a small fraction of the budget for the construction, but since the majority of components are produced in the northern part of the world, costs of shipping are really high and can increase their cost up to 400%.

This emphasizes that making a universally valid estimate of absolute value of those voice of cost, would be impossible beside little meaningful. A time by time setting, in relation to the chosen target region, should be preferred.The constructors themselves tend to make an estimate of the costs on the single plan, rather than refer to tabulated values.

A third typology of cost concur to the overall amount of the electric line expenditure. It is called *Direction Change Cost (DCC)* and aims to take into account the additional costs associated with deviation tower when a nonstraight path is followed. They concern technical complications introduced by curves which could be relevant for transmission lines, but negligible for MV and LV lines.

With the exception of *DCC*, all of the costs are expressed as costs per kilometers. The transition cost $f(l_k)$ between two neighboring cells p_k and p_{k-1} can be computed as shown by Monteiro et al.

$$c^{LU}_{p_k,p_{k-1}} = (c^{NGC}_{p_k} + c^{SC}_{p_k,p_{k-1}} + \frac{c^{TCC}_{p_k} + c^{TCC}_{p_{k-1}}}{2}) \tag{1.16}$$

$$f(l_k) = d * c^{LU}_{p_k,p_{k-1}} + c^{DCC}_{p_k,p_{k-1},p_{k-2}} \tag{1.17}$$

Where: *LU* stands for *"length unit"*; d is the distance between p_k and p_{k-1}; $c^{SC}_{p_k,p_{k-1}}$ the cost associated with the local slope of the terrain and $c^{DCC}_{p_k,p_{k-1},p_{k-2}}$ those related to direction change. Note as TCC are calculated as the mean value between TCC of the two connected cells.

A different approach has been adopted in *Onsset*, which does not plan the electric grid, but exploits the geographic information about the chosen area to make valid estimates on the cost of connection to the main grid. Instead of assign values of cost, it prefers to define a global *penalty factor* which will increase the cost of connection. The information gathered from the ground (Distance from the road, distance from the nearest substation, land cover, altitude and slope) are classified and combined together, taking into account the different impact of each aspect on the final cost.

$$\begin{aligned}\text{Combined Classification} =& 0.05 * \text{Distance from road class}\\ &+ 0.09 * \text{Distance from substation class}\\ &+ 0.39 * \text{land cover class}\\ &+ 0.15 * \text{Elevation class}\\ &+ 0.32 * \text{Slope class}\end{aligned} \tag{1.18}$$

As eq.1.18 show, Onsset considers slope and land cover much more impactful on the line cost than for example the distance from a road. This strategy, is therefore rooted on the analysis of the relative weight of one voice against the others: despite certainly less detailed, it is more suitable for a commercial use since the user must provide only the basic line cost per unit of length.

1.6 Demand assessment

In energy planning, particularly in rural electrification planning, assessing the demand is a fundamental process which requires high carefulness and socio-dynamical analysis. This means quantifying electric use on consumers side in terms of energy consumed and/or power required. The bigger issue is linked to the fact that the real nexus between energy access and socio-economic development is still being investigated. When dealing with rural electrification, intervention contexts are generally areas where, due to the lack of electricity, people do not have any kind of electric appliance. This factor coupled with usually very low financial capacities of rural households makes assessing even the short-term energy needs a truly complicated task. Additionally, being electrification's first goal to drive development, the outcome is supposed to be higher financial capacities which generally leads to the acquirement of the new appliance which on its part leads to higher energy demand. Thus, considering the off-grid systems limited capacity, in addition to the baseline load demand evaluation, a meticulous load forecasting has to be performed in order to predict how the load is going to evolve during a previously defined timeframe. Since that every class of appliance has its peculiarities in terms of load over a one-day-long window, the evident difficulty in this is anticipating which would be the appliances consumers will adopt over the years, a process strongly depending on region-specific sociological patterns. The best way to make such an analysis would pass through analogous projects' historical data. Unfortunately, being the sector relatively new, this kind of data generally does not exist.

The approach is therefore based on a combination of context analysis, data gathering and statistical methodologies.
Consumers are generally divided into three main categories: households, public services, and productive uses. For each of them, a list of possible livelihood needs with the linked appliances and specific power and energy needs is drafted. The total energy need for a portfolio of consumers is:

$$E_c = \sum_{j}^{User\ class} N_j \cdot (\sum_{i}^{Appliances} n_{ij} \cdot P_{ij} \cdot h_{ij}) \tag{1.19}$$

Where:	**i**	:	Type of electric appliance
	j	:	Type of User Class
	N_j	:	Number of users within class j
	n_{ij}	:	Number of appliances of type i within class j
	P_{ij} $[W]$	:	Power rate of appliance i within class j
	h_{ij} $[h/day]$	:	Functioning duration of appliance i within class j

$\sum_i n_{ij} \cdot P_{ij} \cdot h_{ij}$ represents the energy consumed by the single user i $\in$ class j; whereas $N_j \cdot (\sum_i n_{ij} \cdot P_{ij} \cdot h_{ij})$ is the overall energy needs of the class j.
This formula is useful to have broad indications about the order of magnitude of the energy need of a system but is not enough to do an accurately defined analysis which should end up with the definition of the most suitable power system configuration. Advanced estimates require the introduction of the concept of load curve, defined as the electric power required as a function of time. Its estimation passes generally through a bottom-up approach, in which each appliance within each user class is initially treated in a separate way, with a gradual aggregation ending with a unique whole load curve. Three common methodologies are exploited; they are shown below ordered with an increasing precision and resulting increase in time-consumption level.

Basic Approach Load curve is evaluated in the most direct way:

$$P(t) = \sum_{j}^{User\ class} N_j \cdot (\sum_{i}^{Appliance} n_{ij} \cdot P_{ij}(t)) \leftarrow \begin{cases} t \in Fw_{ij} \rightarrow & P_{ij}(t) = P_{ij} \\ t \notin Fw_{ij} \rightarrow & P_{ij}(t) = 0 \end{cases} \tag{1.20}$$

Fw_{ij} → Functioning windows in a day of the appliance ij

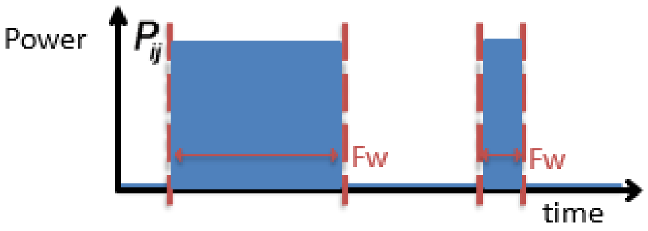


Figure 1.19: Example of daily functioning windows of a generic appliance

This approach has the advantage of being really simple and straightforward. It can be applied when having highly detailed information about exact functioning times. It gives the possibility of differentiating profiles based on cyclical specificities like weekends and seasonality of appliances usage. The main disadvantage is that it considers all the n_{ij} appliances switched ON at the same time: this implies over-estimation of the load Power Peak, and usually significant discrepancies between the estimated and the real profile, particularly when dealing with lowly-predicable appliances' frequency and duration of usage.

Intermediate Approach The second approach tries to manage these last considerations substituting the previously used functioning duration with the notion of load factor, defined as:

$$f_{L,ij} = \frac{E_{C,ij}}{24h \cdot P_{ij}} = \frac{h_{ij} \cdot P_{ij}}{24h \cdot P_{ij}} \tag{1.21}$$

$$h_{ij} = f_{ij} \cdot 24h \leq \sum duration(Fw_{ij} \tag{1.22}$$

f_{ij} is the *load factor* of the appliance ij The reasoning behind is that an appliance

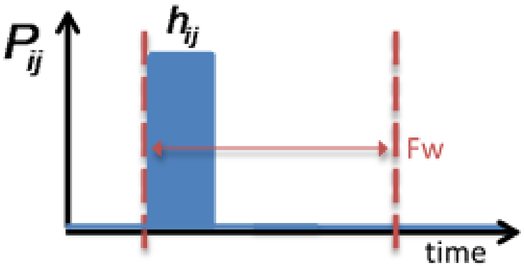


Figure 1.20: Functioning time of a generic appliance

will not be switched ON during its entire functioning window (see figure 1.20), but due to the intrinsic variability relying on its usage it is not possible to predict when it will occur (figure 1.21). Consequently, average power is estimated, and the accounted total energy is spread over all windows durations as shown in figure 1.22.

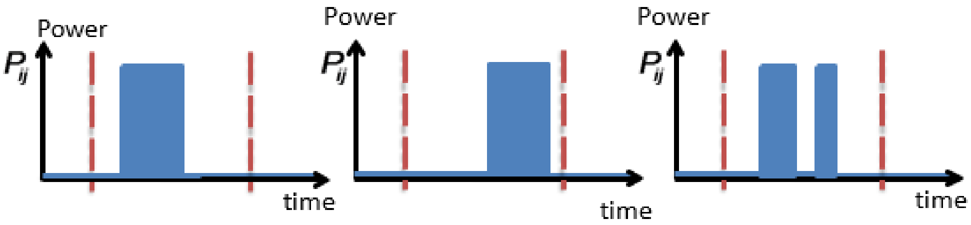


Figure 1.21: Example of variability in functioning profiles related to the same appliance

$$E_{L,ij} = P_{ij} \cdot h_{ij} \tag{1.23}$$

$$P_{AV,ij} = \frac{E_{L,ij}}{\sum duration(Fw_{ij}} \tag{1.24}$$

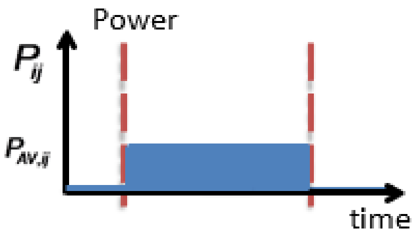


Figure 1.22: Average power of the appliance ij in the selected functioning window

The result is a load curve with an underestimated Power Peak, because even if like in the previous approach appliances are considered switched ON during all the functioning time, they are seen as operating at their nominal power, but at an average power taking into account the difference between the functioning window of an appliance and it actual functioning duration. Moreover, even if indirectly, it considers the coincidence relationship between appliances belonging to the same classes.

Advance approach Ascertained that both first approached implies high degrees of approximation, a third more comprehensive methodology is generally adopted. It addresses switch statuses uncertainty with probability using a stochastic method and defines it accordingly with a probability distribution, namely the Specific Probability Distribution Functions. For instance, if the appliance is actively involved in the power peak creation (figure 1.23a) it will be sampled with a normal distribution. If instead, it does not contribute to peak (figure 1.23b), it is sampled as uniform distribution. Secondly, the frequency and duration of usage could be modeled in or-

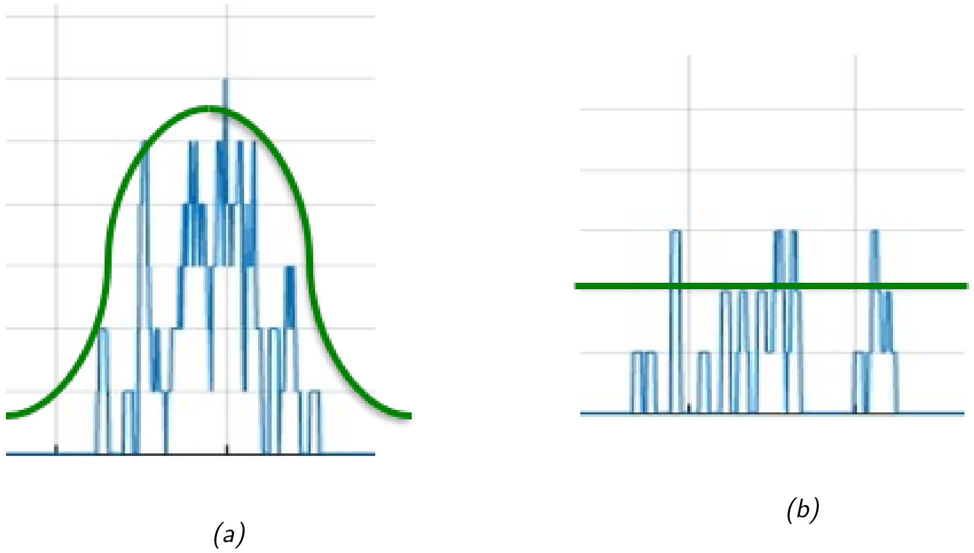


Figure 1.23: a) Appliance normal distribution sampling, b) Appliance uniform distribution sampling

der to simulate their functioning. The key addition is the fact that a new parameter is added, which is the minimum working time one appliance has been switched on: D_{ij}.

$$\sum duration(Fw_{ij}) \geq h_{ij} \tag{1.25}$$

$$f_{C,j} = \frac{PP_j^{Actual}}{\sum P_{ij}} = \frac{PP_j^{Actual}}{PP_j^{MAX}} \tag{1.26}$$

$$f_{L,j} = \frac{E_{L,j}}{24h \cdot PP_j} \tag{1.27}$$

Finally, the fact that generally there is an inter-class linkage between appliances usage, is taken into account correlating the Load Factor $f_{L,j}$ to a new parameter called Coincidence factor $f_{C,j}$ and the number of users. Combining these three innovative innovations, through an iterative process a load curve for each user class. The fundamental equations and the iterative process is represented below:

$$f_{L,j} = \frac{E_{L_j}}{24h \cdot PP_j^{Actual}} \tag{1.28}$$

$$f_{C,j} = f(N_j, f_{L,j}) \tag{1.29}$$

$$PP_j^{Actual} = f_{C,j} \cdot PP_j^{MAX} \tag{1.30}$$

This is the overall correlation linking together $f_{L,j}$, $f_{C,j}$ and PP_j^{Actual}. The main advantage of this approach is that its input data are, even being numerically more than the one required by the previously analyzed approaches, easier to be estimated thanks to direct correlations between appliances' usage peculiarities. Moreover, it gives a more accurate estimation of the power peak of a load profile. Specifically related to assessment in rural contexts, from one side it implies relatively easier questions to be inserted in a survey, and from the other side it is more resilient against the high uncertainty of load demand in unelectrified or partially electrified rural areas.

1.7 Energy Generation sizing

Hybrid microgrid systems are composed by multiple distributed resources working in a parallel configuration. They are generally the most cost-effective solution to electrify remote rural areas laying far from the grid.

Hybrid systems are usually based on locally available renewable energy sources and for this reason the big deal in optimizing this kind of systems is the necessity of sizing a power block relying on mostly unpredictable sources. Their goal is supplying energy to communities or rural income-generating activities, thus power production and demand need to be constantly aligned.

Two main solutions can be used to overcome the problem of unpredictability: storage systems and backup gensets. The first ones allow to accumulating excess energy in order to use it when resources are not available, whereas the seconds increase flexibility by adding a non-intermittent power sources which can be switched on when particularly-high peaks occur or, again, when energy is needed in concurrently with lack of renewable energy resources availability. The result is a reduction of the the otherwise very high costs linked with these technologies.
It is intuitive how control systems are paramount in managing such systems, in order to continuously match production and demand and achieve a veritable techno-economic optimization. Hybrid microgrid systems can operate not only in isolated/island mode, but also in a grid connected one in which energy can be imported or exported from and to the main national grid, depending on its lack or excess and even price differences.

Chapter 2

State of Art - Tools for rural electrification planning

Currently, several tools are available on the market helping a broad-spectrum of stakeholders planning energy electrification strategies[41]. Each of them addresses one or multiple issues in energy planning in rural areas of the developing world, through multidimensional perspectives. They have been developed by different entities, with diametrically opposed characteristics, for instance:

- developed for public or private sector;
- proprietary or open source license;
- top-down or bottom-up approach;
- regional-specific or global dimension;

GISEle is the result of an accurate analysis through which we defined which were the important gaps missing in this tools portfolio. With GISEle we wanted to open a new trend, bringing from one side analytical innovation, and from the other promoting the development of an integration platform enhancing synergies among tools and identifying the direction we all need to aim, in other to solve the three components of the energy conundrum: energy access, energy security, and climate change.
A resume of the main aspects characterizing the tools here described, are depicted in table 2.1 and 2.2 at the end of this chapter.

2.1 REM

The Universal Energy Access Research Group at MIT in collaboration with the Comillas University in Madrid is working on a computation model to approach the problem of determining least-cost electrification modes for rural electrification, which is called *Reference Electrification Model* (REM). Despite still in developing phase, it has been already tested helping to plan electricity networks in India, Colombia, Kenya, Rwanda, Uganda and other developing countries. Its main characteristics are:

- Broad spectrum:
  Differently from most of the available tools REM does not focus on a single activity but embrace in a single tool all the steps required to design a new energy network. It provides an answer to which is the best electrification mode (grid connected, micro-grid or isolated system), estimates cost of electrification and makes a preliminary design of recommended system including a wide range of design and technological choices.

- Works with geo-referenced data considering the location of individual loads, not only the indicative position of micro-grid site.

- Includes a clustering process to not connect all the customers in a single energy system, but divide them in groups which will be individually analyzed and classified as candidate off-grid systems or grid extension projects.

- Static model:
  It considers a single future year, and produces system designs and cost estimates based on serving the electricity demand in that year. However, it takes into account some year-to-year effects in a simplified way in order to estimate the lifetime performance of a system (degradation in solar panel and battery performance)

- allow that the demand is not served as a whole, because it may produce too expensive solutions.

In most of these aspects REM resembles what GISEle aims to do.
Its main weaknesses resides in the spatial design of the electric grid and in the limited variety of energy resources considered. In the REM version described by Douglas Ellman in [42] only two types of energy sources are considered: solar photovoltaic (PV) and diesel generator sets. The lack of wind and hydroelectric energy in the set of generation technologies is important and strongly affects the results. This is true especially for hydropower since, due to its high reliability in production, when a source proximate to the target area is available it is often exploited.
Such aspect is related to the developer's choice to integrate the design of local generation into the broader algorithm of REM with a self-developed code, instead of relying on mature and well-maintained commercial software. It allows to automatically produce local generation designs for a large number of systems spread over a large area, since the sizing step is part of the main code, but introduces the risk of coarser solutions. For what concern the model the electrical network, instead, REM

does not rely on it's own algorithms but leans on a specific algorithm called RNM. It takes account of many of the technical aspect involved in the design of the grid, but less attention has been given to the routing process across the territory as will be better discussed in the next paragraph. As it will be described later on, GISEle includes the generation-sizing process within its code too, so to be able to replicate it iteratively each time it required; however, the involvement of a well structured, open-source python based tool available in literature has been preferred to the development of a new code.

2.2 RNM

The Reference Network Model (RNM) is a very large-scale planning tool, designed by the *Instituto de Investigation Tecnologica* of Comillas University, which plans the electrical distribution network using GPS coordinates and power of every single costumers and distributed energy sources (DER) [43]. It designs the high, medium and low voltage networks, planning both substations and feeders. In doing that it makes considerations and assumption about:

- TECHNICAL ASPECTS: Voltage limit, capacity constraints, continuity on supply.
- GEOGRAPHICAL CONSTRAINTS: forbidden ways through (such as lakes or nature reserves) and a street map which is automatically generated considering the input location of customers

RNM has two available operating modes which allow to design networks in both *greenfield* and *brownfield* environment. The former is used to sketch networks in virgin territories, lacking any kind of electric infrastructure: it designs the new grid by considering the interconnections with transmission substations as the supply points together with distributed generation (DG) connections. The latter, instead, is devoted to network expansion purposes to accommodate additional demand: it takes the existing distribution network as input and obtains the required reinforcements and new facilities to connect the expected new loads and DG connections [44].

The approach proposed by GISEle will results in a more detailed topological analysis of the target area since, together with lakes, natural reserves and roads considered by RNM, it takes also into account the impact of rivers, ground slope and land cover in determining the optimal path of grid connections. Nevertheless, the first version of GISEle described in this lecture, has not been equipped yet with calculus related to the grid electrical balance involving resistances, reactances and bus voltage.

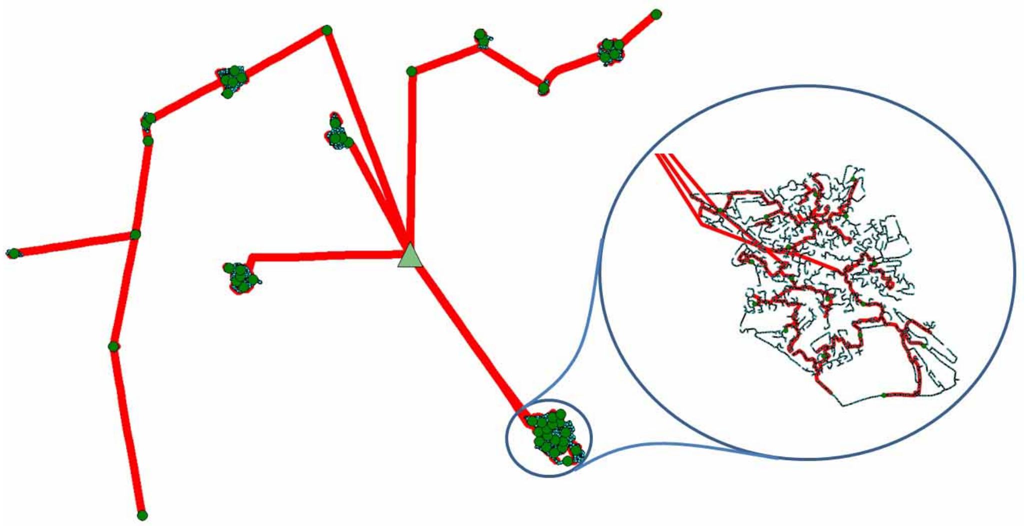


Figure 2.1: Example network built with RNM in rural areas

Thick red lines are Medium Voltage (MV) feeders, thin black lines are Low Voltage (LV) feeders, the green triangle is the HV/MV substation, green circles are MV/LV Transformer substations and small points are LV customers.

2.3 ECOWREX

ECOWREX is the platform developed by observatory for Renewable Energy and Energy Efficiency of the *Economic Community of West African States (ECOWAS)*[45]. It is aimed at improving existing knowledge and mitigating information barriers towards the development of the energy sector in the ECOWAS region. The tool has two main objectives:

- Provide targeted, timely and statistical information on the energy resources (especially in the field of RE and EE) including RE resources, policies, projects, power plants and other relevant information about the ECOWAS Region, to support in decision making;

- Build up a network of energy experts and cooperation among key local and international players to share knowledge and experience on best practices and technical know-how from around the world;

ECOWREX basically is a web-based GIS visualization platform delivering key data, analytics and insights for high-level stakeholders, supporting investment planning. Its high usability, the GIS based approach and its promotion of data sharing which characterize ECOWREX work in its support, but, on the contrary, it is only addressed to West Africa area and it is not suitable for data management but only for visualization purposes.

2.4 OnSSET

OnSSET (*Open Source Spatial Electrification Tool* [5]) is a bottom up optimization energy modeling tool, that estimates, analyzes and visualizes the most cost-effective electrification strategy.
It is a veritable software, with an open-source license and developed in Python. The tool relies on a geospatial approach, dividing the whole study area into a number of cells proportional to the desired resolution of the problem. By overlapping a multitude of layers, each corresponding to a dataset, the tool is able to return to the user the best electrification strategy among the stand-alone, mini-grid and grid extension options, and in case of an off-grid output which technology is the most suitable for each location.
The great innovation it the use of georeferenced data is the possibility of promoting region-specific policies. Thanks to the huge progresses in the computational field, we can now relatively easily deal with huge amount of data, so called big data, and use them in order to deal with complex multi-criteria decision making problems.
One of the issue we highlighted using OnSSET was the fact that each cell was treated as independent entity, so, for example, a very useful layer like the population density is not being used at its full potential, because during computations the tool identified populated cells, but was not designed to really find and deal with of densely populated areas, which implies looking at adjacent cells. A second weakness is linked to the power block sizing: the energy demand is the result of simple multiplication between the estimated population of a cell and the desired *ESMAP Multi-Tier Framework for Measuring Energy Access* in a 2030 horizon, so it's not a real optimization because from one side we don't have a realistic load profile and from the other side the energy resources too are yearly average values. The consequence is that a sizing output with an high degree of approximation, there is no space for energy mixes, so the LCOEs can be sensitively far from the real value. A third problem is that little space is given to the grid infrastructure, both from technical and economical points of view, even though the asset may have strong impact in such systems economics, especially in isolated rural contexts. In any case, there is no doubt that OnSSET is one of the best attempts as energy planning tools and has been from the beginning the reference in terms handling spatially distributed data.

2.5 Homer Energy Pro

As stated by the name, *Hybrid Optimization of Multiple Energy Resources*[5], the main domain of application of this software is the power block sizing of off-grid energy systems. Its high reliability in terms of techno-economic analytical concep-

tion, and an even higher level of user-friendliness make him the world leader in the distributed generation and microgrid modeling sector. Nonetheless it has some weaknesses that limit its degree of application.
First of all it is a proprietary software, so the source code is inaccessible, making it de facto impossible to be integrated it into a dynamic and automated sizing process. Secondly, although allowing strong flexibility in terms of technology, it focuses only on the power production side, totally neglecting any required distribution or transmission grid, which leads to often highly underestimated Levelized costs of energy.

2.6 RE2nAF/PVGIS

Renewable Energies Rural Electrification Africa is an open tool firstly developed for visualization purposes, with an intuitive web interface, in which a series of key data have already been uploaded, like infrastructures such as roads, power plants and grid, and others like population and solar irradiation. It gives the possibility, one country at time and based on different combinations of specific diesel and solar PV peak power costs, to divide the area into two different profitability zones for solar and diesel generators.

2.7 Network Planner

Columbia University developed this tool which has the purpose of effectively planning electrical networks, but only from a top-down approach. Standalone grids are only estimated as a result of a combination of community population and extension, and the in any case all grid are results of a simple bidimensional network problem in which, firstly, despite having for example very different populations, every point seems to have the same weight, and secondly lines do not consider morphology complexity, and connect directly each consequent point leading both to unrealistic physical representations of the most appropriate infrastructure and too high approximations in terms of costs due to grid lengths basically corresponding to aerial distances. Moreover, its structure makes it suitable only for users which are at the end of a whole study, as the required input data are not basic ones, but already structured data deriving from previously made deep electrification planning analysis. Nonetheless, the concept was highly valuable, but at the moment the whole project seems having been abandoned from more than five years.

2.8 Calliope

Calliope has been developed by KTH of Stockholm and self-defines as *"a multi-scale energy systems modelling framework focusing on flexibility with high spatial and temporal resolution"*.
Its first strength is to be fully open source, written in Python language but designed in order to deal mainly with easily writable text files as .csv and .yaml ones. These characteristics allow it to be easily managed also by coding illiterate users, but most importantly it makes it a building block for higher level frameworks promoting valuable synergies. Moreover even if not directly working with GIS data, it is designed to provide spatial optimization allowing to add a key dimension to each component of the energy system, from resource to users, basic requirements when we would like to optimize interactions and flows between nodes of energy system.
Calliope can deal with both small and large scale systems optimization, making it a suitable tool from both micro-grid scales, in which the nodes could correspond for instance to households and PV power generators, and multinational energy systems in which the nodes could represent entire cities and multi-GW sized powerplants.
Some highlightable weaknesses concern the fact that electric power and load flows are not taken into account in the grid sizing and a relatively poor dynamics: an example is the fact all factors are exogenous, for example the grid configuration needs to be already known to the user, making it at the moment a simple standalone application, limiting its true potential.

2.9 Off-Grid Energy Market Opportunities

[46] This open web-based tool also has insights delivering purposes. For each country it is possible to select different socio-economic criteria on which the evaluation will be based, such as energy-reliant and strategically relevant facilities, population density and logistics infrastructures. Secondly the user is asked to indicate the supposed market penetration and the per household average revenue per year connected to the offered services in the selected area. The final output is a map resulting for the previously stated constraints with highlighted their potential in terms of annual total revenue estimate.
This tool has the advantage of being totally market-oriented but has a big weakness in relying on a very limited portfolio of data, which leads to results which from one side are very poor in terms of significance, and from the other suffer from high degrees of approximation.

2.10 LoadProGen

This tool, developed by Politecnico di Milano, is one of the best on the market in terms of assessment of Load Profiles for multi-consumer-class systems and incorporates the third of the approaches previously described in the energy assessment section. LoadProGen is developed as Matlab packages so it is theoretically free but indirectly requires Matlab license which instead is not. It has a basic but very intuitive user interface. It gives the possibility to define several user classes, and for each one to define appliances. Besides, each appliance needs:

- *Nominal Power Rate [W]*;
- *Functioning cycle [min]*: Minimum duration of the functioning cycle for a defined appliance;
- *Functioning time [min]*: total functioning time, during a day, for the considered kind of appliance;
- *Random variation of functioning time [%]*: uncertainty associated to the total functioning time (Rh);
- *Random variation of functioning window [%]*: uncertainty associated to the functioning windows (Rw);
- *Specific Cycle [ON/OFF]*: presence or not of a specific cycle for the considered appliance. When ON is selected, the nominal power of the appliance and the duration of its cycle cannot be changed but the shape of the power profile can be modified clicking on Cycle. Remember that the functioning cycle has always to be expressed in minute and it will not be considered during elaboration if a sample time longer than a minute is selected.
- *Functioning windows*: interactively selected along a day, in which timesteps (hours, minutes or seconds) the appliance generally operates.

When all classes and appliances have been configured the model can be run. LoadProGen gives the possibility to compute more than a single load profiles in order to give the user the possibility of appreciating the impact of high-uncertainty appliances on the whole profile and, in particular, in correspondence of peaks.

Tool	Description	Strenghts	Weaknesses
ECOWREX	Monitoring for investment support	- high usability - GIS based - Promotes data sharing	- Only West Africa area - Only visualization purpose
OnSSET	Forecasting for energy policy	-GIS based - Separation between electrification strategies	- Single cell approach - No sizing - No hybrid configurations
HOMER Energy Pro	Power generation resources optimization	- High reliability in techno-economical conception -high level of user-friendliness	- Proprietary software - Focus limited to power production
RE2nAF/PVGIS	System analysis for energy policy	- Open source - Intuitive web interface	- Focus limited to PV and diesel generation

Table 2.1: Resume of depicted tools A

Tool	Description	Strenghts	Weaknesses
Network Planner	Network planning	- Optimal grid planning	- Only bidimensional analysis, with no morphology account - No population density accounting - Only grid extension
Calliope	Multi-scale energy systems modelling	- Open source - Strongly GIS based - Multi-dimensional - Relatively high userfriendliness	- Exogenous grid infrastructure
Off-Grid Energy Market Opportunities	Market potential analysis for investment decision	-Open source - Strongly business oriented	- Limited data portfolio
LoadProGen	Load profile generation	- Userfriendliness - Resilient statistica formulation	- Open source code but Matlab based
REM	Support in electrification strategy planning	- GIS based - Multi-activity - Complete analysis	- Only PV and genset technologies
RNM	Network spatial planning	- High number of aspects considered - GIS based - Electrical balance	- Low detailed routing

Table 2.2: Resume of depicted tools B

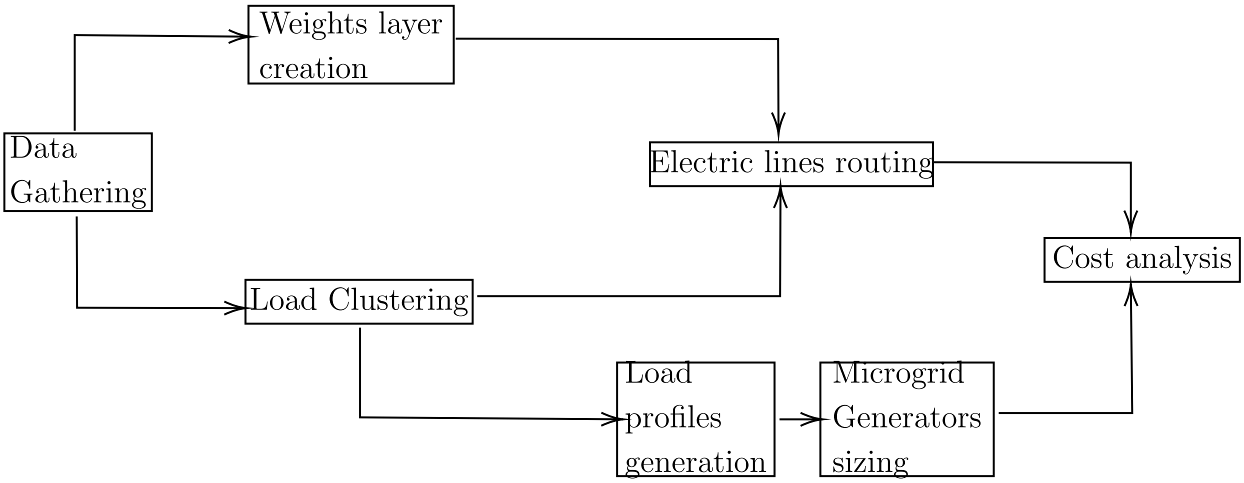
Chapter 3

Rural areas electrification: the approach proposed

The main aim of the thesis is to develop a complete procedure for the planning of rural electrification strategies, rooted on the analysis of spatial data, which runs through all the passages leading to the identification of the optimal techno-economic solution to bring the energy where it lacks. Exploiting the potentialities of GIS environment in which it operates, the approach proposed (hereinafter named GISEle) is able to consider also the spatial characteristics of the proposed solutions:

- The spatial distribution of consumers and generation plants;
- the detailed topology of the electric grid which would connect them together and, optionally, to the existing national grid;

Those aspects allow GISEle to supply more accurate evaluations about the final costs related to different possible solutions. What follows is the description of the method and the algorithms composing the procedure.



The procedure starts with the collection of all the data needed for the analysis: information about energy resources, population distribution and terrain characteristics

are gathered and combined together into a discrete representation of the target region made of a regular point mesh. The information embedded in each point are managed in order to define a *"penalty factor"*, index of the difficulty of building an electric line through that specific point. Alongside, *GISEle* accomplishes an analysis on population distribution. In this step a clustering algorithm identifies the densely populated area, and divides the population in groups called *"clusters"* which will constitute independent energy communities. Those isolated households and populated points, located in remote regions far from the defined community areas are classified as *outlier* and discarded. Combining the information about the spatial distribution of the households within a cluster, together with the ground characteristics expressed through the *"penalty factor"*, the electric lines routing algorithms designs the optimal, least-cost, electric network able to connect each household of the cluster in a unique energy grid. Furthermore, it considers also the opportunity to connect the cluster directly to the HV national transmission line, calculating costs and optimal routing path to realize such link.
Once the electric infrastructures have been defined, *GISEle* moves forward providing an evaluation of the cluster's energy needs generating daily load profiles and, combining the energy demand with the energy resources availability, defines the optimal configuration of the generation system able to reliably fulfil the energy request. Finally, *GISEle* conclude its works providing an economic analysis of the two possible possible electrification strategies:

- Isolated Micro-Grid
- Grid connected energy system

The choice is made evaluating the cheaper and most promising solution by comparing the respective values of LCOE.
Below, each steps is depicted in detail.

3.1 Data gathering

The first step is to collect all the data needed for the analysis of electrification strategies.
Spatial data are usually collected within datasets with different levels of details made available by several organization through internet platforms. *Energydata.info*[47], for instance, is an open data platform launched recently by The World Bank Group and several partners, trying to change energy data paucity. It has been developed as a public good available to governments, development organizations, non-governmental organizations, academia, civil society and individuals to share data and analytics that can help achieving universal access to modern energy services. Every day governments, private sector and development aid organizations collect

data to inform, prepare and implement policies and investments. Yet, while elaborate reports are made public, often the gathered data remain locked. This is true especially for very high quality data which required significant investments to be gathered. Furthermore, when it comes to underdeveloped countries, sometimes digital datasets does not exist at all.

Therefore, it is not always easy to find all the data and, if a high resolution is needed, probably they are made available only for a fee. The OnSSET website[48] published a list of available platforms where different categories of spatial data can be gathered. All the datasets exploited by GISEle are listed in table 3.1 with additional information about the data typology: raster or vector. The Administrative boundaries represent the reference layer of each activity described in this methodology: they define the target area to which the GISEle analysis is addressed. This area can be delimited by political borders like province, regions or even entire states or also by technical borders: for instance an unelectrified area enclosed within HV transmission lines. Usually, the datasets obtained do not perfectly fit the target area: some of them contain information at global level; others, instead, cover only part

DATASET	**TYPE**
Population distribution	Raster
Administrative boundaries	Vector Polygon
Existing grid network	Vector line
Planned grid network	Vector line
Electric substations	Vector Point
Roads network	Vector line
Water bodies (lakes)	Vector Polygon
Protected areas	Vector Polygon
GHI	Raster
Wind speed	Raster
Elevation map	Raster
Land cover	Raster
Hydrologic soil type	Raster
Monthly rainfall precipitation	Raster
Monthly mean Temperature	Raster
Monthly mean minimum Temperature	Raster
Monthly mean maximum Temperature	Raster

Table 3.1: Geo-spatial datasets for GISEle procedure

of the area of interest so multiple files must be combined together. Therefore, each dataset is clipped using the administrative boundaries layer as mask so to obtain information fitted on the desired area of interest. Looking at the list of geo-datasets depicted in table 3.1, it becomes evident the lack of the river network. As stated in chapter 1.2 introducing the concept behind the CN-method, in the assessment of energy resources the hydro-power potential is the most difficult to estimate.
The website http://gaia.geosci.unc.edu/rivers/ [49] offers detailed maps of the global river network with embedded information about width and depth. This dataset is not the result of direct observation, but it has been modeled by means of an hydrologic analysis of terrain morphology combined with satellites observation as explained by Andreadis et al. in [50]. Figure 3.1 shows an image of the resulting Mozambique river network. Despite the modeled rivers' map follows with good approximation the real course of the rivers validating the hydrologic model adopted, it provides only annual mean estimation of water flow rate which is not sufficient to make estimates about hydro-power potential. Water availability can change a lot during the year, especially in tropical regions, where seasons are distinguished in dry and wet: months of good hydroelectric production could be followed by months of shutdown caused by the lack of water, and this must be taken into account for a good analysis of energy resources. For our purposes, therefore, monthly values of water flow rate are needed, together with available hydraulic head.

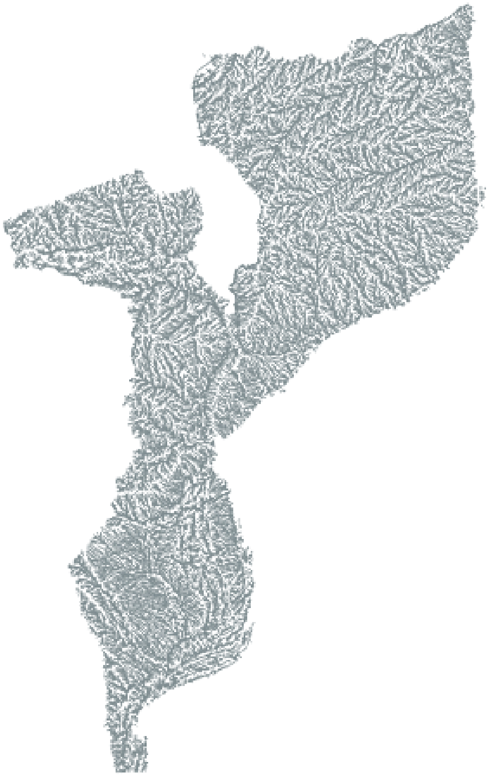


Figure 3.1: Mozambique river network

Hydropower potential assessment

In order to create a spatial layer representing the monthly hydropower potential in each point of the target region, three steps must be accomplished:

- Determine the water runoff, namely the amount of water neither absorbed by the terrain or by vegetation nor evaporated in atmosphere which, therefore, flows in surface.
- Determine how surface water moves across the territory conveying and giving rise to creeks which than becomes rivers.
- Make estimates about the hydraulic head available in a reasonable neighborhood of each river's point.

All of these tasks can be addressed combining and manipulating the information embedded in the datasets depicted above, within a GIS environment. ArcGIS provides several packages devoted to hydrologic analysis that makes the task easier; but since we chosen to implement the procedure in an open environment, all the analysis here presented can be replicated within QGIS.

Water runoff

Section 1.2 describes the way to calculate the surface water runoff through the *SCS-CN Method*. The same procedure is therefore applied and replicated for each month of the year. The following layers are needed:

- Land cover
- Hydrologic Soil Group
- Potential Evapotranspiration
- Rainfall precipitation
- Administrative boundaries

Differently from the others, potential evapotranspiration is not part of basic layers set because it is not always available, therefore it needs to be defined.
Among several equations to estimate PET, the Penman-Monteith equation 3.1 adopted by the FAO (FAO-PM) is currently widely considered as a standard method: it does not require estimations of additional site-specific parameters but its major drawback is its relatively high need for specific data for a variety of parameters

which are especially lacking in developing countries (i.e. wind speed, relative humidity, solar radiation, etc.).

$$ET_0 = \frac{0.408 \cdot \Delta(R_n - G) + \gamma \frac{900}{T+273} \cdot u_2(e_s - e_a)}{\Delta + gamma(1 + 0.34 \cdot u_2)} \tag{3.1}$$

ET_0	:	reference evapotranspiration $[mm \cdot day^{-1}]$
Rn	:	Net radiation at the crop surface $[MJ \cdot m^{-2} \cdot day^{-1}]$
G	:	Soil heat flux density $[MJ \cdot m^{-2} \cdot day^{-1}]$
T	:	Mean daily air temperature at 2m height $[°C]$
u_2	:	Wind speed at 2m height $[ms^{-1}]$
e_s	:	Saturation vapour pressure [kPa]
e_a	:	Actual vapour pressure [kPa]
$e_s - e_a$	:	Saturation vapour pressure deficit [kPa]
D	:	Slope vapour pressure curve $[kPa \cdot^{\circ} C^{-1}]$
g	:	Psychrometric constant $[kPa \cdot^{\circ} C^{-1}]$

The *Consortium for Spatial Information CSI*[22] compared the performance of four other temperature based methods to Penman-Monteith: Thornthwaite (1948), Thornthwaite modified by Holland (1978), Hargreaves et al. (1985), Hargreaves modified by Droogers and Allen (2002). From the analysis of results the Hargreaves model was chosen as the most suitable to model PET globally.

$$PET = 0.0023 * RA * (Tmean + 17.8) * \sqrt{TD}\ [mm/month] \tag{3.2}$$

	Tmean	:	Mean temperature of the month $[°C]$
Where:	TD	:	Daily temperature range $[°C]$
	RA	:	Extra-terrestrial radiation

TD is calculated as the difference between average monthly maximum and minimum temperature. RA, instead, represents radiation on top of atmosphere expressed in mm/month as equivalent of water evaporation.

In the available dataset solar radiation is expressed in $\frac{MJ}{m^2\ day}$, but can be converted in $\frac{mm}{day}$ with the following formulas:

$$Radiation\ [depth\ of\ water] = \frac{Radiation[Energy\ surface]}{\lambda \cdot \rho_w} \tag{3.3}$$

Where	λ	:	latent heat of vaporization $[\frac{MJ}{kg}]$
	ρ_w	:	density of water $[\frac{kg}{m^3}]$

Considering $\lambda = 2.45\frac{MJ}{kg}$ and $\rho_w = 1000\frac{kg}{m^3}$ the final equation results:

$$Equivalent\ evaporation\ [\frac{mm}{day}] = 0.408 * Radiation[\frac{MJ}{m^2\ day}] \tag{3.4}$$

Once all the datasets have been defined, they are firstly clipped using Administrative boundaries layer as mask; then the methodology described by the CN-method is followed to obtain a final runoff dataset which characterize each cell composing the target region. Figure 3.2, is an example of resulting layer.
Additional consideration:

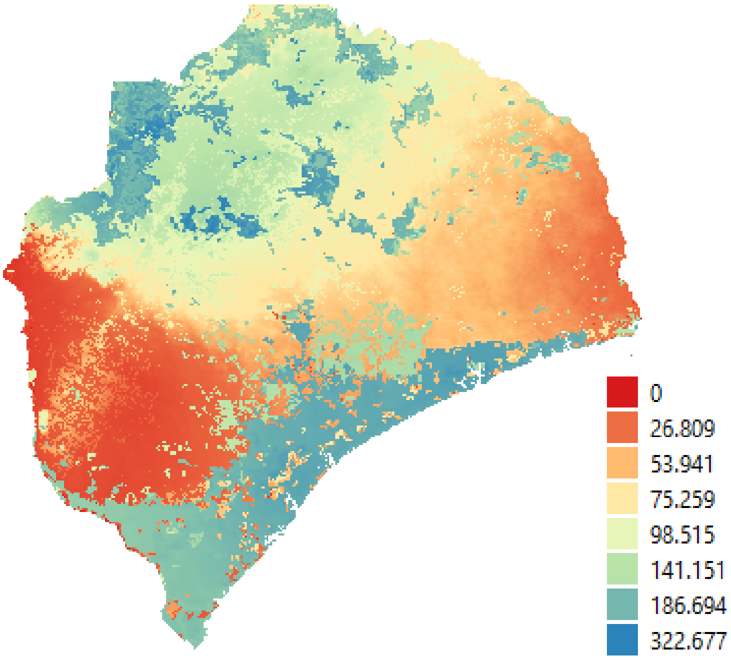


Figure 3.2: Runoff layer of Zambezia region in northern Mozambique

Legend's values are expressed in mm of water

- The SCS-CN method classifies land cover in 16 different typologies supplied by the USGS National Land Cover Database (NLDB) and shown in figure 3.2. In the event that a different classification has been adopted in the provided Land-cover dataset, a reclassification must be made.

- Same reasoning is applied to the Hydrologic soil group: the official classification divides the soil type in 4 different categories (A, B, C, D), but sometimes dual Hydrologic Soil Groups (A/D, B/D, and C/D) are given for certain wet soils that could be adequately drained. The first letter applies to the drained and the second to the undrained condition.

The curve number (CN_{II}) for each Land-cover/Soil-type combination is assigned following the table 3.2 built from [21].

National Land Cover Dataset Types	ID	Curve Number for HSG			
		A	B	C	D
Open water	11	100	100	100	100
Developed, Open Space	21	39	61	74	80
Developed, Low Intensity	22	54	70	80	85
Developed, Medium Intensity	23	77	85	90	92
Developed, High Intensity	24	89	92	94	95
Barren Land	31	77	86	91	94
Deciduous Forest	41	30	30	41	48
Evergreen Forest	42	30	55	70	77
Mixed Forest	43	36	60	73	79
Shrur/Scrub	52	35	56	70	77
Grassland/ Herbaceous	71	49	69	79	84
Pasture, Hay	81	30	58	71	78
Cultivated Crops	82	67	78	85	89
Woody Wetlands	90	100	100	100	100
Emergent Herbaceous Wetlands	95	100	100	100	100

Table 3.2: Curve Number Table

Hydrologic analysis and streams delineation

Once the estimate of the runoff water has been accomplished, we have to determine how it moves across the territory. Following the morphological characteristic of the terrain, surface water moves from one cell to another, accumulating gradually and forming streams and rivers.

A hydrologic analysis, starting from the elevation layer (also called *Digital Elevation Model DEM*, performs a study of the terrain shape, defining how water behaves inside it and providing as result useful information about the hydrology of the region like flow accumulation, watershed basins and rivers' paths. At the basis of almost all the calculations we performed is the layer of the runoff direction (flow direction).

There are several alternative mathematical models to determine the water behaviour which mainly divides in two groups: those that consider a one-dimensional flow (commonly called single flow directional algorithms) and those that consider a two-dimensional flow (multiple flow direction algorithms). In other words, the first considers that the flow of a cell, due to the slope of the terrain, moves to one, and only one, other cell, while the second considers that the flow can spreads to more than one contiguous neighbouring cell. An example of one-dimensional method is the D8: this is the method used by ArcGis and ArcHydro. The flux is calculated according to the greatest slope between the cell considered and its 8 contiguous cells. Differently, in the Multiple Flow Direction (MFD), also called FD8, according to the slope, all cells located below the cell concerned will receive a portion of the flow. From the DEM layer, the *Watershed function* available in QGIS calculates the slope and the direction which the ground is facing, and combining them with the runoff just calculated, produces:

- The *"Accumulation layer"*: containing the absolute value of the amount of overland flow that traverses each cell;
- The watershed basins in which the region is divided;
- The map of the rivers. All the cells in which the amount of accumulated water overcome a limit defined by the user are classified as "rivers"[1].

Meaningful values of accumulation are measured only in correspondence of rivers, indeed it is on a river path that we investigate to build a hydro power plant. Considering the accumulation and elevation values measured along the rivers' path, the following steps lead to the definition of hydropower potential.

1. Calculate the water volume flow rate in each river's cell:
$$Q = Volume\ Flow\ Rate\ [\frac{m^3}{s}] = \frac{\frac{Accumulation\ [mm]}{1000} * Cell\ size}{seconds\ in\ the\ month} \tag{3.5}$$
2. Define equidistant points along the river path: stated the elevation values, find the river point in a defined neighborhood with the minimum elevation so to obtain the *Available hydraulic head* Δh $[m]$.
3. Calculate the hydropower potential with the formula:
$$Hydropower\ Potential\ = Q * \alpha * \rho * g * \Delta h * \eta_{power-plant} \tag{3.6}$$

α : reduction coefficient

ρ : Density of water $[\frac{kg}{m^3}]$

g : gravity acceleration $[\frac{m}{s^2}]$

η : Conversion efficiency of the plant

Figure 3.3 rewiews all the steps described so far to produce the hydro-power potential dataset.

Figure 3.4 by means of a satellite picture witnesses how modeled river topology well reflects reality.

[1] Due to very unequal distribution of water precipitation along the year, the same analysis performed in different months will produce different river maps: whole rivers disappear during dry season because runoff does not reach the minimum threshold. To be coherent in the analysis, the river map produced in the most rainy season is considered

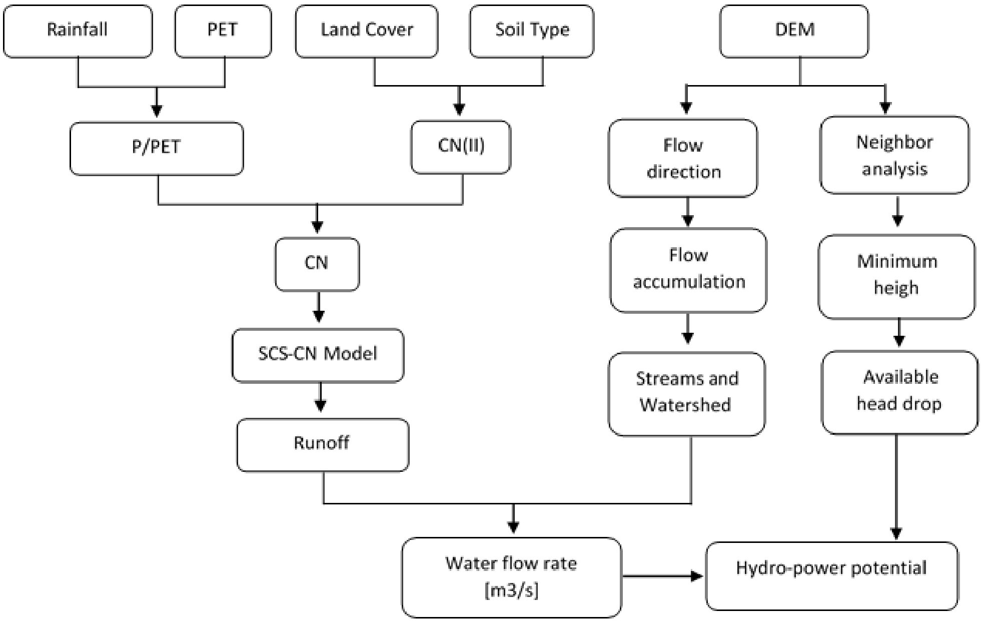
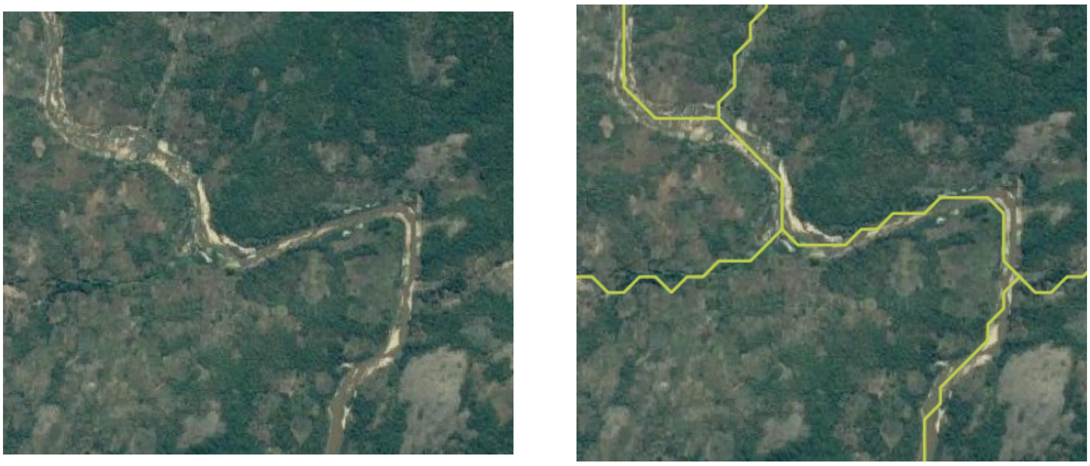


Figure 3.3: Workflow of the hydrologic model to estimate hydro-power potential

Hydrologic model validation

The introduction of a corrective coefficient α becomes necessary to deal with the errors in water runoff estimation introduced by the CN-method. The model adopted, indeed, has been originally developed to predict the ground response to single rainfall events in small agricultural or urban watersheds. Its application for runoff estimation to larger time intervals, like weeks or months, and wide regions, is possible and adopted in several hydrological models and research works ([20], [17]), but introduces inevitably strong simplifications and increases the error. Furthermore, it does not consider secondary aspects, difficult to measure, which affect the real runoff amount: underground sources, storage, water deployment for agricultural and social purposes. α derives from the comparison of the obtained results with real point measurements of monthly flow rate, resulting from antecedent measurements campaigns. GRDC [15], provide river discharge time series data from pretty much all over the world. Results of water flow rate, deriving from the application of the hydrologic analysis just described to selected regions of Mozambique and Tanzania, have been compared with 6 real measurement campaigns dating back to the second half of '900: one in Mozambique (Rio Messalo) and the others in Tanzania (Simiyu river, Ruvu river and Sigi river). Estimated flow rate are often greater than the average flow measured, but the definition of a meaningful coefficient α results difficult due to the high dispersion of measured values as shown in figure 3.5. In march 1971 the average detected flow rate was about 300 $\frac{m^3}{s}$, the next year lower than 50 $\frac{m^3}{s}$ and in 1973 150 $\frac{m^3}{s}$. Finally a coefficient $\alpha = 0.03$ has been considered appropriate.



(a)

(b)

Figure 3.4: Comparison between real river path (a) and modeled one (b)

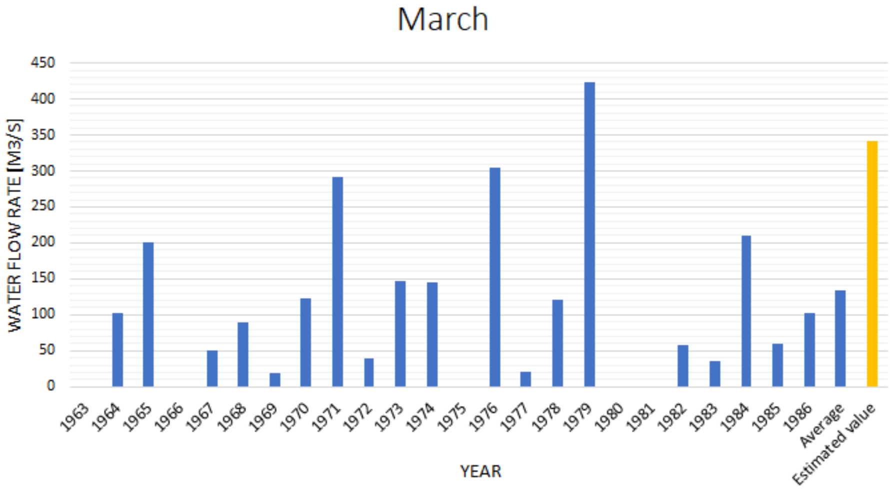


Figure 3.5: Results of march water flow rate measurements in Est. Nairoto Montepuz station along the Rio Messalo (MOZ)

3.2 Data Management

Once all the necessary datasets have been defined, they can be combined together in a discrete representation of the target region: a python routine, executable in the python environment of QGIS, has been written at this purpose. The user must provide:

- The desired resolution
- the EPSG code of the Projected Coordinate System to which reproject the datasets.

At first, all the datasets are clipped on region's administrative borders, resampled at the desired resolution and reprojected to the desired CRS. Then, the steps are the followings:

- *Nodes creation*: The population raster layer is converted into a point vector one. Each point is positioned in the center of a cell and take the cell's population value as attribute.

- *Slope*: by means of a specific QGIS function, from the analysis of DEM data, the raster layer containing values of slope in each cell is created.

- The information contained in each of the following raster file is transferred to the node mesh.

  - Elevation
  - Slope
  - Land-cover
  - River flow rate (If a point does not belong to a river the value is Null)

  Each node acquires as attribute the the value of the cell within which it falls.

- *Road distance* For each node, the distance to the closest road is calculated and assigned as attribute.

- *Water bodies and protected areas* An attribute states if a node falls within a lake or not. The same analysis is accomplished with regard to protected areas in which it is forbidden to build electric lines.

- *Substations*: Substations are localized through a point vector layer.
  With regard to planned transmission lines, a substation every 5 km is assumed; if the substations layer is not available, the same assumption is applied also to existing transmission lines. Nevertheless, these points do not necessarily match with nodes, therefore the presence of a substation is assigned as attribute to the node, as closest as possible to the real substation location.

At the end of the process, each point embeds information about:

Projected coordinates (X, Y)

Elevation

Slope

Land-cover

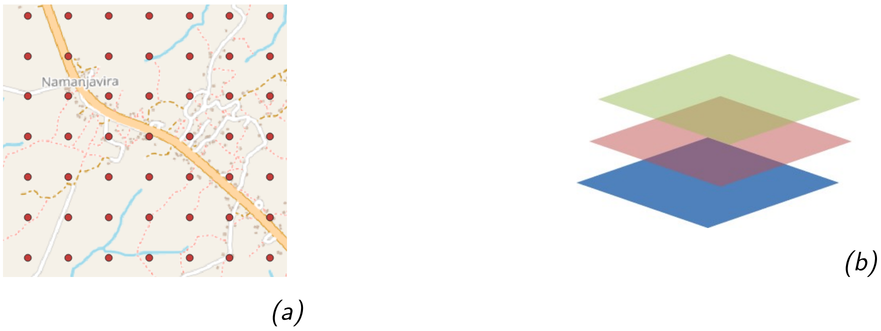
River-flow rate

Distance from road

Affiliation to water bodies or protected areas

Presence of electric substation

Datasets usually contain errors within them, representing lack of data, which are faced filling hole through interpolation of surrounding cells values. That presented in figure 3.6 is a view of the regular point mesh discretizing the target area.



(a)

(b)

Figure 3.6: View of the regular point grid over the village of Namanjavira (Mozambique), resulting from the dataset combination

Weighting process

The information gathered and embedded in each node, constitutes a set of the spatial aspects and ground characteristics which impact costs of an electric line realization. They contribute, in different extent depending on their values, to the increase in line cost per unit length with respect to the ideal condition. To be coherent, and avoid to consider voices that could be too specific and difficult to collect not all the technical and geographical aspects depicted by Monteiro at al.[11] have been taken into account inside the analysis. Furthermore, many of them are important in HV transmission lines planning, whereas become negligible for MV lines. Therefore, costs connected to direction change, corrosive environment, icing problem, wind rafals and the interconnections with communication lines and roads were not considered. For the same reasons, since costs changes a lot country by country, the concept of a *"penalty factor"* applied to a base cost previously set has been preferred to provide an absolute value to each voice of cost. In our approach the following contribute have been considered:

- **Distance from road**
- **Land cover**
- **Slope**
- **River flow rate**
- **Water bodies**
- **Protected Areas**

In each node, these values are analyzed and combined together to return a *penalty coefficient* affecting the electric line which will run through it. The penalty coefficient is initially set to 1 and is augmented by the contributions of each single voices. The final value responds to the question: "How may times more than the basic would cost my electric line? "

$$Penalty\ factor = 1 + \sum_{i}^{penalty\ aspects} penalty_i \tag{3.7}$$

Roads are fundamental for the logistic: poles, electric components, tools needed for the excavation works are all brought to the site by means of roads. If roads lack, they usually must be built to accomplish the work. The further is a point from the road network, the higher will be the costs to connect it. Nevertheless, it has a superior limit that we set equal to 6 times the basic cost, reached at 1 km from the road. Figure 3.7 shows its profile.

Slope penalty, instead, has an exponential profile: the steeper a terrain is, the harder will be to reach and build an electric system on it. The exponential curve is designed to return a value of 1 at a slope of 35 degrees as shown in figure 3.8.

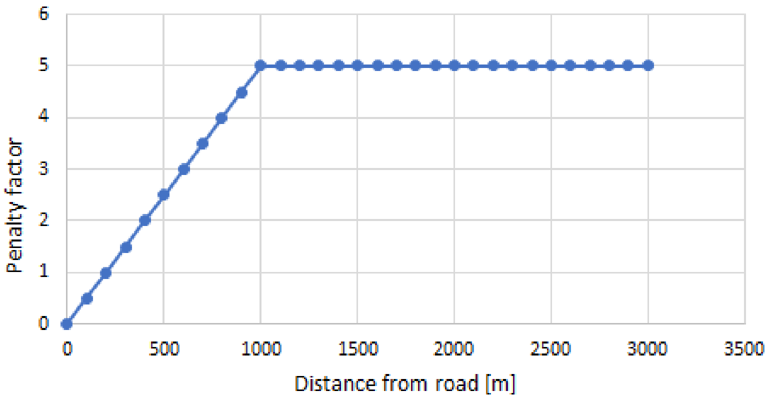


Figure 3.7: Penalty factor associated to road distance

All the other penalty components are not described by means of functions: a value of penalty is directly assigned to each class of terrain as table 3.3 illustrates.

The importance of the road proximity must be underlined: ground characteristics do not matter if road is really close to the site[2].

If road distance < 100m → Penalty coefficient=1

What shown in figure 3.9, is the visual result of the weighting process applied to the administrative place of Namanjavira, chosen as case study of our work.

[2] The threshold value must be set in relation to the resolution adopted. 100m has been chosen for the analysis of datasets with cell size = 200x200m.

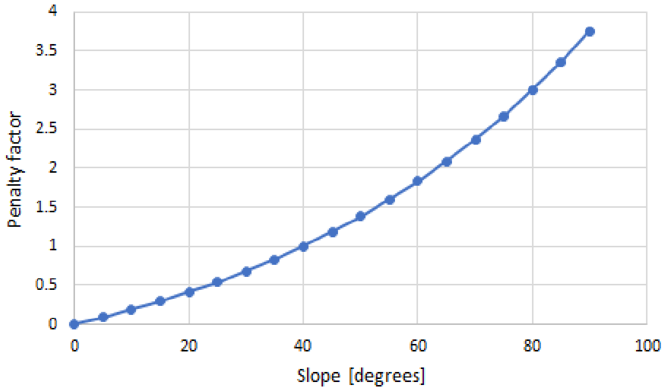


Figure 3.8: Penalty factor associated to road distance

3.3 Clustering

After having collected and processed the necessary data, the first step consists in clustering the population and hence the electric loads of the entire area. This process has to be done considering the population attribute of the previously created regular points grid. It is an exploratory process with which, the algorithm needs to browse the entire target area in order to find densely populated area suitable for electrification.
The main objective of the clustering process is to move beyond the basic approach with which each cell is considered by itself, neglecting the strategic value of closely located highly populated areas: the algorithm will proceed by identifying valuable groups of cells instead of only one cell at a time. Using a simple map it would have been possible to identify as intervention areas all the cities avoiding this clustering step. This top-down approach would have had some criticalities:

- In remote rural areas there are often highly populated areas which are not formally recognized yet;
- Knowing the position of a city and even its population is not enough to design an adequate infrastructure: firstly because its internal population density can considerably vary from one area to another, and secondly the shape and effective extension are not defined and vary from one year to the other.

The population clustering, moreover, is a fundamental step of GISEle procedure to reduce the computational burden of our model. The routing procedure applied to the whole region of interest, indeed, would require a big amount of time because of the high number of points to be considered; furthermore it would create an electric line connecting each populated point, independently from its location, included the most isolated ones. Clustering, solves these issues, by classifying those single isolated points as outliers, and grouping all the others in different clusters which will

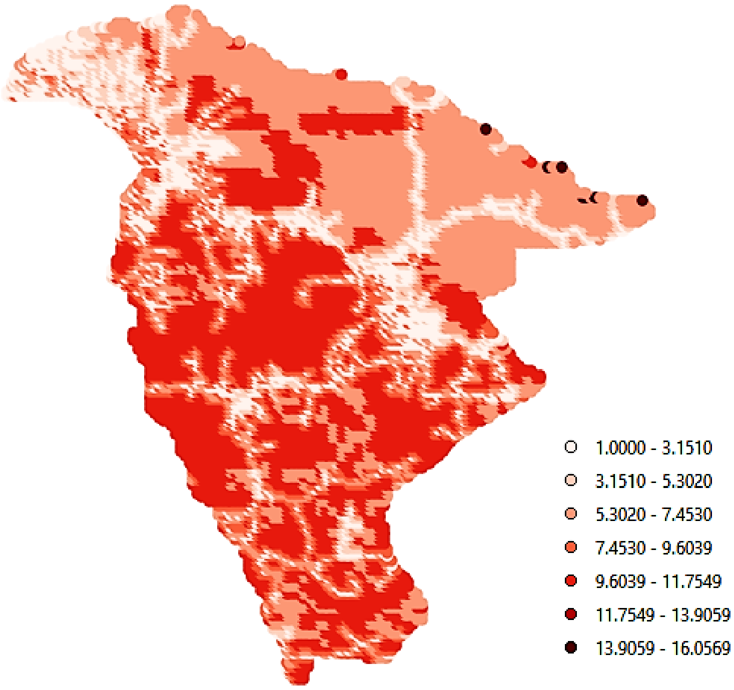


Figure 3.9: Penalty factor distribution of Namanjavira administrative place

be considered as unique, separated energy communities. The following routing algorithm will than be applied to each cluster, drawing its specific electric grid topology. In this way, the amount of data involved in each routing process, ad so the memory required to manage them, will be limited. Clustering algorithms are fast: generally they outweigh routing algorithms in terms of computational time. Therefore their application speeds up the global procedure. Considering the analysis made in the Clustering's State of the Art Section, a specific methodology had to be adopted in designing the GISEle model. Specifically, we previously discussed several clustering algorithms categories. Based on the already explained specific characteristics and weaknesses, we will now discuss which one of them could or could not be suitable approaches for the GISEle model's purposes:

- Hierarchical Clustering Techniques for instance are the less suitable solutions, as no hierarchy relations seems to exist between data points representing population distribution. Moreover, the adoption of this type of algorithm would have led to huge computational time considering the target resolution level.

- Centroids models, like k-means, have two disadvantages that make them unsuitable for GISEle: firstly it is unrealistic to think that the user could be able to a priori define the required number of clusters and, secondly the result is the fractioning of the whole given area in Voronoi polygons with only convex shapes, not representative of areas where the population is so low that electrification would not be cost-effective.

- Model-based clustering algorithms, despite having many strengths, are really difficult to apply because it is difficult to effectively find a probability function able to fit the population distribution in the studied area. This is why this category is rarely used in human population clustering: some attempts have been made in developed countries [51], where urban plans exist, while in remote rural areas it is harder to find any underlying scheme in houses' spatial distribution. Nonetheless, we consider this option to be a possible future development: combining an image recognition machine learning algorithm able to identify the houses and an accurate model-based clustering, it would be possible, based on yearly subsequent satellite images, to forecast not only the population growth of a community but particularly to infer its future spatial evolution.

- Graph-based class was, considering our specific goal, the first choice in the development process. Being our final purpose to solve the electrical grid routing problem, the idea was creating one unique grid and in a second time deleting worthless paths. The result would have been spatially-separated graphs representing optimal off-grid medium voltage lines. The main problem is that clustering algorithms are much more efficient than routing algorithms: this is the main reason why clustering and routing have been considered as two separated steps, as the first one allows to reduce the computational time of the second one.

- Grid-based and Density-based clustering algorithms were then the best two options. They are quite similar but, considering the previously detailed data gathering and processing step, it is easy to see how the fundamental stage of Grid-based clustering was performed in advance on raw data: the data space is partitioned in the grid-like structure. Each cell's attributes are in this case collapsed into its central point for sake of simplicity.

Accordingly, *GISEle's* clustering algorithm is a density-based algorithm, and specifically a slightly modified version of *DBSCAN*, with a particular adaptation based on an important input data property: points do not represent a singular element like in general clustering problems but have a fundamental property which is the population. Thus, *DBSCAN* algorithm has been selected: a particular adaptation of this algorithm in which the point weighting criterion is precisely its population attribute was implemented. Therefore, the *minPts* input parameter changes its meaning from minimum number of points to be found in a neighborhood to define it as core neighborhood, to minimum number of people. Firstly, this allows the algorithm to create clusters having the exact communities' extension shape. Additionally, the ability of this algorithm of identifying scarcely populated areas allows to neglect vast zones with few people, prioritizing highly populated ones.

Moreover, the abovementioned main weaknesses of this algorithm (not being totally deterministic, the hardship in defining concept of distance in high-dimensional data and the implicit assumption of clusters having similar densities), does not affect the suitability of this algorithm for our purposes. In particular the not totally deterministic solution is not a problem because it does not affect the final electrification status of a point, as it only could cause its cluster membership assignment.
DBSCAN's input parameters, *eps* and *minPts*, need to be defined by the user. To facilitate this task, the model asks the user to define investigation ranges for both, and displays three tables showing the results of each couple of parameters' combinations in terms of:

- number of resulting clusters;
- % of clustered people over the target area total population;
- % of clustered area over the total target area;

In this way, the user can choose the preferred parameters' combination to be entered in *GISEle*, clusters are created, and finally the model asks the user to combine (if necessary) the output clusters, based on a graphic interface showing their distribution in the total space. An example of the output of the clustering process on a fully rural area is shown in figure 3.10. It is possible to see the great advantage given by the autonomous optimal cluster shape detection of the *DBSCAN* algorithm: cluster points are enclosed in an area surrounding the populated village and respect the administrative borders. More details in terms of clusters' characteristics will be given in chapter 4 in which we will apply it to a concrete case study.

3.4 Electric grid routing

The discrete representation of the target region obtained by means of the nodes mesh originally created, has proved sufficient to perform the analysis of population distribution just explained, which led to the definition of people's clusters constituting the future single energy communities. To achieve the next goal an improvement becomes however necessary. In order to design the electric grid topology which will connect all the consumers of a cluster, the allowed interconnections between the nodes, together with their intrinsic cost, have to be defined. Once done, the graph, representing the basic framework where to run the mathematical algorithm depicted in section 1.4 will be completed.

A regular graph is built allowing each node to be connected only with its eight neighboring nodes (figure 3.11).

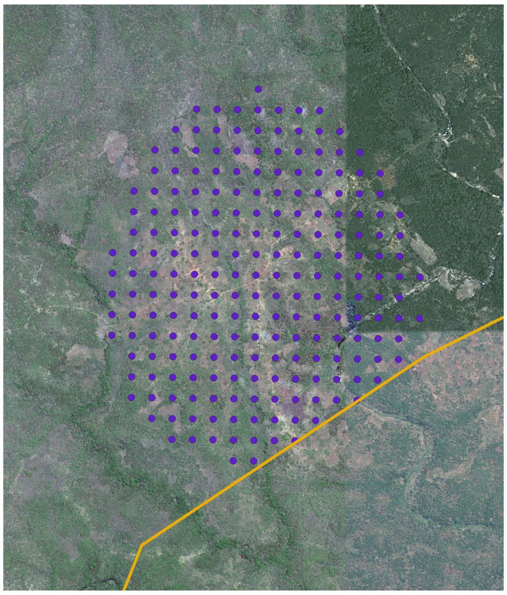


Figure 3.10: Population-weighted DBSCAN algorithm's output cluster

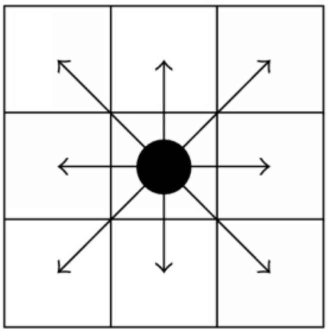


Figure 3.11: Connections of one cells with its eight neighbors

And the cost of the single graph's edge connecting two nodes i and j is defined as:

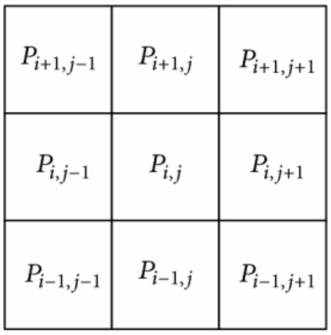
$$C_{ij} = L_{ij} \cdot \frac{UC}{1000} \cdot \frac{p_i + p_j}{2} \tag{3.8}$$

Where:

- C_{ij} is the cost of the connection ij in US dollars.
- L_{ij} is the distance between two terminals i and j.
- UC is the unitary base cost of an electric MV line expressed in $\left[\frac{\$}{m}\right]$
- p_i and p_j the penalty factors associated to the two terminals

The penalty factor of the connection is nothing but the arithmetic mean of the penalty factors associated to the two terminals.

Due to the nature of node mesh generation process in QGIS, nodes are not defined as members of a matrix with values of row and column n_{ij}, therefore it is not possible to identify neighbors by means of reciprocal position as shown in picture 3.12 A different approach has been adopted, based on distances.



$P_{i+1,j-1}$	$P_{i+1,j}$	$P_{i+1,j+1}$
$P_{i,j-1}$	$P_{i,j}$	$P_{i,j+1}$
$P_{i-1,j-1}$	$P_{i-1,j}$	$P_{i-1,j+1}$

Figure 3.12: Neighbors relative position identified by means of rows and columns

The distance between two consecutive nodes is equal to the resolution adopted, and is also equal to the minimum distance registered between each couple of nodes ij with $i \neq j$; considering this known value, the *"neighboring nodes"* can be defined as:

Def 3.4.1 *Given a set of equispaced nodes N, a node $i \in N$, d_{ij} the distance between each couple of nodes ij and "res" the minimum distance measured. Then, a node $k \in N$, with $k \neq i$, is a neighbor if:*

$$d_{ik} \leq \sqrt{2} \cdot res \tag{3.9}$$

The factor $\sqrt{2}$, allows to include within the set of neighbors also those nodes which borders in diagonal direction.

Therefore, a distance matrix including the whole set of nodes is firstly build, and only those connections which satisfy the previous statement are considered. Finally, the *Edge cost matrix* is created, in which:

- The cost of selected edges is calculated through the equation 3.8.
- The cost of all the remaining connections is set to an almost infinite value equal to +999999999

Now the graph is ready to be employed in Minimum Path analysis.

Considering a single cluster, every populated point within it (with a value of population greater than 0) becomes part of the set of *terminal nodes* of a Steiner tree problem. Its solution will lead to the definition of the electric line topology with

minimum cost, connecting all the populated point of the cluster in a single energy system.
The *Networkx* package for python provides a function for the approximate solution of a Steiner tree problem[52] which asks as input parameters:

- A weighted graph G(V, E, w)
- The set of terminal nodes L

And returns the graph composed by only the edges which constitute the final electric grid able to connect all the cluster's populated point in a unique energy community. To reach the solution, the *steiner-tree* function adopts the MST approach, and the theory behind it, depicted in chapter 1.4, stated that the time complexity of this algorithm is

$$O(V * L^2) \tag{3.10}$$

dominated by the construction of the metric closure. It is therefore function of both the total number of nodes composing the graph (V) and the number of terminal nodes (L). The application of this function to a graph covering the whole target region would require huge amount of time and, especially, of RAM which the majority of common PCs are not equipped with; so, to reduce the size of the problem, only a limited portion of the graph enclosing the cluster in object is considered. The dimensions of such "box" are proportional to cluster width and are defined as follows:

1. The width of the cluster is defined analysis the coordinates of all the cluster's points.
$$Cluster\ width = \sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2} \tag{3.11}$$
2. A parameter *"Ext"* is defined as the distance between the cluster and the borders of the box which surrounds it. Its value changes in relation to cluster's width.
$$Ext = \frac{Cluster\ width}{5} \tag{3.12}$$
This value has been empirically defined in order to avoid *"memory errors"*.
3. The box is built with dimensions defined as:

$$\begin{aligned} X_{max}\ box &= x_{max} + Ext \\ Y_{max}\ box &= y_{max} + Ext \\ X_{min}\ box &= x_{min} - Ext \\ Y_{min}\ box &= y_{min} - Ext \end{aligned}$$

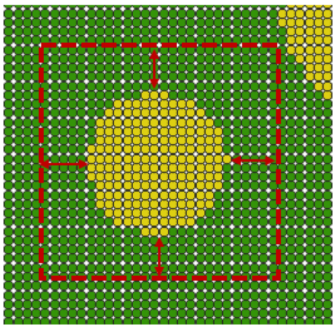


Figure 3.13: Example of box built around a cluster

Cluster points in yellow, graph's nodes in green. Red dashed lines define box's borders.

Through this expedient, only the nodes enclosed within the box are used to constitute the graph of the Steiner Problem, so reducing the V term of time complexity equation. Figure 3.13 shows a visual example of the result. Nevertheless, the main term of 3.10 is represented by terminal nodes L which, once the cluster is defined, cannot be reduced; for big clusters, despite the reduction of V, the computational time results still too high to make the Steiner Tree Problem manageable. A new, modified and lighter, approach is therefore required.

New approach for electric grid routing

The authors developed a new approach for electric grid routing which combines the potentialities of MST and Dijkstra algorithms to produce an approximate solution to the Steiner problem presented above.
The strength of the approach here presented, consists in the ability to split the original problem in multiple subproblems which can be faced separately. Instead of dealing with all the terminal points of a cluster in a single operation, each subproblem will involve only two terminals which needs to be connected, so limiting the amount of data that needs to be simultaneously managed by the RAM. Furthermore, in this way the time required to reach the solution of a single subproblem becomes acceptable: $O(m + n \log n)$.
Here, the developed strategy is depicted in detail.

MST At first, an initial, rough model of the electric connections is traced by applying the MST to the overall set of cluster's populated point. This is not an onerous operation: the *"minimum spanning tree"* function[53], provided by the *"scipy"* python library, returns the minimum spanning tree of the undirected graph, computed using the Kruskal algorithm, so taking a time equal to $O(m \cdot \log n)$.
Note: It is not possible to define meaningful values of cost for MST connections, since terminals can be very far aparth from each other; therefore we let it work on

distances. The weight of each connection w_{ij} is defined as the 3D distance between points i and j, which takes also into account the difference in altitude.

Lines classification An analysis of the MST solution is then performed, making a distinction between *Short lines* and *Long lines.*

- *Short lines* are those segments of the MST which connect two neighboring points. This happens when 2 neighboring points are both populated as figure 3.14 shows. Those connections are considered "valid" because their cost can be defined through the equation 3.8: therefore, together with their embedded cost values, they become part of the final solution tree and the nodes connected by these lines starts filling the list of points connected by the grid, named *"Lights-on"*.

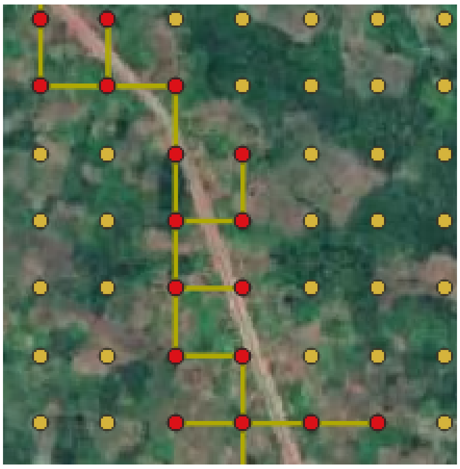


Figure 3.14: Connections between neighboring populated points

Def 3.4.2 ***Lights-on List*** *is the list containing the nodes, terminal and non-terminal, which have been reached by the electric grid.*

Def 3.4.3 ***Short-lines*** *is the dataset composed by all the segments connecting two neighboring points (therefore with length not greater than* $\sqrt{2} \cdot res$*), which constitute the final solution tree.*

- As *Long lines* are classified all the other segments of the MST connecting non-neighboring points (so with length greater than $\sqrt{2} \cdot res$) like that shown in figure 3.15. These, cannot be part of the final solution because, in this form, their cost cannot be directly determined. A further processing is required.

Long Lines Management The aim for the long lines is to define the *Minimum Cost Path* in the weighted, undirected graph which models the land area of interest,

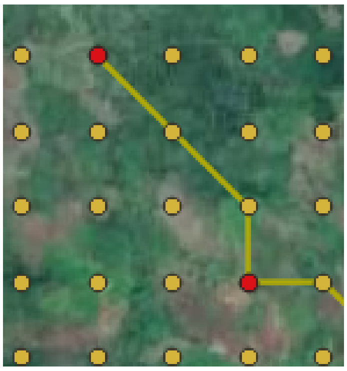


Figure 3.15: Example of connection between two non-neighboring points

able to connect the two terminals of the line. This is precisely the goal why Dijkstra algorithm has been developed.
The long lines are therefore sorted in increasing order of length and then processed one by one, from the shortest to the longest in the following way.

1. *GRAPH BORDERS DEFINITION*
   The first task, as done for the Steiner solution, consists in reducing the size of the problem by limiting the area under analysis. At this purpose a box is built around the area enclosed between the *source* and *target* vertexes.

$$\begin{aligned} x_{max} &= max(x_{source}; x_{target}) \\ x_{min} &= min(x_{source}; x_{target}) \\ y_{max} &= max(y_{source}; y_{target}) \\ y_{min} &= min(y_{source}; y_{target}) \end{aligned}$$

   Also in this case, as for clusters, the value of the extension *"Ext"* is function of cluster's width, now it is function of the distance between source and target d_{st} (equal to the line length).

$$\begin{array}{lll} d_{st} < 1000m & \rightarrow & Ext = d_{st} \\ 1000m \;< d_{st} \leq 2000m & \rightarrow & Ext = \frac{d_{st}}{1.5} \\ d_{st} > 2000m & \rightarrow & Ext = \frac{d_{st}}{3} \end{array}$$

   Also in this case the different factors have been set to avoid *"memory errors"*, and the box's extension is defined with:

$$\begin{array}{lll} X_{max}\ box & = & x_{max} + Ext \\ Y_{max}\ box & = & y_{max} + Ext \\ X_{min}\ box & = & x_{min} - Ext \\ Y_{min}\ box & = & y_{min} - Ext \end{array}$$

Only the nodes enclosed within this box will constitute the graph.

2. *WEIGHTING OF THE EDGES*
   Defined the borders of the graph, the edges connecting neighboring nodes are defined and weighted with the methods previously described, and the graph results therefore completely defined; however, some correction in the edges' cost must be made.
   Before launching the Dijkstra algorithm to find the Least Cost Path connecting *source* and *target*, it returns useful to check if any of the graph's connections has been already realized. If any edge is part of the *Short Lines dataset*, indeed, it means that it has been already included in the solution, therefore, its specific cost is set equal to 0.

   Theorem 3.4.1 *Given a weighted, undirected graph G=(V,E,w), a set S of short line connections already part of the solution, and being e_{ij} an edge $\in$ G.*

   $$If\ e_{ij}\ also\ \subset S \rightarrow w_{ij} = 0 \tag{3.13}$$

   This operation leads to two consequences:

   - It avoids that the cost to realize a section of line is considered more than once;
   - The Dijkstra Path connecting source an and target will be encouraged to exploit already realized electric lines;

3. *LEAST COST PATH*
   The Networkx's function *dijkstra path*[54] requires as input:

   - A weighted, undirected graph G=(V,E,w)
   - Source node
   - Target node

   And returns the sequence of nodes constituting the shortest path connecting source and target, which is possible to translate in the sequence of short lines, each one connecting neighboring nodes, composing the path. Each section p_{ij} of the founded path is taken into consideration:

   - If $p_{ij} \subset S \rightarrow$ It is already part of the solution, therefore it is discarded;
   - If $p_{ij} \not\subset S \rightarrow$ It is a new section of the final tree. p_{ij} is added to *Short lines* dataset, and the nodes i and j to the *Lights-on* list.

This process is repeated for every *long lines*: at the end every terminal node of the cluster will result linked to the electric grid, which can be displayed as a whole in QGIS environment.
Considerations:

- Being the Long Lines set ordered in ascending way of line length, iteration after iteration, the extension of the graphs increase and the analysis involves bigger and bigger number of nodes: therefore, the algorithm progressively tends to slow down.

Once the process reaches the end, the Short lines set S contains all the segments which constitute the final electric grid of the cluster.
In some individual cases, the algorithm could lead to the formation of internal cycle in the final grid. Normally, they do not represent a technical problems to the operation of the network, in fact they would increase the reliability of the electric furniture in the event of failures. But since the objective of the algorithm is to *"find the least cost network connecting all the populated nodes"*, the presence of cyclical structures is index of imperfect solution (figure: 3.16).

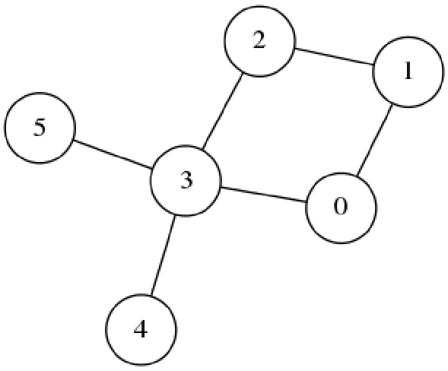


Figure 3.16: Graph structure containing a cycle

To face this problem, a python routine able to detect cycle structure has been implemented. It identifies the imperfection and deletes the most costly connection which is part of the cycle.
Whether the final grid is calculated with the Steiner function or through the new, modified approach, the resulting graphs are analyzed in the same way and the total length and cost of the grid are so calculated:

$$Cluster\ Grid\ Length = \sum_f length_f \tag{3.14}$$

$$Cluster\ Grid\ Cost = \sum_f w_f \tag{3.15}$$

Where f are the edges included in the final least cost grid, and w_f the cost associated to each of them. Figure 3.17 shows an example of resulting grid.

Having been developed in the course of this work, the performance of the new approach to the Steiner problem, both in terms of computational time and accuracy of solution, have not been undergone in-depth study yet. For these reasons, in the case of small cluster for whose analysis the Networkx's *steiner function* is applicable, the authors let the software free to calculate an approximate solution with both the methodology presented, and to finally choose the cheapest one.

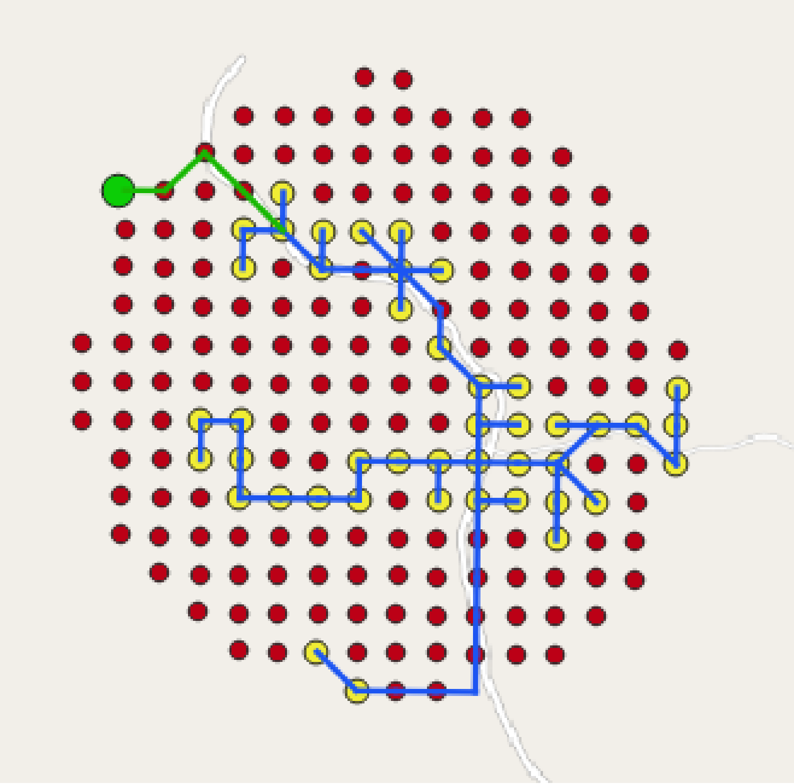


Figure 3.17: Example of electric grid (blue line) connecting the populated points (yellow) within a cluster (red points). The green line represents the further connection to the closest HV substation (green point).

HV Transmission Line Connection

The Dijkstra algorithm, adopted for the routing of long lines, is also useful to establish the path of the electric line which could connect the cluster's internal grid to the nearest HV transmission line substation. The two terminals, start and arrival of the connection, are defined by means of a distance matrix. The rows represent all the cluster's points reached by the internal network: both terminals and non terminals; the columns, instead, the set of the electric substations, existing or expected, which lie within the administrative boundaries of the target region.
The matrix's cell containing the minimum value, therefore, provides information about the cluster's electrified point which is the closest to the nearest HV substation. In other words, with a single analysis it is possible to identify both:

- The HV transmission line substation which is closest to the cluster in object;
- The cluster's point, connected to the internal grid, which is as close as possible to the defined substation;

These will be the two terminals of the desired connection. From here on out, the procedure to define the least cost path is the same adopted to manage long lines which exploits the Dijkstra algorithm.
Once the box defining the graph extension has been set, together with the edges' cost, the cost of those edges which are also part of the cluster electric network already built are set automatically to 0. This concept could be better understood looking at figure 3.17: all the blue segments (elements of the cluster electric network), have already been realized; therefore, in the graph studied to determine the least cost path to the substation, the cost of such specific segments will be null.
Often, in rural areas of developing countries, villages and communities can be located in very isolated areas, far from the closest HV substation as shown in figure 3.18. These situations, lead to deal with very wide graphs, which requires significant amount of time to be analyzed.

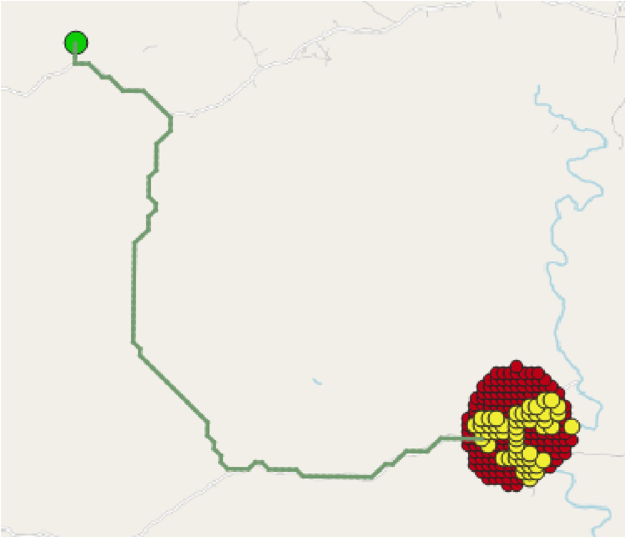


Figure 3.18: Example of isolated cluster, far from the closest HV substation

3.5 Load curves estimation in GISEle

Being *GISEle* designed to be as much as possible autonomous in its exploratory purpose, it is evident how the definition of loads cannot be done before running the model: considering that it is impossible to know clusters' number, geographic positions and extents, in advance. Accordingly, loop inferences must be performed inside the model itself, in order to determine each cluster's load profile based on valuable proxies.

The energy needs assessment of clusters is based on the definition of reference profiles: in the initialization phase, the user will be required to insert previously computed reference load profiles. This approach assumes that geographically close communities share socio-economic analogies in terms of basic energy needs, so in first approximation, their load curves will approximately have similar shapes.
GISEle takes the reference load profile as input, so the user is free to use the preferred approach in computing it. In addition to the profile, the user will have to specify a proxy, the most suitable being the number of households in the community to be used as reference. The reason behind this choice relies on assumption that, in developing countries' rural areas, communities are basically household-centric, meaning that their core is made of rural households. The other classes are a direct consequence of the house numerosity, in particular:

- Public services have generally fixed number in same-region communities, but differ in size and capacity based on the served population, which can be linked to the number of households assuming an average number of people per dwelling. Two close communities with very different populations will both have probably only one health center and one school, but their dimensions, and consequently energy needs, will have different capacities that can be assumed proportional respectively to the number of person and kids in their corresponding community;

- Basic productive uses (like for instance milling machines) generally follow similar rules, being their size dependent on community's population needs, like milling machines.Two important exceptions are represented by commercial activities and higher-level entrepreneurial activities. The first category is strongly affected by the location of the community, because strategically located communities, close for example to busy highways, generally have much more commercial activities compared to same sized communities far from roads. Entrepreneurial activities, instead, depend not only on the geographic strategic value of the location, but mainly on the unique peculiarities of the community and surroundings in terms of resource, opportunities, and on the sociological characteristics. There are in facts fundamental preconditions to the growth of entrepreneurial activities like literacy race, social structure, presence of micro-credit schemes and other region-specific cultural features that could facilitate the birth of a collective mindset among the community members.

Three considerations have to be done about the GISEle's adopted approach:

1. Firstly, being our main purpose addressing energy access, ideally for all, the considered time-frame is restricted to a medium-term horizon. Consequently, the reason no long-term forecasting is performed in it, is that the idea is to work with a fix one-year-long window representative of the energy load few

years after electrification. This gives the possibility of considering all basic public services, households' appliances which will be bought immediately and few years after the electricity arrival and lastly the existing productive uses plus some "early-adopter" entrepreneurial activities. Any further upscale of the systems is not now among the purposes of *GISEle*;

2. The resilience of the load generation process strongly depends on the input data the model receives: if the user has the possibility of inserting several input references loads, the smallest their geographic resolution, more likely each cluster will be assigned with a plausible load profile.

3. Load forecasting dynamics have a so high degree of complexity that they would require an entire thesis to be dedicated them. Therefore, being true we consider our assumptions sufficiently resilient, for the purpose of this work these issues will not be directly attacked. That said, there are very good chances that further development of this model will try to deal with them.

3.6 Sizing

Resources

After having defined the clusters energy needs the successive step is sizing an optimal power system capable of supplying the energy when needed. Being the final comparison of this proposed methodology between off and on-grid configuration, it is paramount to size an optimal system based on available resources. We have identified solar, wind and hydro as most cost-effective renewable resources for off-grid rural electrification. Regarding the first two ones, to date GISEle is able to extract the location of each cluster and automatically download the corresponding solar and wind resources from RenewablesNinja[55], an opensource web tool developed by Imperial College London and ETH Zurich, in order to be used in the subsequent veritable sizing part. This database has been chosen as our main source thanks to the fact that it gives open API access allowing GISEle to automatically contact the renewables.ninja API and download the necessary data.
Concerning the hydroelectric potential, the most promising site in the nearby of the cluster is selected from the hydropower potential layer obtained through the hydrological analysis depicted in 3.1. Only one, because the sofware *Homer Energy Pro* exploited for a first approach to the problem, allows to consider a single set of data per each energy resource typology. This characteristic, do not represent a big handicap in regard to solar irradiation and wind speed, since their values can be considered constant on the whole cluster area without incurring in big errors. The impact is instead greater for the hydro resources because, if a cluster energy system has easy access to multiple promising sites, only one can be chosen and the

others discarded. When Calliope will be completely integrated inside *GISEle*'s code, multiple hydro resources may be taken into consideration.

Methodology

As long as sizing is not the main goal of this project and being it a problem that has already been widely addressed in the literature, we have decided to let this part be somehow outsourced. In the previous chapter 2 two tools were described at this scope, highlighting their advantages and disadvantages: *Homer Energy Pro* and *Calliope*.

Homer Energy Pro has the advantage to be a highly consolidated tool designed specifically for solving optimization problems based on resources and loads.
Unfortunately, it has the disadvantage of having a proprietary license, which from one side opposed the open-source purposes of our project and, from the other side, being impossible to have access to its source code, it is impossible to effectively insert it in a fully autonomous model like *GISEle*: this means that the code would have to pause at every iteration, to give to the user info to be inserted in HOMER Energy Pro which will optimize the power system, and finally the user will have to give back the output to *GISEle* which would then continue with its loops until all clusters' power systems have been sized.
It is clear how this process would be unfeasible when running on vast areas like an entire country in which it is quite realistic to think clusters would be several hundreds.

Calliope Through *Calliope*, instead, being its code accessible and written in Python language, we were able to customize it in order to make it complementary to our purposes. This tool's main disadvantage, which is also its greater value in a GISEle's long-term perspective, is linked to the fact that its main purpose is not sizing autonomous systems but optimizing medium and large-scale interconnected systems. For this reason, the definition of technologies is trickier than in HOMER Energy Pro: for instance, regarding components like batteries which are fundamental in an off-grid system from both technical (their cycles' modelling plays a paramount in renewable energy generators' sizes) and economical (their cycles' modelling is paramount in evaluating its actual lifetime and thus forecasting future replacements and linked future investments cash flows).
Firstly, the possibility of inserting its sizing algorithm transversely to *GISEle* allows to automatize the process regardless the number of clusters. Moreover, its advantages, ranging from its georeferenced approach to the problem with the possibility of defining nodes and edges, to its multidimensionality allowing to model systems

from pico-level to multinational grids allow to think about future deep integration between the two models. If applied to distribution field, its framework composed by nodes and edges would reflect respectively households, power generation components and low voltage lines; at an higher, national level nodes could be cities and power-plants, whereas edges the high voltage lines.
In light of the above, in the development of this thesis project both approaches were used mainly in order to evaluate the differences in their respective outputs, and hopefully, to validate Calliope with respect to a more resilient software like HOMER Energy Pro guaranteeing in this case the legitimacy of using it in further *GISEle*'s releases.
Whatever approach is considered, the result of the analysis will be a set of different possible energy solutions able to fulfil the energy demand of the cluster grid for the future 25 years. Each setup is presented with embedded informations about:

- Number and characteristics of each item which will be part of the energy system;
- LCOE;
- Initial capital expenditure;
- Overview of operating costs over the entire operating horizon;
- Fraction of renewable energy penetration;
- Amount of pollutant emissions;
- Amount of fuel required;

Together with a multitude of other minor information detailing both technical and economical characteristics of the solution proposed. The proposed configurations are usually sorted in an increasing LCOE order: indeed, the primary objective of this specific step in electrification strategy planning is: stated the available technologies, the energy resources and the amount of energy required along the year, to find the optimal combination of energy technologies able to fulfill the demand pursuing the minimization of the energy cost. Among all, the LCOE is the main parameter considered: the one on which, in the next final step, the decision about the optimal electrification strategy to follow will be based. Despite costs of renewable energy technologies are constantly decreasing, renewable energy sources are often still more expensive than fuel. In the cheapest energy solutions, indeed, Genset always covers a primary role in energy generation because it avoids the costs of expensive batteries storage systems. In each cluster, *GISEle* selects three different setups which will be subjected to the final evaluation:

- Solution 1: the cheapest one;

- Solution 2: the one still exploiting fossil fuels , but with the highest fraction of renewable penetration.
- Solution 3: the cheapest solution providing 100% renewable energy.

3.7 Costs evaluation and decision making

The final output delivered by *GISEle* is a cluster specific comparison between the two main possible electrification strategies:

1. Isolated Micro-grid;
2. Grid-connected energy system.

Isolated micro-grids are not connected to the HV national grid, therefore they do not have to sustain the costs related to the siting of the power line linking the cluster's internal grid to the nearest HV substation. Nevertheless, they deal with energy costs which are usually higher than that proposed by the national service. As underlined in the beginning of this lecture, the LCOE returned by the sizing algorithms, and expressed with equation 1, considers only costs linked to generation devices feeding the micro-grid. In order to improve the quality of solution, *GISEle* includes in the formula also the capital expenditure for electric lines routing. The new final expression of the energy cost becomes:

$$LCOE_{micro-grid} = \frac{C_{grid}}{total\ energy} + LCOE_{gen} \tag{3.16}$$

C_{grid} : Capital cost for cluster internal electric network

total energy : Forecast of the total amount of energy produced and sold by the energy system in 25 years of operation

$LCOE_{gen}$: Levelized Cost of Electricity returned by sizing process

The alternative to the micro-grid structure consists in connecting the cluster network directly to the national HV transmission line. This solution give access to electric energy at lower cost (since produce by big power plants which can benefit from economies of scale), but requires to realize the further electric line linking the cluster to the closest substation. The energy cost related to this option is therefore:

$$LCOE_{connected-grid} = \frac{(C_{grid} + C_{con})}{total\ energy} + COE_{NG} \tag{3.17}$$

With respect to equation 3.16, the new terms included in equation 3.17 are:

C_{con}	:	Capital cost of electric connection between cluster and HV line
COE_{NG}	:	Cost of Energy provided by TSO.

Finally, the two alternatives are compared and the optimal strategy is defined.

Category	**Type**	**Penalty factor**
Land-cover	Tree Cover, broadleaved, evergreen (>15% tree cover)	1
	Tree Cover, broadleaved, deciduous, closed (>50% Tree cover)	5
	Tree cover, broadleaved, deciduous, open (15-40% tree cover)	1
	Tree Cover, needle-leaved, evergreen	3
	Tree Cover, needle-leaved, deciduous	3
	Tree Cover, mixed leaf type	3
	Tree Cover, regularly flooded, fresh water	8
	Tree Cover, regularly flooded, saline water	8
	Mosaic: Tree cover/Other vegetation	2
	Tree Cover, burnt	2
	Shrub Cover, closed-open, evergreen	1
	Shrub Cover, closed-open, deciduous	1
	Herbaceous Cover, closed-open	1
	Sparse Herbaceous or sparse Shrub Cover	1
	Regularly flooded Shrub and/or Herbaceous Cover	6
	Cultivated and managed area	1
	Mosaic: Cropland/Tree Cover/Other natural vegetation	1
	Mosaic: Cropland/ Shrub or Grass Cover	1
	Bare Areas	1
	Water Bodies	9
	Snow and Ice	7
	Artificial surfaces and associated areas	1
Water Bodies	Yes	10
	No	0
Protected areas	Yes	9999999
	No	0
River	Yes	9
	No	0

Table 3.3: Penalty factors

Chapter 4

Case study

4.1 Mozambique: context

Country overview

Mozambique has an area of 801.590 km^2 and a population of 28.861.863 inhabitants, of which 15.061.006 women (see National Census of 2017) with an average population density of 36.1 inhabitants per km^2. The 65 % of the population lives in rural areas, with a specific growth rate of 2.5% per year. The country is well endowed with natural resources: arable land, forests, fish resources, water, mineral resources and solar energy. Its economy has recorded an average annual increase of around 7.5% from 2005 to 2015, without however managing to face deep development challenges, since the rapid macroeconomic growth was not matched by a significant reduction in the poverty of its population: social inequality is very pronounced (Gini coefficient equal to 45.7, UNDP, 2013) and there is a substantial absence of an adaptation plan of the consumption of natural resources, in terms of efficiency and conservation. To date, Mozambique is still one of the poorest countries in the world, with about 54.7% of the population living below the poverty line, as shown by the low value of the country's Human Development Index (0.418 in 2016) which put it in the 181st place out of 188 nations. The Mozambican population is characterized by an extreme and systematic fragility: the majority lives in a condition of high vulnerability and chronic crisis caused by systemic factors such as poor primary sector productivity, strong population growth (+ 43% in the last decade, + 80% in the last twenty years), the persistence of strong socio-economic tensions (which are exacerbated with increasing inequalities), the depletion of natural resources and the absence of other resources to ensure inclusive and equitable sustainable economic growth.

Energy sector today

As of 2016, national access to electricity was 24.2% (World Bank), and the average time required to get a connection to the electricity grid went from 91 days in 2016 to

68 in 2017 (WB). The Mozambican energy landscape appears intricate and unclear. Until today the regulator of the energy sector has been the MIREME (Ministério dos Recursos Minerais e Energia), a role that, however, was supposed to be transferred entirely to the CNELEC (Conselho Nacional de Electricidade) in 2017, which would have become fully independent. Nonetheless, this transfer of power never effectively occurred and every crucial decision concerning energy in the country has to be submitted to the Energy Minister's cabinet. At the same time, the Government established the FUNAE (Fundo de Energia), an administratively and financially autonomous public institution with the role of supporting and developing the management of energy resources, being responsible for the off-grid electrification field, to be complementary to EDM (Energia de Moçambique), the National Grid operator. Currently, the southern network of the national territory is not connected to the central and northern parts. The current National Electric Grid accuses several problems in terms of reliability, as it suffers from numerous blackouts, which besides being a considerable cost for EDM, represent a brake on the economic development of the entire country. One of the reasons lies in the used tariff regime (flat-rate tariff) which is applied throughout the network, regardless of the socio-economic potential of each area. Although there is a strong will on the part of the Mozambican government to electrify the rural areas of its national territory, the prohibitive costs estimated for the extension of the REN (between 15 and 25 thousand euros per km), clearly show the importance of investing, especially in the short term, in off-grid systems, based on the abundant renewable energy resources present in the country. Investments still struggle to be implemented mainly due to a legislation that is unclear and at times adverse to the private sector involvement in the energy sector: Mozambique is the African state with the lowest investments in renewable energies, being 2.2 million dollars between 2009 and 2014. The major barriers to the success of rural electrification in general and off-grid systems are resumed in tab 4.1. They are the outcome of a survey carried out in 2011 by the World Renewable Energy Congress, questioning some of the major players on the national energy scene (including EDM and FUNAE). Although 8 years have passed since that survey, the

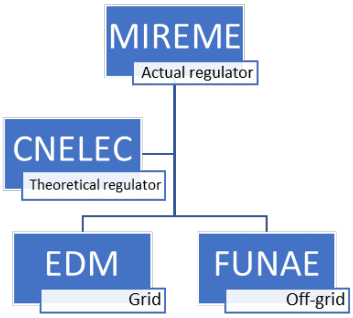


Figure 4.1: Structure of energy sector in Mozambique

	Rural electrification	*Rural electrification*	*EDM*	*Funae*
Institutions and stakeholder performance	*General*	*Off-grid*	Do they agree?	
Low quality institutions	x		Y	N
Inadequate planning skills	x (only donor)		N	N
Sector top-down management	x (only EDM)		Y	Y
Economics and Finance				
Poor rural market with few productive uses	x		Y	N
High diesel costs		x	Y	Y
Dependence on donor	x		Y	N
Social Dimensions				
Poverty and low purchasing power	x (only EDM)		Y	N
Poor system maintenance culture	x	x	N	N
Low capacity of PV solar systems		x (only FUNAE)	N	Y
Poor access to necessary components	x (only EDM)	x (only FUNAE)	Y	Y
Poor generation capacity	x		Y	N
Diffusion and adaptation of technology				
Lack of local entrepreneurship	x (only EDM)		Y	N
Rural infrastructure				
Dispersed population	x	x	Y	Y
Devastating cyclones	x		N	N

Table 4.1: Barriers to rural electrification

stagnation of the energy sector make that no substantial differences have been found during this project's assessment. Moreover, a key issue that hampers the country's electrification is the current legislation which, besides being unclear, does not allow private operators to sell energy within the country. This turns out to be a considerable barrier, as it is now the opinion of most that the involvement of the private sector is the real key to create fertile ground for large-scale implementation of micro and mini-grid systems that can meet the energy needs of rural communities far from being reached by the national electricity grid. Finally, specific incentives and policies for the off-grid sector, which are fundamental pillars for a real development of the sector, are almost non-existent.

Being investments in the sector the lowest in the continent, FUNAE has so far exercised the role of implementing agency, rather than supporting third-part's rural electrification initiatives. For instance, when internal funds allowed it, it built several off-grid micro-grids. In the first time, diesel generators were the preferred technology: due to their low investment costs they allowed to electrify a relatively high number of communities. The systems were strongly subsidized, energy was available for few-hour timeframes and consumers had to pay a fixed monthly fee far from being close to its cost-reflective value. In the long run, subsidies were removed and due to the unsustainability for consumers even in only purchasing the fuel, most

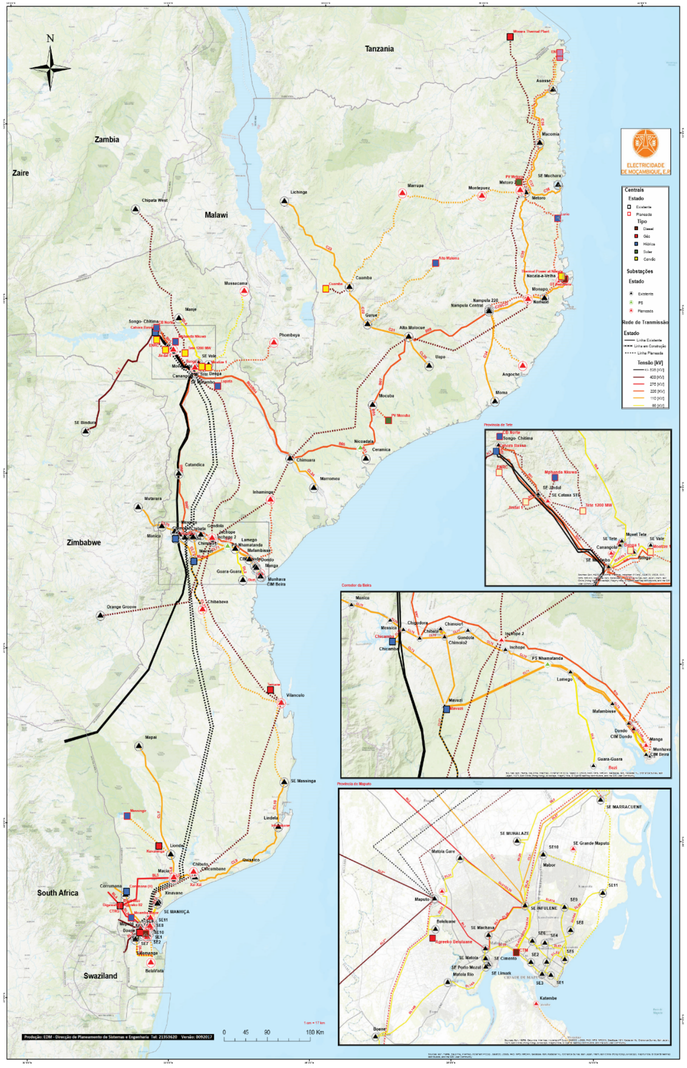


Figure 4.2: Actual energy infrastructure of Mozambique with lines and power plants

of these systems failed.

Consequently, due to critics from the communities, FUNAE abandoned diesel generators and switched to renewable energy and, in particular, to solar-PV technology. It started building some small micro-grids with a capacity generally around 5 to 10kWp feeding a few dozen homes and public functions. High investment costs obviously reduced the replicability of the systems, but low annual costs allowed these systems to be financially self-sustaining when considering the investment as non-repayable. The business model was a community management model in which community committees were created, with at least a president, treasurer and an on the spot trained technician chosen among community's young men. Tariffs had basically the same structure as the previous scheme. This model allowed some systems to have a longer lifetime but faced some obvious issues that cannot be effectively associated to any sort of sustainability. The investments had to be entirely a burden of the Mozambican State and this makes a replication towards universal energy access impossible, especially for a country with such limited financial resources. Additionally, this fact combined with homogeneous and non-cost-reflective tariffs, creates a chronic economically unsustainable situation: every unpredictable event (failures) or even predictable ones (ordinary replacements of exhausted components) are insurmountable walls limiting the lifetime of the systems. Figure 4.2 represents the current Mozambican electric infrastructure.

Energy sector tomorrow

FUNAE drafted the *"Atlas of Renewable Energies"* [56] and the *"National Portfolio of renewable water and solar energy projects "*, with the aim of creating the conditions for the advent of foreign investments. They are the outcome of a two-year-long country-wide mapping initiative, in which renewable energy resources and electrification priorities in terms of socio-economic strategic value were mapped, and consist of a long list of prefeasibility assessments of potential off-grid projects (micro-grids, mini-grids and standalone systems), with locations, suggested technologies, estimated available powers and strategic values.
Although the remarkability of this initiative, it is necessary to underline some criticalities: data are not directly accessible (we asked but their policy do not include any data disclosure option), no methodology is comprehensively explained (especially concerning resources assessment like hydroelectric power potential), and significant discrepancies exist between values in the printed version and in the online version, and also between different web-pages. The Atlas, and particularly the work behind it, is one of the reasons why *GISEle* project started, and although it is not completely clear how valuable it should be considered, it will be a of reference scenario for results' comparisons. The main reason is that it gives valuable information in terms of geographic distribution of potential projects, and it could be used as an important term of reference. Even with the existence of the Atlas, foreign investments did not have an appreciable increase and the cause can mainly be identified in the energy sector regulatory framework. As already stated, it is unclear and somehow hostile to private sector involvement. Moreover, the key players, which should work in synergy, act in completely separated ways, leading to frequent conflicts of interests and uncertainty even when the Ministry seemed to be opening the field to experimental innovative business models. An exhaustive example is the Titimane project. In 2015, MIREME, in collaboration with EDP (Energias de Portugal), and UNEP (United Nations Environment Program), chose the village of Titimane, as the site for a hybrid solar/biomass mini-grid, which was to be the first step in the creation of a fertile soil that would favor the implementation of similar projects with the involvement of the private sector. Unfortunately, a few months after the work began, EDM decided to bring the village of Titimane into its electrification plan for the following two years, thus causing the project to fail in the bud. This

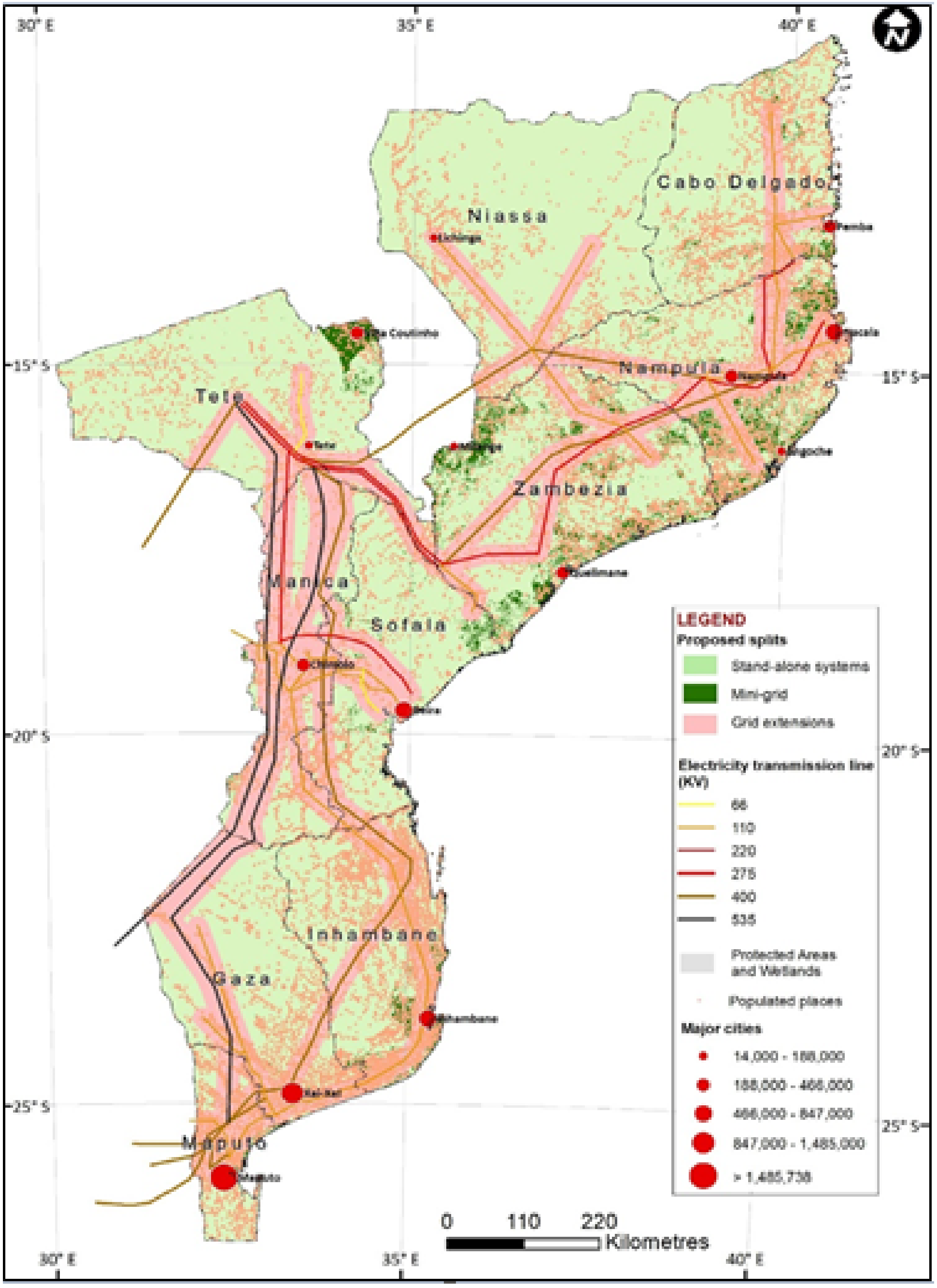


Figure 4.3: Graphic view of Atlas resume

episode highlights how the lack of communication between the various players in the sector makes Mozambique a very risky country for this kind of investments. To date the Mozambican Government, as a result of repeated interlocutions with international stakeholders decided to launch a vast program of reforms in the sector, resumed in the *"Nova Lei da Energia"* (New Electricity Law or NEL), a new law that is expected to disrupt the energy sector and hopefully to drive Mozambique towards an energy-driven sustainable development. This law was supposed to enter into force in 2019 but due to incumbency of the national elections, the ruling party decided to postpone its implementation to the following year. Its final content is actually not fully disclosed, but it should substantially privatize the energy sector, giving the possibility to private players to actively entering it, and in particular, concerning the off-grid sector, new business models should be allowed, opening to co-financing, long-term licensing and cost-reflective tariffs setting. The last one is especially crucial, as nowadays it is the main barrier in terms of sustainability of off-grid electrification strategies: the cost of energy is the same all over the country, regardless the region-linked socio-economic peculiarities and the reasons behind this choice are fundamentally socio-political. This tariff is the EDM's tariff, set several years ago when over the 90% of the energy produced in the country came from the 4GW *Cahora Bassa* hydroelectric power-plant, leading to a specific cost of energy very low due to huge scale effects. With the rise of energy consumption, more-costly power-plants were built and the true energy cost rose too, but the tariff slightly changed, creating a situation in which it is unsustainable even for EDM to operate. Obviously, a tariff which is unsustainable for EDM cannot be effective for basically isolated off-grid systems which are generally linked to true specific costs of energy more than 5 times the EDM's tariff. The responsibility over the regulatory framework and tariff approval will be transferred to a new entity, ARENE (Autoridade Reguladora de Energia), to which each project will have to submit specific tariff schemes for an official approval.

4.2 Area under study

Mozambique administrative division is structured in four fundamental level:

- Province;
- District;
- Posto Administrativo;
- Localidade.

The *Zambezia* Province concentrates many of the criticalities in terms of development and energy. It has an area of 103.478 km^2, it is the second most populated province with 5.164.732 inhabitants, corresponding to the 18.5% of the whole country population and is one of the most densely populated provinces of Mozambique (48.7 inhabitants/km^2). Its population growth rate is also particularly high considering a 35% increase over the last 10 years. The population pyramid shows that 64% of the population is under 24 years old.

Zambezia is the most relevant province both from political consensus and development goals perspectives since it hosts one fifth of the country population, the 93% of which is living in rural areas, characterized by a very low resilience in the face of climate change and external shocks and where 70.5% of the population lives below the poverty line. Therefore, most of the government strategic efforts are concentrated on it.

Within this, the *Mocuba* District has a strategic value, being a crossroad in the middle of the northern part of the country, with highways coming from:

- *Quelimane* (place of respectively both the only and the most important Zambezia's airport and port),
- *Milange*, (link with Malawi),
- *Nampula*, the most developed Province (excluding Maputo City).

Mocuba has three *Posto Administrativo*, *Mocuba City*, *Namanjavira* and *Mugeba*: the target area for this analysis will be the second one. The reasons that lead to this choice are several:

1. One of the authors of this thesis is nowadays actively working on an electrification project in this area (the Italian Cooperation for Development Agency (AICS)'s ILUMINA Program implemented by the Italian NGO COSV), so a collaboration has been set up with the implementing agency, and the data gathering process was somehow more reliable.
2. Being *GISEle* designed to deal with both *greenfield* and *brownfield* areas, in order to effectively guarantee the validity of the case study's results, all data had to be most possible accurate. Unfortunately, the dataset related to the existing distribution network path in the country is not available. For this reason, *Namanjavira* had the most suitable configuration because despite being completely unelectrified (contrarily to the other two *Posto*), the *EDM*'s planned transmission network has two lines crossing its borders.
3. Being the uncertainty in terms of computational burden of the final model relatively high, concentrating on one smaller area gave us the possibility of testing all algorithms' outputs, despite having a 200m spatial resolution, quite high compared to other tools that generally don't go below 1km.

4. The hydropower dataset, for which similar considerations from 2. and 3. can be made, was an ulterior limiting factor. In fact, as already explained, data were not available, and an entire parallel methodology had to be developed not to neglect one of the most abundant available resource in the intervention area. Unfortunately, being this methodology quite a burden, dealing with bigger areas would have been highly time-consuming despite not adding so much in terms of value to *GISEle* and in general to this thesis work.

	Zambezia		**Namanjavira**	
	Number of sites	Total power	Number of sites	Total power
Solar	111	0.2 GW	0	0 MW
Wind	700	0.750 GW	0	0 MW
Hydro	1191	10 GW	16	77.13 MW

Table 4.2: The FUNAE's Renewable Energy Atlas identifies in Zambezia and specifically in Namanjavira:

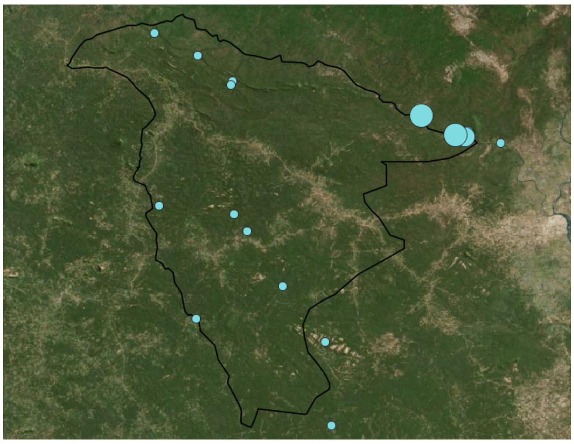


Figure 4.4: Renewable Energy Atlas' Projects in Namanjavira

4.3 Resources

The following resource assessment is the combination of our independent data collection process with the FUNAE's Renewable Energy Atlas. This dualism allows us to appreciate not only the primary energy availability, but also set an order of magnitude benchmark to the effective potential in terms of power generation. The data reported in FUNAE's Atlas will be subsequently recalled in order to appreciate eventually occurring convergences or divergences between our results and the main actual reference document (for all stakeholders) in electrification strategy planning in Mozambique.

Source	**Name**	**Priority**	**Power [MW]**
Hydro	Garanha	Yes	9.99
Hydro	Garanha2	Yes	14.6
Hydro	Namanjavira1	No	21.53
Hydro	Namanjavira2	No	18.69
Hydro	Namanjavira3	No	3.95
Hydro	Namanjavira4	No	1.87
Hydro	Namanjavira5	No	1.08
Hydro	Namanjavira6	No	0.36
Hydro	Namanjavira7	No	0.32
Hydro	Namanjavira8	No	0.32
Hydro	Namanjavira9	No	0.24
Hydro	Namanjavira10	No	0.14
Hydro	Namanjavira11	No	0.12
Hydro	Namanjavira12	No	0.07
Hydro	Nigau	Yes	3.06
Hydro	Paroma	Yes	0.79

Table 4.3

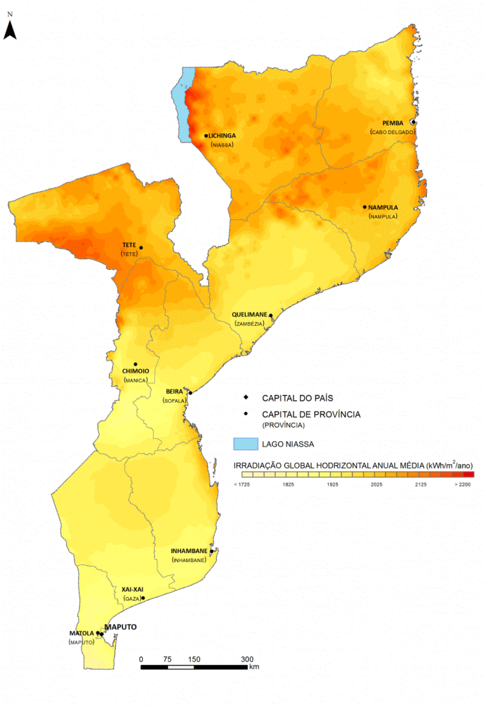
As it will be seen, the Atlas includes the whole hierarchy of powerplants' scale, in which the vast majority are big grid-connected renewable-based powerplants: as to the present *GISEle* is not taking into account on-grid powerplants, in the final results' comparison we will focus only on off-grid systems.

4.3.1 Solar

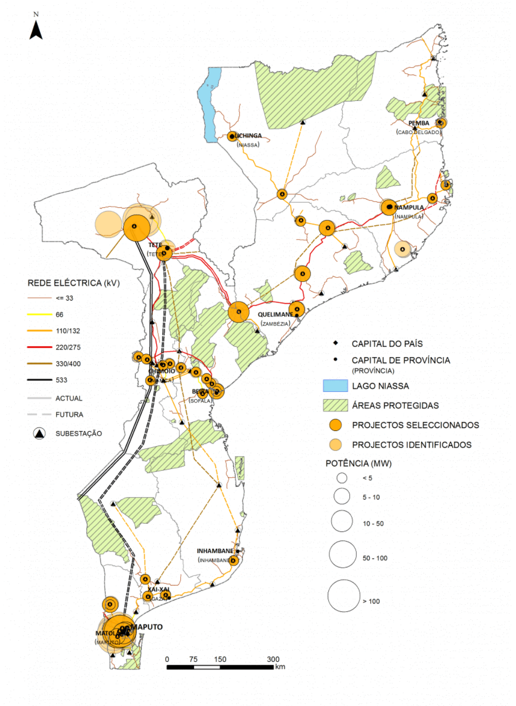
Mozambique has great solar power potential all over the country. In particular, northern districts have global horizontal irradiation average values above 2.000 $kWh/m^2/year$.

The *Renewable Energy Atlas* estimated a total 23TWp of national potential, 2.7 GW was estimated to be near existing substations, with significant potential remaining in off-grid areas. Such availability of resources led to an estimated average cost of energy for off-grid solutions at the 10,000 villages assessed of approximately USD 375 per MWh for hybrid solar and USD 600 per MWh for a 100% solar-battery system. Hybrid solar energy systems are therefore the most economical option for most settlements where hydropower is not available. Solar-battery systems seem to be the most economical 100% renewable option except in regions with high agricultural productivity for biomass.

The methodology adopted is supported by the combination of 11 *Instituto Nacional de Meteorologia* and 17 *World Radiation Data Center* (WRDC) meteorological stations' gathered data, between 1970 and 2000. In addition, 16 pyranometers were installed in 12 locations spread along the country having a two seconds sampling interval and a tenminutes average internal record for more than a year. Combining these data with satellite measurements of albedo, Linke turbidity, cloud index and topography the following global horizontal irradiation map has been produced:



(a) Annual mean horizontal irradiation in Mozambique



(b) Identified areas for solar energy based power plants

4.3.2 Hydro

Mozambique presents a mean annual rainfall of 940 mm concentrated in the months of December and March. The high concentration of rainfall in some months invariably results in floods and in droughts in the remaining months of the year. However, despite the irregular hydrological regime, mean annual runoff tends to be very high The country has a strong hydropower potential at all project scales, with existing experience in large hydropower projects. The *Renewable Energy Atlas* identified a further 1,446 potential sites (18.6 GW) of which 433 could power mini-grids, being within 10km of a mini-grid region, with a total potential of 2.1GW. From these 433 sites, 672 MW may be generated from priority projects, with the largest potential of 245 MW in the region of *Nampula*, followed by 171 MW in *Sofala*, and 165 MW in

Zambezia. Of the 1.44 GW of non-priority potential, 521 MW exists in Zambezia, and 455 MW in Nampula. Figures 4.7 below show the distribution of hydro sites within the provinces identified in the Atlas.

FUNAE claims this analysis relied on 265 already geo-referenced existing hydro schemes, 700 stream gauge and 1400 rainfall gauge stations, but unfortunately these data were not accessible, so there is no actual evidence of their validity. Moreover, although it is undeniable that Mozambique and especially the Zambezia province have high abundance of waterways, the main criticality in assessing the hydropotential the fact that during dry season most of the rivers have negligible flows. This seems to be confirmed by actual data we found on the Messalo river located in northeastern Mozambique.

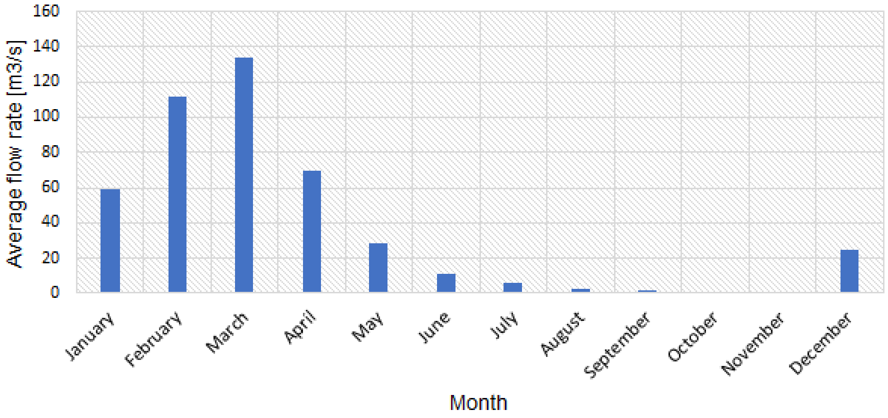


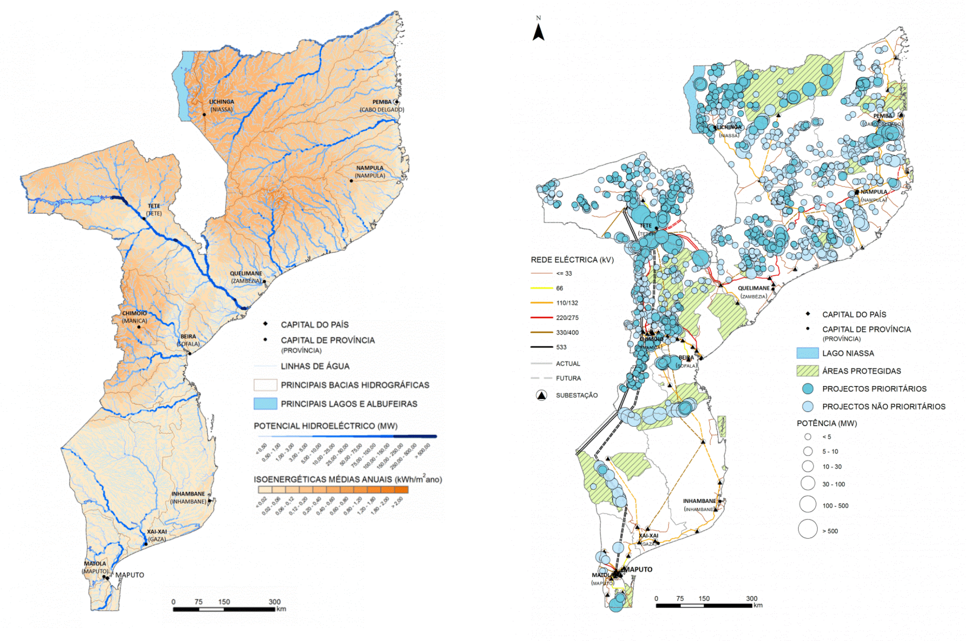
Figure 4.6: Monthly mean flow rate of Rio Messalo measured at Est. Nairoto Montepuz Station

4.3.3 Wind

The Mozambican wind sector remains underdeveloped, with only a single 300 kW turbine installed in Praia, Inhambane in 2012. As a result, there is a current lack of skilled technicians and local manufacturing of wind components. However, the resource mapping and sites visited as part of the Atlas should help stimulate the sector and there are already plans to develop a wind farm close to Inhambane.

Wind resources in Mozambique are significant, with an estimated 230MW of high potential sites identified in the Renewable Energy Atlas and average wind speeds of over 7m/s in northern and coastal regions such as Maputo and Gaza. The projects in these provinces have over 3,000 equivalent hours at rated power (NEPs). Projects were also identified in the provinces of Sofala, Cabo Delgado, Zambezia, Inhambane and Tete, with over 2,500 NEPs. Other areas have low to moderate intensity, with speeds between 4 and 6 m/s, and are therefore not cost competitive. Figure 11 shows the mean wind speed at 100m.

The methodology followed by FUNAE in this specific assessment is the mesoscale



(a) Mozambique river network

(b) Identified areas for hydro power plants

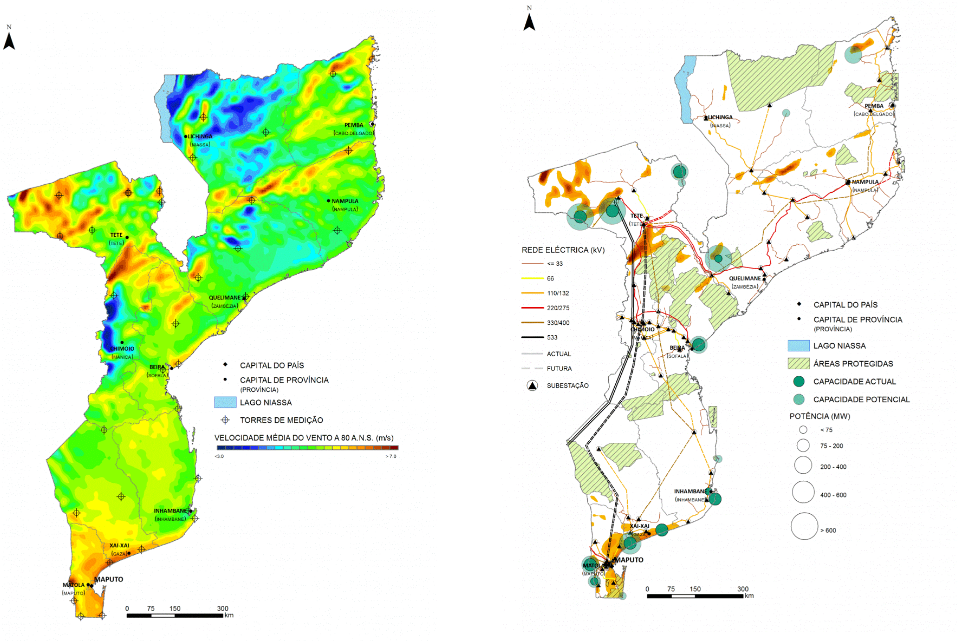
Figure 4.7: Hydro power potential presented by the Atlas

model MM5, based on global data from Reanalysis Project NCEP/NCAR. Based on this wind mapping and the major environmental, legal, technical and topography constraints, 35 locations were selected for wind measuring meteorological stations installation.

4.4 Loads

As already explained, *GISEle* asks the user to provide pre-computed load profiles in order to exploit them as references when assessing a plausible load for each cluster. As said the case study area has been selected as one of the students is an active part of a development cooperation electrification project located in it. Consequently, thanks to the favorable opinion of the implementing agency in sharing data about the beneficiary community, a real case region-specific load profile has been developed. The data gathering process has been conducted by the student itself with the help of many local people, through detailed surveys about energy needs. Moreover, a great advantage this community had is that it was previously electrified through a diesel-generator powered off-grid micro-grid, even if at the time only less than thirty households (of more than 700 in the core of the community) had actually access to the grid. This precondition simplified the data gathering process because community members were already aware about the possible uses of electricity:

- households which once had electricity still had their appliances and even in



(a) Mozambique wind map

(b) Identified suitable areas for wind power plants

Figure 4.8: Wind power potential presented by the Atlas

case of foreseen substitutions it was relatively easy for them to analyze their needs and general functioning windows and times

- people which never had electricity at home anyhow had a sufficient knowledge level to analyze which would have been their primary desires once energy eventually arrived in their community.

Combining these data about households and existing productive activities with necessities linked to public services, which are relatively easy to evaluate, allowed to define the whole community load profile. For this purpose, *LoadProGen* tool was adopted.
Below (tables 4.4 and A.11) are reported as example the starting data and key parameters used to generate load profiles of a health center and a standard household. 11 additional type of consumers where considered in the simulations and the details are reported in appendixA: primary school, secondary school, places of worship, administrative headquarters, police station, world vision office, public lights, merchants, tailors, barbers, premium households.
The time windows indicated below are the result of iterative optimization cycles. For reasons of synthesis we report here only the time windows that for technical-economic reasons have been chosen as definitive after having tried various options, based on feedbacks from the beneficiaries and the technical details of the plant that will be subsequently proposed.

Health centre

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	15	25	16-5	0	13	13	0
Ordinary Internal Lights	20	20	16-5	0	3	12	10
Extraordinary Internal Lights	10	25	0-24	0	2	8	50
Computer	3	100	0-24	0	2	14	30
Medical Electronic Equipment	1	300	0-24	0	2	7	50

Table 4.4: LoadProGen configuration parameters for health centers

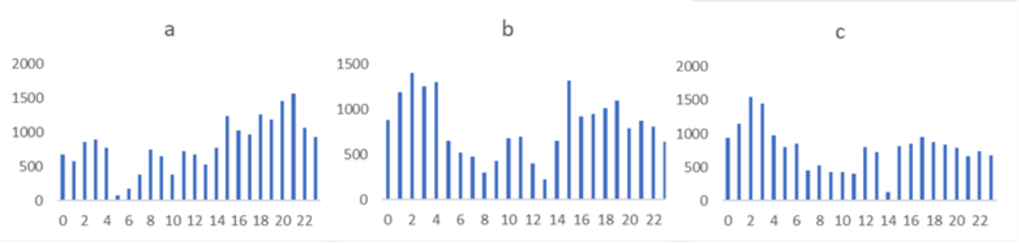


Figure 4.9: Three examples of health center daily load profiles

Given the peculiarities that it presents, predicting the progress of the health centre's load is definitely unfeasible. Three of the potential profiles are shown in figure 4.9, and as we can see they are completely antithetical, as the peak of the demand could be one day at night and the following day in the afternoon, or two peaks could occur over the same day.

In figure 4.10 it is possible to qualitatively appreciate the trends over the 24 hours of the three main consumer categories (standard household, premium household and merchants).

As explained in the *GISEle* methodology description, its load has been used as base for a proportional scaling over each cluster/community. The previously processed population layer has "person" as unit of measure: for this reason *GISEle* asks which is the average number of people in a household in the studied area is (in this case the average value for the *Zambezia* province was used, i.e. five) and divides the population of each cell by this value estimating how many house are realistically present in it.

Additionally, the *Zambezia* province is a good example of what was described before about rural community framework: public services are built by the government based on the dimension of the served area, where for example schools or health center may have an operating radius above 15 km.

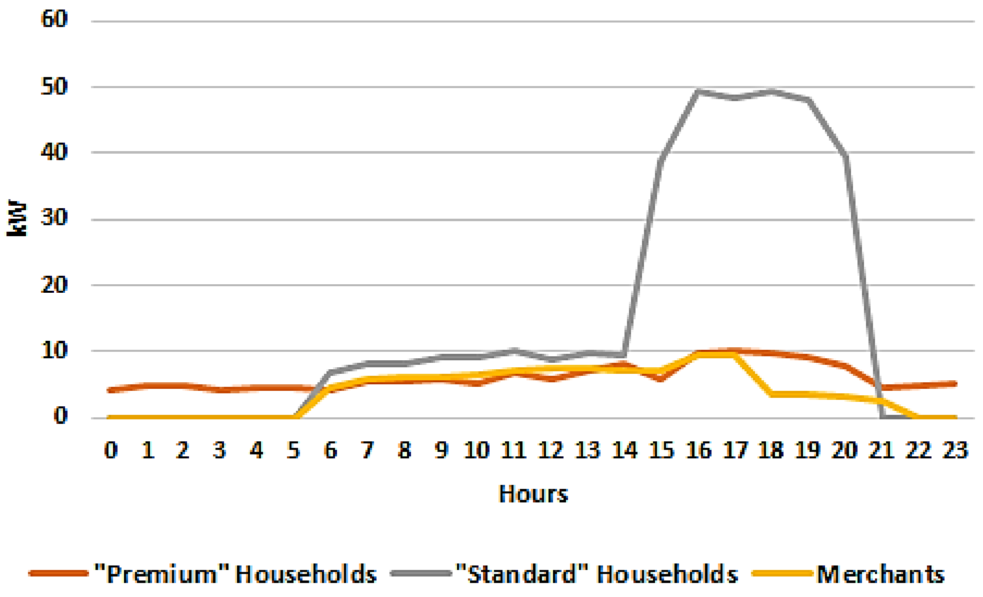


Figure 4.10: Main consumers load profile

4.5 Components specificities & Costs

Mozambique has no consolidated market linked to this sector, therefore quantifying reasonable costs for all components is a complicated task. Moreover, actual costs are highly influenced by the strong METICAL/DOLLAR exchange rate volatility and the logistics costs to deliver such valuable assets in lowly connected rural areas.

In order to stimulate the sector, in 2011 FUNAE created a manufacturing plant near *Maputo* in which solar cells are assembled. The other kind of assets are generally imported from outside the country, mainly China, India, South Africa and Germany.

A market analysis on import and logistics costs revealed how, due to this lack of maturity of the sector, their uncertainty was high, ranging from between 30 and 90 percent of the asset value.

Grid specific costs had high variability too, but differently from other assets, this discrepancy can also have been exacerbated by the fact that these data were collected by the implementing agency directly addressing private sector operators, whose seeing in it the possibility of a future subcontracting certainly tried to influence the budget predefinition process by overestimating actual costs. A reasonable cost of 20000 dollars per unit kilometers of MV line has been estimated after carried out some interviews with local manufacturers. Such value is also in line with the cost data indicated by ARERA (*Autoritá di Regolazione per Energia Reti e Ambiente*) for the realization of new electric connections to the main grid [57].

Below, instead, are listed the specifications of components considered in the modelling phase of this specific case study:

LIST OF ENERGY SYSTEM COMPONENTS			
Lead Acid Batteries DC	Capital cost	420	$/kWh
	Replacement cost	420	$/kWh
	O&M	10	$/year
	lifetime	10	years
	Minimum state of Charghe	40	%
	Roundtrip Efficiency	80	%
Li-Ion Batteries DC	Capital cost	770	$/kWh
	Replacement cost	770	$/kWh
	O&M	10	$/year
	lifetime	15	years
	Minimum state of Charghe	20	%
	Roundtrip Efficiency	90	%

Table 4.5: Energy components' characteristics

LIST OF ENERGY SYSTEM COMPONENTS			
Wind turbine (10 kW AC)	Capital cost	70000	$/kWh
	Replacement cost	70000	$/kWh
	O&M	500	$/year
	lifetime	20	years
Hydro turbine (10 kW DC)	Capital cost	112000	$/kWh
	Replacement cost	56000	$/kWh
	O&M	2400	$/year
	Pipe loss	15	%
	Design flow rate	70	l/s
	Minimum flow	50	%
	Maximum flow	150	%
	Efficiency	80	%
Hydro turbine (100 kW AC)	Capital cost	643783	$/kWh
	Replacement cost	322000	$/kWh
	O&M	6000	$/year
	Pipe loss	15	%
	Design flow rate	500	l/s
	Minimum flow	50	%
	Maximum flow	150	%
	Efficiency	80	%
PV System DC	Capital cost	3500	$/kWh
	Replacement cost	3500	$/kWh
	O&M	10	$/year
	Lifetime	25	years
Converter	Capital cost	420	$/kWh
	Replacement cost	420	$/kWh
	O&M	0	$/year
	Lifetime	15	years
	Efficiency	95	%

Table 4.6: Energy components characteristics

These information has been gathered from the Homer datasets of available technologies. They refer to generic components: no brand-specific items have been considered. To not neglect the meaningful impact of high cost linked to logistic, which, as stated before, in underdeveloped countries could also reach the 100% of the intrinsic cost of components, all the voices of cost have been increased by 40% according to the information in matter gathered on-site.

Chapter 5

Results

The rural administrative place of Namanjavira, in the northern Mozambique, has been therefore selected to test the performance of *GISEle* methodology in effectively interpreting and elaborating spatial data to provide a valuable electrification strategy plan for the area, completed with

- A statement of the capital cost needed to realize the project, separated in *"Energy generation costs"* and *"electric network cost"*;
- A forecast of resulting LCOE;
- A visual representation of the electric network topology;

In the following, the results produced by each step of the final global procedure are presented, together with some intermediate ones belonging to the development phase of *GISEle*, which turn useful to highlight the criticalities encountered along the work and help to better understand the choices made.

5.1 Data analysis

5.1.1 Terrain characteristics

Among the 22 different classes of land-cover defined by the *Global Land Cover 2000 Project (GLC 2000)* and adopted in the weighting procedure, Namanjavira is composed by only three as figure 5.1 clearly shows:

- Tree Cover, broadleaved, evergreen (>15% tree cover)
- Tree Cover, broadleaved, deciduous, closed (>50% Tree cover)
- Tree cover, broadleaved, deciduous, open (15-40% tree cover)

Since it results from automatic satellite measurements with a resolution of 911 meters, which is not sufficient to distinguish the wide variety of terrain, certainly, this

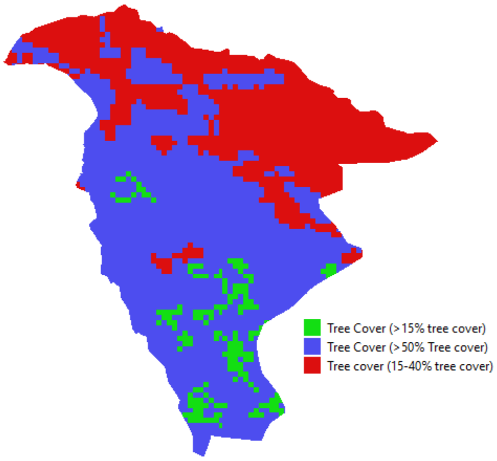

Tree Cover (>15% tree cover)
Tree Cover (>50% Tree cover)
Tree cover (15-40% tree cover)

Figure 5.1: Land cover typology of Namanjavira

assumption does not reflect the reality at all. For example, no cultivated areas has been detected which is quite unusual in a rural province of Mozambique. Nevertheless, these information are still useful to identify the natural setting of the region. The slope layer presented in figure 5.2, almost completely blue, denotes that Namanjavira is a plain area, with the exception of some singolarities. However superposing the river layer realized through the hydrological analysis it becomes clear that the whole region is divided by a ridge located in the northern part, that crosses it from east to west defining the two major watershed basins which characterize the hydrology of Namanjavira: one in the north and the other in the south as shown in figure 5.3.

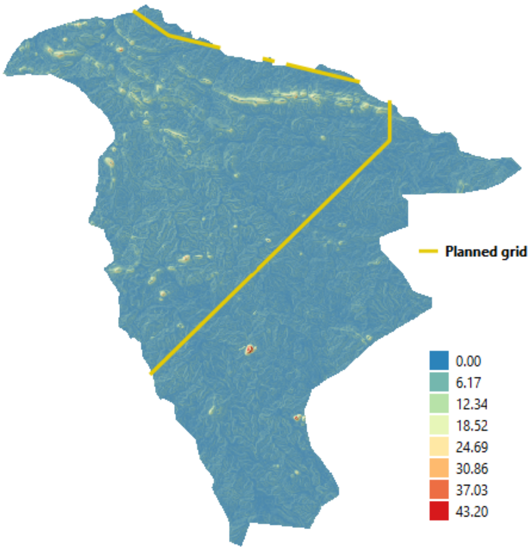


Figure 5.2: Slope layer of Namanjavira presented together with the HV transmission line that the government plan to build in the next future.

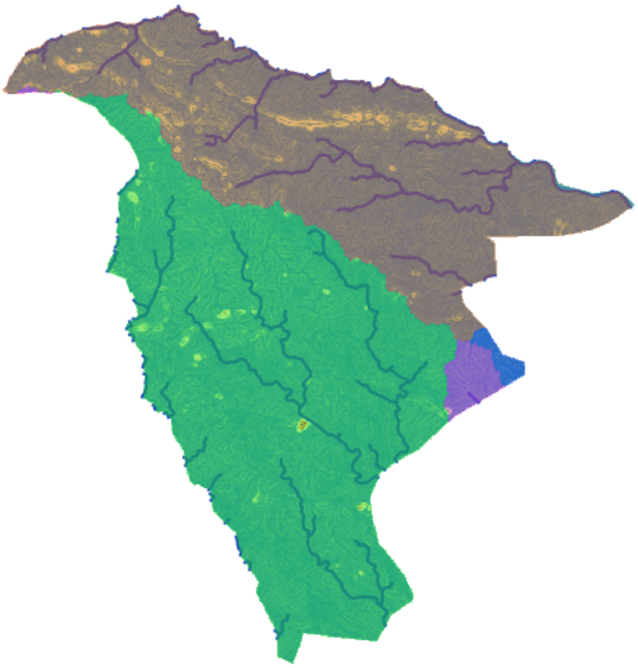


Figure 5.3: Result of hydrologic analysis highlighting rivers' path and watershed basins.

In Europe, or in urban context in general, we use to see people and houses grouped in cities and villages, radially distributed around the center. The downtown represents the core of the commercial and social activity of the community and the place where the population density is the highest, whereas it decreases moving away towards the countryside. In the rural areas of Mozambique, instead, the situation is completely different: rather than be collected in villages, houses are distributed along the main roads. It turns difficult to detect villages' centers because, in such framework, is along the roads that business and commercial activities grows. (figure 5.5) In Namanjavira, the primary connection route runs along the ridge previously defined (figure 5.4), and there the majority of people is concentrated.

As we will see analysing the energy resource, being the main fraction of population located over a hill, i turns difficult to find exploitable water flow in the nearby. The map 5.3 clearly show how water streams born on both sides of the ridge, therefore the availability of water will be limited.

5.1.2 Energy resources

As proof of the last assumption figure 5.6 presents the hydropower potential map resulting from the hydrological analysis. It must be underlined that this is the maximum power which could be extracted exploiting all the available water with an ideal turbine having an efficiency of 100%.

In the nearby of the ridge, indeed, the estimated potential is very limited, suitable only for pico power plants. Values increase to interesting value in the southern part

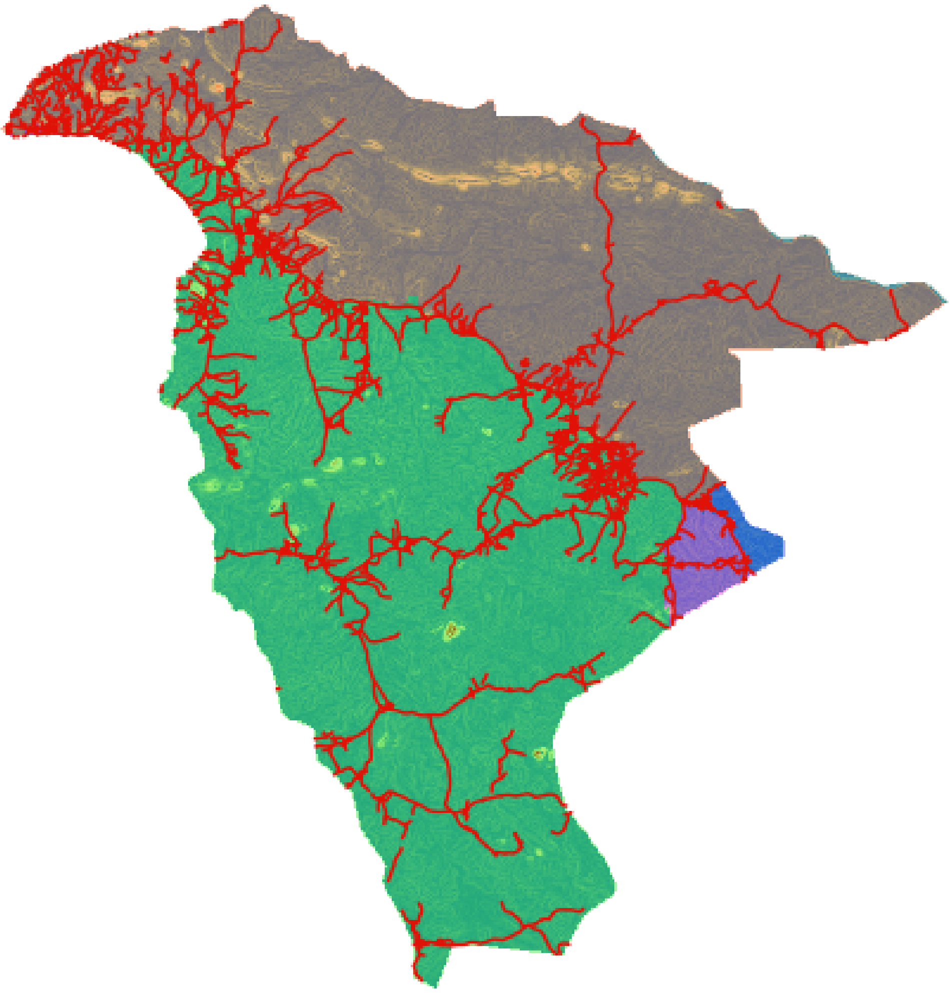


Figure 5.4: Namanjavira Road Network

of the region and at the northern border, where the streams born along ridge's slopes join to a larger river. Moreover data about monthly precipitation, depicted in table below, confirm the high seasonality of water availability discussed in chapter 4.3.2, which undoes the hydroelectric production for several months in dry season.

NAMANJAVIRA RAINFALL VALUES			
January	→	215-309	mm
April	→	80-156	mm
August	→	17-35	mm
November	→	96-153	mm

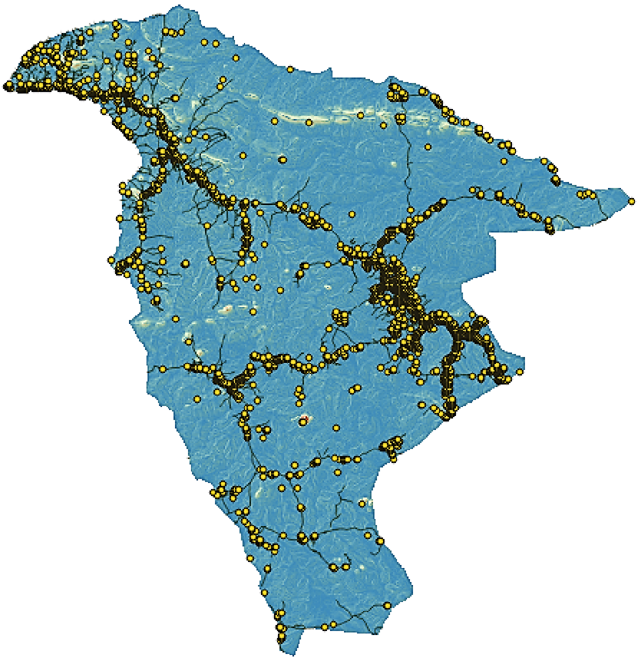


Figure 5.5: Populated points distribution along road network

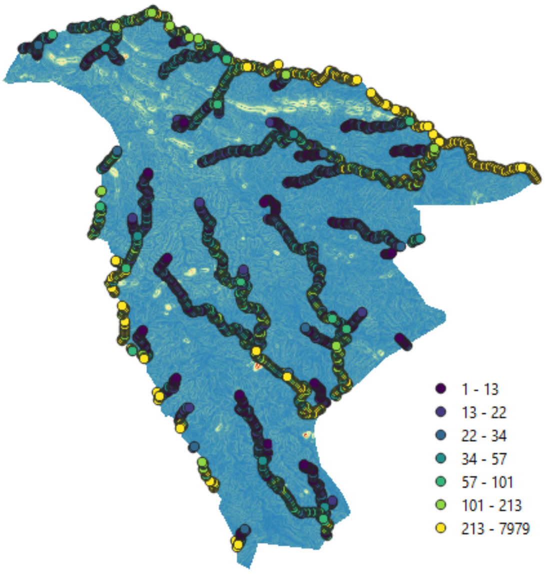
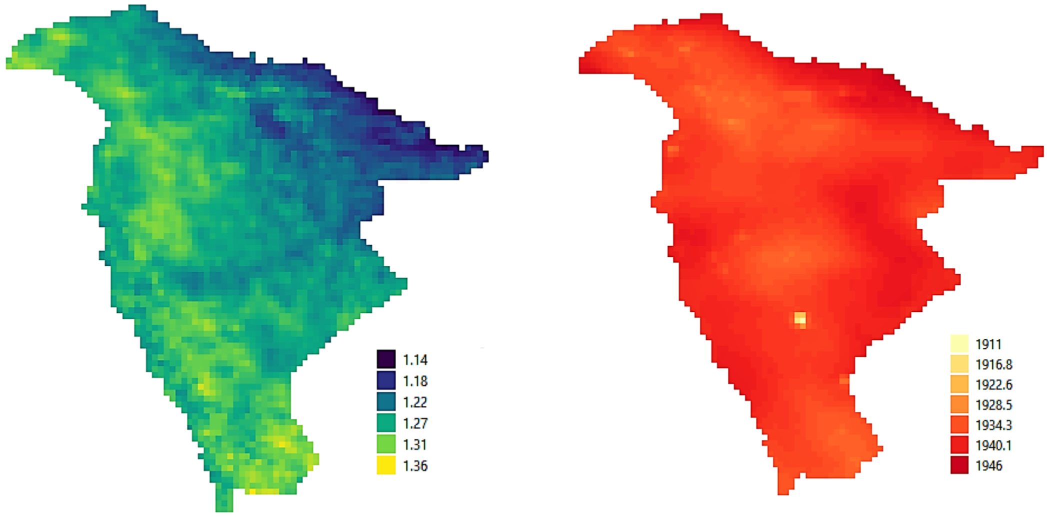


Figure 5.6: Mean annual Hydropower potential of Namanjavira's water streams. Measurements are provided in kW

Whereas wind does not result to be exploitable for energy production since the average mean wind speed is wherever below 2 m/s, the Global Horizontal Irradiation on the area is quite high, ranging between 1911 and 1946 $[kW\ m^{-2}day^{-1}]$ making the sun the primary renewable candidate to feed the electric grid. The point layer resulting from the discretization of the region, whose points embed values of penalty factor has been already shown with figure 3.9 in chapter 3.2.



(a) Average annual wind speed [m/s]

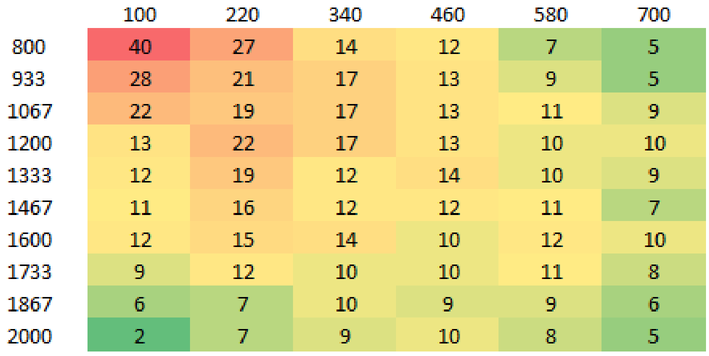
(b) GHI [$kWm^{-2}day^{-1}$]

Figure 5.7: Renewable energy sources availability in Namanjavira

5.2 Clustering

As explained in the previous chapters, the two *DBSCAN*'s defining parameters *eps* and *minPop* have to be chosen by the user in order to fit the characteristics of the target area. Tabs show the output obtained from each of their combination, giving the following spans:

	Min	Max
eps	100	700
minPop	800	2000

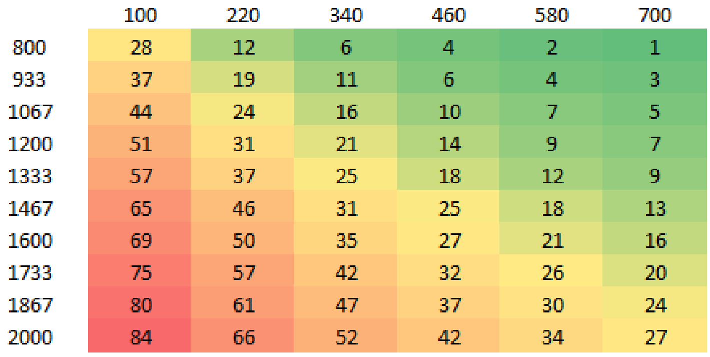


	100	220	340	460	580	700
800	40	27	14	12	7	5
933	28	21	17	13	9	5
1067	22	19	17	13	11	9
1200	13	22	17	13	10	10
1333	12	19	12	14	10	9
1467	11	16	12	12	11	7
1600	12	15	14	10	12	10
1733	9	12	10	10	11	8
1867	6	7	10	9	9	6
2000	2	7	9	10	8	5

Figure 5.8: Number of clusters proposed

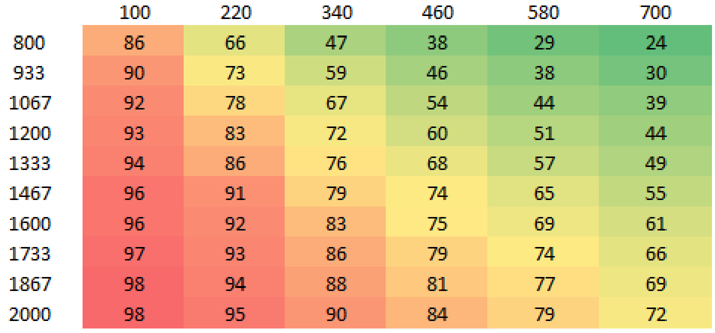
Within them the final combination chosen is:

eps	:	480
minPop	:	1140



	100	220	340	460	580	700
800	28	12	6	4	2	1
933	37	19	11	6	4	3
1067	44	24	16	10	7	5
1200	51	31	21	14	9	7
1333	57	37	25	18	12	9
1467	65	46	31	25	18	13
1600	69	50	35	27	21	16
1733	75	57	42	32	26	20
1867	80	61	47	37	30	24
2000	84	66	52	42	34	27

Figure 5.9: Percentage of region area covered by clusters



	100	220	340	460	580	700
800	86	66	47	38	29	24
933	90	73	59	46	38	30
1067	92	78	67	54	44	39
1200	93	83	72	60	51	44
1333	94	86	76	68	57	49
1467	96	91	79	74	65	55
1600	96	92	83	75	69	61
1733	97	93	86	79	74	66
1867	98	94	88	81	77	69
2000	98	95	90	84	79	72

Figure 5.10: % of total population gaining electricity access

It leads to 13 clusters, covering the 13% of the total area with an approximate value of clustered people of 60%. The reason these values were selected can be seen by comparing the corresponding configurations output in terms of number of clusters, percentages of clustered area and population respectively over the total target area and population. The idea in this selection is:

- To maximize the ratio between the clustered population over clustered area;
- To maintain an as much as possible low value of clustered area percentage in order to avoid single or few clusters covering gigantic areas;
- To identify a parameters' combination capable of clustering a population above a certain percentage;
- To avoid both low (one or two gigantic clusters) and too high (lot of very small clusters) number of clusters;

It has already been anticipated how the functioning of *DBSCAN*, based on the core and border points, could lead to some adjacent final clusters. In order to deal with this issue, clusters were aggregated, and the image 5.11 shows the graphic representation of the associated output. The reasons are both their mutual proximity and relative dimensions. The clusters' aggregation process is linked to the user sensitivity, and obviously leads to different results. Overlapping these clusters to

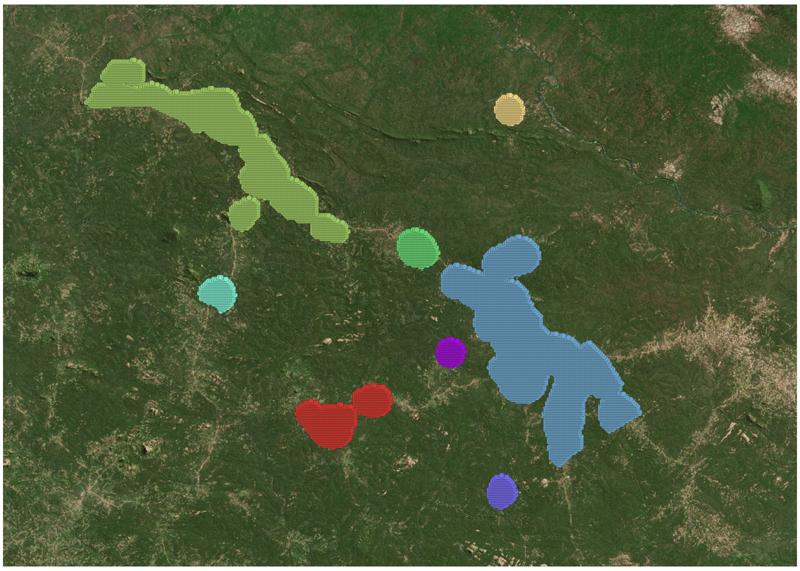


Figure 5.11: Clustering process output

the aerial images from Bing (figure 5.12), it can be seen how the algorithm is able to identify zones in which high numbers of houses are found. Figure 5.13 shows the internal population density distribution in cluster n.2: it is intuitive to the differences between densely populated with respect to border areas. An ulterior validation is the good superposition level of many clusters over the existing formally recognized communities. For instance, clusters n. 0, 2, 5, 6 and 12 (figure 5.14) shows how, despite being not formalized yet, many remote rural areas have experimented relatively high population growth rates and have now population comparable to other recognized one.

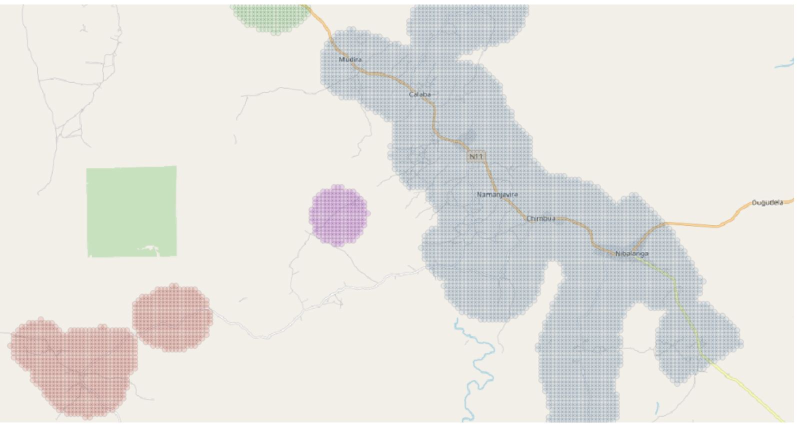


Figure 5.12: Cluster view over Bing map

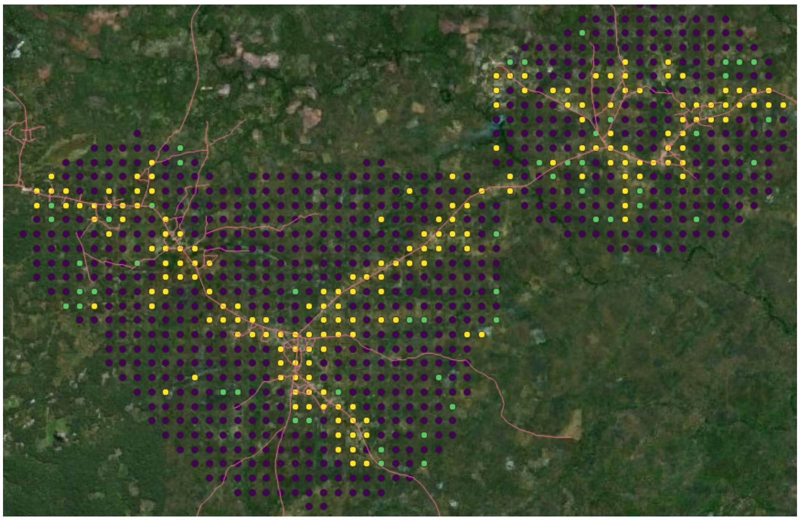


Figure 5.13: Population density

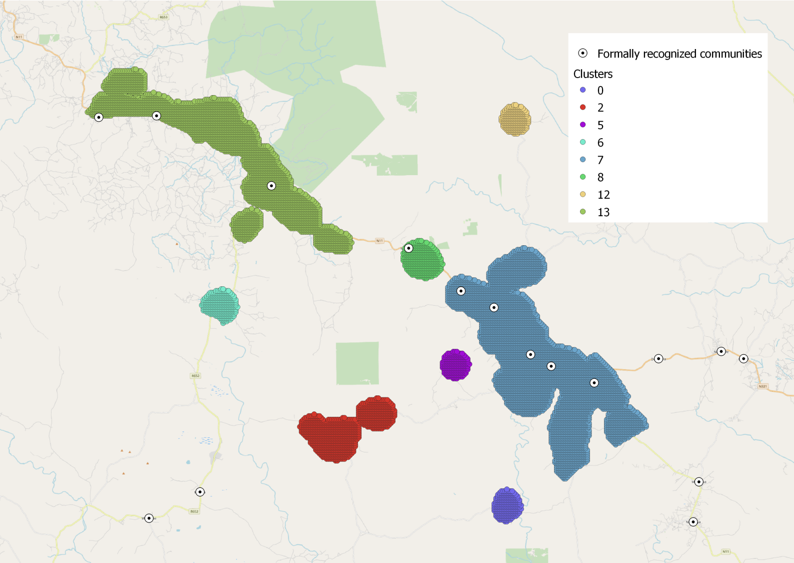


Figure 5.14: Recognized villages

5.3 Grid routing

Initially, we tried to launch the grid routing algorithm on the whole Namanjavira, without accomplish any clustering process. All the points representing at least 10 people were selected and classified as terminal points which must be part of the final grid. The algorithm accomplished the task and returns the grid topology illustrated in figure 5.15.

The grid is wide with a total length of 378 km: in areas with high population density it results to be very thick and connections are short, whereas long lines has been traced throughout the territory in order to connect even the remotest points. This happens in particular in the southern and northern extremes of the map. It is observing such long connections that it results possible to really appreciate the ability of the routing algorithm to find not the shortest, but the least cost path connecting two terminals. Figure 5.16 superposes a portion of the grid located in the southern area of Namanjavira to the *penalty factor* layer showing how the algorithm, instead of follow a rectilinear path across a costly terrain, recommends to exploit where possible the presence of the road to reduce costs.

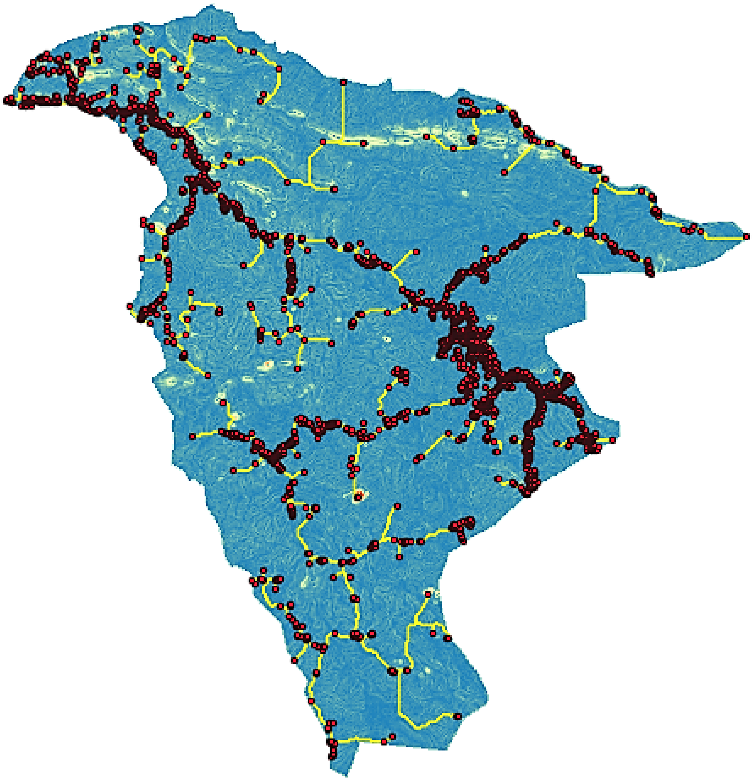


Figure 5.15: Electric network covering the whole Namanjavira

Applying the algorithm to limited areas within the clusters' borders the final topology is completely different (figure 5.17). This time single long connection in remote areas are no more clearly identifiable, since all the populated nodes too isolated has been discarded by the clustering process which classified them as outliers. They could be considered as candidate for stand-alone energy systems. The total length of all the micro-grid realized, one for each cluster, is expected to be lower than that related to the single grid presented in figure 5.15, simply because electrified area seems to be smaller. Reasoning by clusters, the entire southern part of the region and the main portion of northern one, remains unelectrified. But it is false. The limited area of a cluster reduces the global number of points that the grid algorithm must manage: it allow us to increase the number of populated terminal points without incur in memory problems. In this analysis, therefore, all the cluster nodes with value of population greater than 5 (the average number of people constituting an household) have been connected to the final grid.

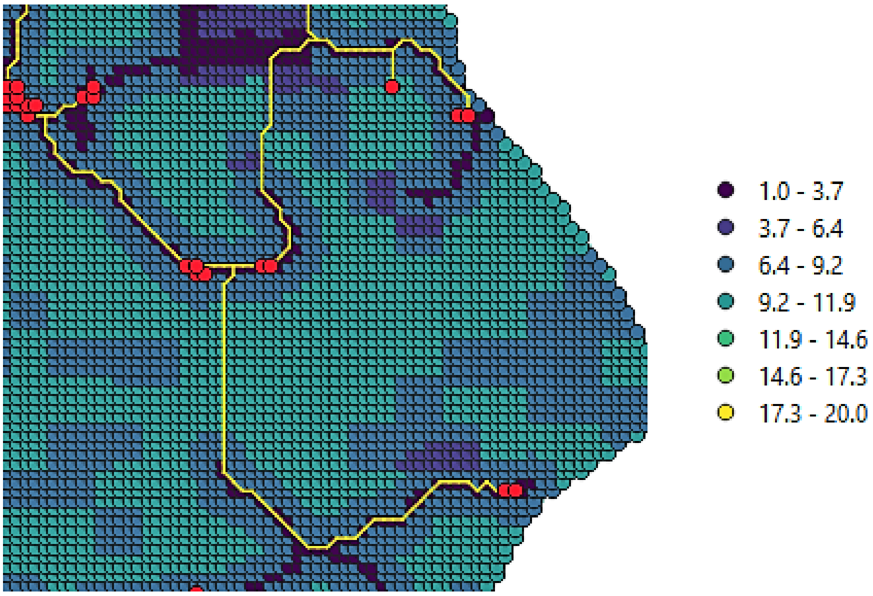


Figure 5.16: Long lines connections over penalty factor layer

This lead to a big increase in connections and despite the covered area is smaller, the total length of electric lines created is 664 km. An example is shown if figure 5.18 related to cluster five, where the yellow lines belongs to the whole regional grid connecting cells with at least 10 people; red lines instead represents the grid modelled within the single cluster including all the points with more than 5 people. Including in the map also those lines which could link each cluster to the national HV transmission line (figure 5.19) it becomes possible to start advancing hypotheses about the best strategy to adopt to in order to feed cluster's consumers.

Clusters 6, 0 and 3 are small and located far from the future transmission line: gain access to national grid make available energy at lower prime, but their low energy demand probably will not justify the realization of a such long connection. Also cluster 13 is quite far from the nearest substation, however its big energy need would make the connection to the grid the optimal economic solution. 5 and 12, instead, are even smaller than 6 or 3, but their proximity to substation probably will bring to prefer grid connection solution to Micro-grid one. The results of the economic analysis carried out, reveal now if the visual forecasts were correct.

5.4 Economic Evaluation

The tables 5.2 to 5.10 depict the summary of the techno-economic analysis performed for each cluster, ending with a comparison between COE values of the two covered strategies:

- HV line connection

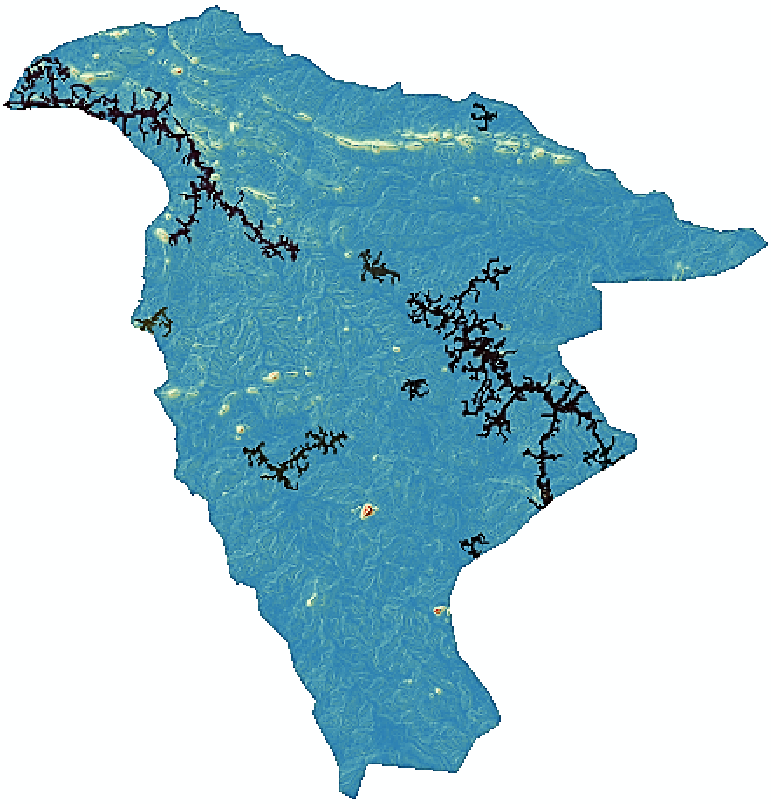


Figure 5.17: Clusters electric grid

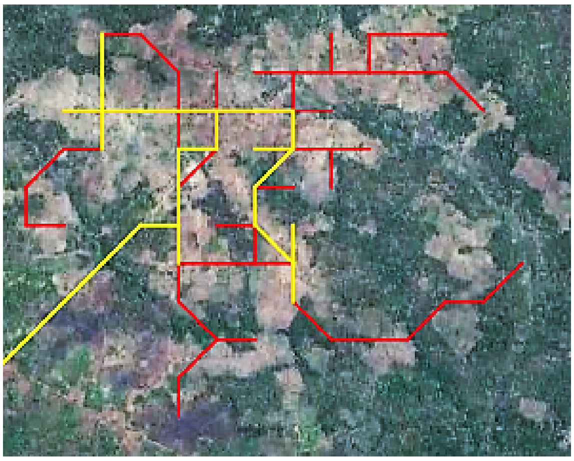


Figure 5.18: Expansion of electric network when the population threshold decreases from 10 to 5.

- Isolated Micro-grid

In order to evaluate the cost of energy in micro-grid mode, every time three configurations for energy generation proposed by HOMER have been taken into account. They differ in percentage of renewable production. Independently from the size of the cluster considered, the solution relying on 100% of renewable energy production reveals to be always the most expensive one: in none of the cases analyzed this kind of configuration turns to be more affordable than grid connection. With respect to

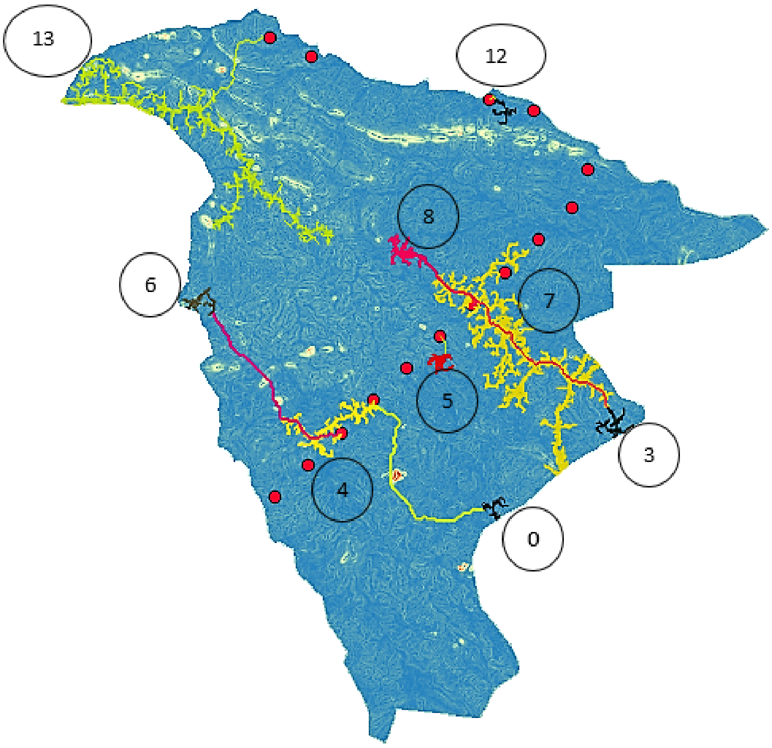


Figure 5.19: Clusters' internal grid and their possible connection to HV line

the forecasts previously made:

- Actually, cluster 6 and 0 are too small and far from the grid to make the grid connection convenient. Nevertheless, the same conclusion is not true for cluster 3. Why? The length of cluster 3's grid connection is comparable to those of the other two, and the same is true also regarding the energy demand. The difference lies in the costs of the grid connections: cluster 3, being located on the main road, could follow a very cheap path to reach the nearest substation, whereas 0 and 3 located in remote areas would need to build the grid connection through impervious ground which almost double the final infrastructural expenditure as shown in figure 5.20.

- Whichever combination of micro grid generation system is analyzed, the HV grid connection results always convenient for cluster 12. The same is true also for cluster 5, with the exception of the cheapest micro-grid solution, whose COE is slightly lower than HV-network one.

- As expected, the high energy demand characterizing clusters 7 and 13 always justify the realization of HV-network link.

The internal grid cost analysis on the overall LCOE reveals it impacts from 50 to 200%, underlying the fundamental importance of an appropriate electric network modelling.

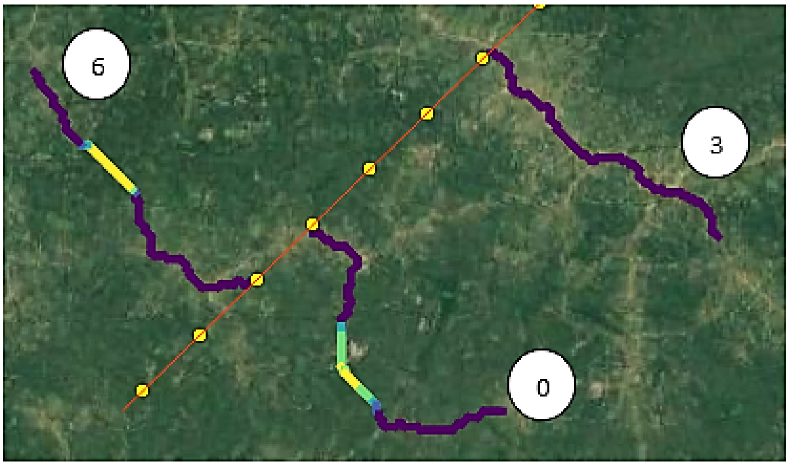


Figure 5.20: Grid connection complexity

$/kWh						
	SOL_1	SOL_2	SOL_3	HV	SOL_3 vs HV	Suggestion
0	1,06596101	1,65856101	1,23156101	2,20641017	-44,18%	SOL_3
3	1,082754086	1,558654086	1,193554086	0,949602549	25,69%	HV
4	1,095637782	1,694837782	1,254937782	0,717694925	74,86%	HV
5	1,40069968	2,04879968	1,62879968	1,474287348	10,48%	HV/SOL_1
6	1,440921303	2,160321303	1,620321303	2,069372963	-21,70%	SOL_3
7	0,843332913	1,628532913	0,844532913	0,513102535	64,59%	HV
8	1,496454942	2,323454942	1,607954942	1,228991736	30,84%	HV
12	1,312093482	1,865993482	1,476993482	0,950283549	55,43%	HV
13	0,819433373	1,629933373	0,874033373	0,484072761	80,56%	HV

Table 5.1: LCOE of each proposed solution for each cluster

Cluster n.	**0**	
N. Houses	**123**	
Electric Infrastructure		
Grid cost	552215,6	$
Grid length	11,79	km
Link cost	1645262	$
Link length	24,54	km
Energy data		
Energy needs	42596	kWh/yr
Energy needs for 25 years	1064900	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 23.9%)		
COE generation from Homer	0,5474	$/kWh
NPC generation	301415	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,14	$/kWh
NPC generation	624562,6	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 67.3%)		
COE generation from Homer	0,713	$/kWh
NPC generation	392765,6	$

Table 5.2: Results for cluster number 0

Cluster n.	**3**	
N. Houses	**212**	
Electric Infrastructure		
Grid cost	1003337,5	$
Grid length	19,41	km
Link cost	499367	$
Link length	21,92	km
Energy data		
Energy needs	74507	kWh/yr
Energy needs for 25 years	1862675	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 24.2%)		
COE generation from Homer	0,5441	$/kWh
NPC generation	524061,8	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,02	$/kWh
NPC generation	984197,6	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 54.9%)		
COE generation from Homer	0,6549	$/kWh
NPC generation	0,6549	$

Table 5.3: Results for cluster number 3

Cluster n.	**4**	
N. Houses	**568**	
Electric Infrastructure		
Grid cost	2889379	$
Grid length	56,26	km
Link cost	0	$
Link length	0	km
Energy data		
Energy needs	201057	kWh/yr
Energy needs for 25 years	5026425	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 25.1%)		
COE generation from Homer	0,5208	$/kWh
NPC generation	1353532	$
SOLUTION 2: 100% RENEWABLE (100 kW Microhydro)		
COE generation from Homer	1,12	$/kWh
NPC generation	2914588	$
SOLUTION 3: Best compromize (100 kW Microhydro, Renewable fraction = 60.4%)		
COE generation from Homer	0,6801	$/kWh
NPC generation	1767817	$

Table 5.4: Results for cluster number 4

Cluster n.	**5**	
N. Houses	**101**	
Electric Infrastructure		
Grid cost	742148	$
Grid length	12,14	km
Link cost	421988	$
Link length	2,25	km
Energy data		
Energy needs	34974	kWh/yr
Energy needs for 25 years	874350	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 23.9%)		
COE generation from Homer	0,5519	$/kWh
NPC generation	249531,7	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,2	$/kWh
NPC generation	249531,7	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 71.2%)		
COE generation from Homer	0,78	$/kWh
NPC generation	352504	$

Table 5.5: Results for cluster number 5

Cluster n.	**6**	
N. Houses	**150**	
Electric Infrastructure		
Grid cost	1193691	$
Grid length	16,54	km
Link cost	1360580	$
Link length	22,86	km
Energy data		
Energy needs	53034	kWh/yr
Energy needs for 25 years	1325850	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 24.7%)		
COE generation from Homer	0,5406	$/kWh
NPC generation	370668,8	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,26	$/kWh
NPC generation	860032,2	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 63.5%)		
COE generation from Homer	0,72	$/kWh
NPC generation	493410,6	$

Table 5.6: Results for cluster number 6

Cluster n.	**7**	
N. Houses	**4647**	
Electric Infrastructure		
Grid cost	15148011	$
Grid length	287,86	km
Link cost	70389	$
Link length	0	km
Energy data		
Energy needs	1644142	kWh/yr
Energy needs for 25 years	41103550	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 32.8%)		
COE generation from Homer	0,4748	$/kWh
NPC generation	10092460	$
SOLUTION 2: 100% RENEWABLE (100 kW Microhydro)		
COE generation from Homer	1,26	$/kWh
NPC generation	26829600	$
SOLUTION 3: Best compromize (100 kW Microhydro, Renewable fraction = 33.9%)		
COE generation from Homer	0,476	$/kWh
NPC generation	10115940	$

Table 5.7: Results for cluster number 7

Cluster n.	**8**	
N. Houses	**193**	
Electric Infrastructure		
Grid cost	1619038	$
Grid length	22,42	km
Link cost	225300	$
Link length	0	km
Energy data		
Energy needs	67923	kWh/yr
Energy needs for 25 years	1698075	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 24.5%)		
COE generation from Homer	0,543	$/kWh
NPC generation	476765	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,37	$/kWh
NPC generation	1203327	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 50.2%)		
COE generation from Homer	0,6545	$/kWh
NPC generation	574690,4	$

Table 5.8: Results for cluster number 8

Cluster n.	**12**	
N. Houses	**109**	
Electric Infrastructure		
Grid cost	722734	$
Grid length	11,06	km
Link cost	39093	$
Link length	1,05	km
Energy data		
Energy needs	37741	kWh/yr
Energy needs for 25 years	943525	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 24.8%)		
COE generation from Homer	0,5461	$/kWh
NPC generation	266449	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,1	$/kWh
NPC generation	534508,1	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 80.5%)		
COE generation from Homer	0,711	$/kWh
NPC generation	346862	$

Table 5.9: Results for cluster number 12

Cluster n.	**13**	
N. Houses	**2818**	
Electric Infrastructure		
Grid cost	8220257	$
Grid length	226,53	km
Link cost	281096	$
Link length	0	km
Energy data		
Energy needs	996596	kWh/yr
Energy needs for 25 years	24914900	kWh
Energy cost from national grid	0,142857143	$/kWh
SOLUTION 1: CHEAPEST WITH GENSET (Renewable fraction= 24.1%)		
COE generation from Homer	0,4895	$/kWh
NPC generation	6306808	$
SOLUTION 2: 100% RENEWABLE (10 kW Microhydro)		
COE generation from Homer	1,3	$/kWh
NPC generation	16740490	$
SOLUTION 3: Best compromize (10 kW Microhydro, Renewable fraction = 35,3%)		
COE generation from Homer	0,5441	$/kWh
NPC generation	7010000	$

Table 5.10: Results for cluster number 12

Chapter 6

Conclusions

In a global framework where guaranteeing electricity access to all is set to be one of the main challenges that the international community is going to face in the next future, *GISEle* places itself among the available approaches to be adopted when allocating resources and capitals at this purpose.

Nowadays State of the Art tools and methodologies often neglect important aspects of rural electrification strategy planning. This is mainly due to not holistic approaches to the problem, usually addressing those aspects in a separate manner, not taking into account their intrinsic interconnections. Spatial dimension of the target areas has a significant impact on rural infrastructures' total cost, therefore, when ignored, the final results can lead to wrong strategy implementations. When discerning between different electrification strategies, not considering electric networks as a fundamental component of energy supply systems is, from the authors' perspective, an error which needs to be avoided, since it leads to wrong evaluation of the LCOEs and consequently of theoretically cost-reflective tariffs.

GISEle project's goal is to create a methodology capable of having a more holistic approach in rationalizing investments and resources deployment in tackling energy access in the developing world. This procedure is suitable for all kind of stakeholders: international organizations, private sector and governments. // Its final framework is composed by several steps. Starting from GIS data, spatial characteristics of the target area are efficiently analyzed in order to extract valuable insights and design the optimal electric network topology. Finally, discrimination between standalone and grid-connected configurations is performed for each single energy community and a comprehensive optimal electrification strategy is suggested.

The fundamental algorithms on which the proposed methodology is based, are specific adaptations introduced in order to optimally fit the addressed problem.

Within GIS data management environment, the total area is segmented into a regular grid of quadratic cell, each one having geo-specific attributes concurring in defining its suitability for infrastructural building. A remarkable achievement is the consistency of a totally open-source data-based procedure. Resources, loads and

terrain morphology publicly available datasets have sufficiently high resolutions to obtain valuable results: their significance has been evaluated on field and the outcome is that for the principal ones the
Thereafter, the weighting process, based on the definition of the *penalty factor*, representing the relative additional costs linked to spatial exogenous factors, has been developed as the the foundation on which the optimal networks are designed.
Parallelly, clustering the population and hence the electric loads of the entire area needs to be performed. This process has to be done considering the population attribute of the previously created regular points grid. It is an exploratory process with which, the algorithm needs to browse the entire target area in order to find densely populated area suitable for electrification.The main objective of the clustering process is to move beyond the basic approach with which each cell is considered by itself, neglecting the strategic value of closely located highly populated areas: the algorithm proceeds by identifying valuable groups of cells instead of only one cell at a time.
Then a new approach for electric grid routing combining the potentialities of MST and Dijkstra algorithms to produce an approximate solution to the Steiner problem was developed. Its strength relies in the ability to split the original problem in multiple sub-problems which can be faced separately. Instead of dealing with all the terminal points of a cluster in a single operation, each sub-problem involves only one points couple at a time, thus sharply reducing the corresponding computational burden. This algorithm is used is iteratively applied on each cluster defining both its internal electric network, and the eventually necessary infrastructure connecting it to the closest national electric grid's substation.
A new approach for energy assessment process automatization has also been developed through the definition of reference consumers load profiles and identification of optimal region specific proxies. In the presented case study, the reference loads have been modelled through *Load Pro Gen* tool based on data collected along an actual energy needs assessment process and the exploited proxy has been the total households number.
Finally, optimal power supply systems are sized through *HOMER Energy Pro* tool, and associated costs are combined with clusters' internal and external grids costs in order to perform the final discrimination in terms of energy electrification strategy.

The obtained results from GISEle deployment in Namanjavira province are promising and create pre-conditions for the application of the same approach not only for the optimal electric grid routing but also to identify suitable locations for on-grid power plants. Lastly it is worth to emphasize how the Mozambique *Renewable Energy Atlas*, has been an extremely both time and resources-consuming process, which with the help of *GISEle* could be done in a much more efficient way, having as added benefit, much more structured results as effective costs of the energy access, and concrete infrastructure design.

Appendix A

Load profiles' input parameters

In the present Appendix, the configuration parameters used to create LoadProgen daily load profiles for various user categories, are reported.

Primary School

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	7	25	16-21	0	5	5	0
Internal Lights	15	20	16-21	0	5	5	0
Fridge	1	60	8-18	0	10	10	0

Table A.1: LoadProGen configuration parameters for primary schools

Secondary School

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	5	25	16-21	0	5	5	0
Internal Lights	25	20	16-21	0	5	5	0

Table A.2: LoadProGen configuration parameters for secondary schools

Places of Worship

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	3	25	16-21	0	7	7	0
Internal Lights	8	20	16-21	0	7	7	0
Sound System	1	20	16-21	0	2	6	30
Phone Charger	1	5	16-21	0	1	3	90

Table A.3: LoadProGen configuration parameters for work place

Administrative Headquarters

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	2	25	16-18	0	2	2	0
Internal Lights	6	20	16-18	0	2	2	0
Computer	2	50	8-18	0	2	6	20
Phone Charger	4	5	8-18	0	1	3	90

Table A.4: LoadProGen configuration parameters for administrative headquarters

Police Station

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	1	20	16-5	0	12	12	0
Internal Lights	4	15	16-5	0	4	6	0
Phone Charger	1	5	18-5	0	1	3	30

Table A.5: LoadProGen configuration parameters for police stations

World Vision Office

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	5	20	16-18	0	1	1	30
Internal Lights	9	15	16-18	0	1	1	30
Fridge	2	70	8-18	0	1	3	0
Phone Charger	4	5	8-18	0	1	5	20
Electronics	5	100	8-18	0	1	5	20

Table A.6: LoadProGen configuration parameters for world vision office

Public lights

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
Street Lights	30	30	16-5	0	13	13	0
Community Lights	30	25	16-22	0	7	7	0

Table A.7: LoadProGen configuration parameters for public lights

Merchants

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var	
		W	h	%	h	h	%	
External Lights	2	25	16-22	0	6	6	0	
Internal Lights	3	20	16-22	10	6	6	0	
Freezer	1	300	10-18	0	8	8	0	Sound System
1	20	10-22	0	2	6	30		
Phone Charger	1	5	10-22	0	1	3	90	
Fan	1	80	10-22	0	3	6	10	

Table A.8: LoadProGen configuration parameters for merchants

Tailors

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	1	25	16-19	0	3	3	0
Internal Lights	3	20	16-19	0	4	6	10
Phone Charger	1	5	10-19	0	1	3	90
Radio	1	5	10-19	0	3	6	25
Fan	1	60	10-19	0	3	6	30

Table A.9: LoadProGen configuration parameters for tailor

Barbers

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	1	20	16-19	0	7	7	0
Internal Lights	3	15	16-19	0	4	6	10
Radio	1	5	10-19	0	1	5	25
Phone Charger	1	5	10-19	0	1	3	90

Table A.10: LoadProGen configuration parameters for barbers

"Standard" Households

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	1	15	16-21	0	7	7	0
Internal Lights	3	15	16-21	0	4	6	10
Radio	1	5	10-21	0	1	5	25
Phone Charger	1	5	10-21	0	1	3	90
Fan	1	25	10-21	0	1	4	30

Table A.11: LoadProGen configuration parameters for standard household consumers

"Premium" Households

	Number	Power	Time Window	FT var	Functioning Cycle	Functioning Time	TF var
		W	h	%	h	h	%
External Lights	1	20	16-21	0	5	5	0
Internal Lights	3	15	16-21	0	3	4	20
TV	1	100	10-21	0	2	7	20
Fridge	1	70	10-21	0	11	11	0
Radio	1	5	10-21	0	1	5	25
Phone Charger	1	5	10-21	0	1	3	90

Table A.12: LoadProGen configuration parameters for premium household consumers

Bibliography

[1] IEA. About energy access, 2016.

[2] IEA. International Energy Agency - Energy Access Outlook 2017: From poverty to prosperity. *Energy Procedia*, 94(March):144, 2017.

[3] Monique Kremer and Peter Van Lieshout. *Doing Good or Doing Better development policies in a global izing world.* 2009.

[4] IEA. https://www.iea.org/sdg/electricity/.

[5] http://www.onsset.org/#.

[6] Magdi Ragheb. Economics of Wind Power Generation. In *Wind Energy Engineering: A Handbook for Onshore and Offshore Wind Turbines*, pages 537–555. Academic Press, jan 2017.

[7] Rajesh Govindan, Tareq Al-Ansari, Anna Korre, and Nilay Shah. Assessment of technology portfolios with enhanced economic and environmental performance for the energy, water and food nexus. In *Computer Aided Chemical Engineering*, volume 43, pages 537–542. Elsevier, jan 2018.

[8] GIS wiki.

[9] A Gentle Introduction to GIS.

[10] Z Sumic, T Pistorese, and S.s Venkata. Automated underground residential distribution design–Part 1: Conceptual design. *IEEE Transactions on Power Delivery - IEEE TRANS POWER DELIVERY*, 8:637–643, 1993.

[11] Cláudio Monteiro, Ignacio J. Ramírez-Rosado, Vladimiro Miranda, Pedro J. Zorzano-Santamaría, Eduardo García-Garrido, and L. Alfredo Fernández-Jiménez. GIS spatial analysis applied to electric line routing optimization. *IEEE Transactions on Power Delivery*, 20(2 I):934–942, 2005.

[12] Jaime Cevallos-Sierra and Jesús Ramos-Martin. Spatial assessment of the potential of renewable energy: The case of Ecuador, 2018.

[13] Esri. ArcGIS.

[14] Esri. QGIS.

[15] GRDC. `https://www.bafg.de/GRDC/EN/Home/homepage{_}node.html;jsessionid=F70EE3B04577F983E6EDE61925FE9E97.live21304`.

[16] Soil Conservation Service. https://www.nrcs.usda.gov/wps/portal/nrcs/site/.

[17] I. Sahu and A. D. Prasad. ASSESSMENT OF HYDRO POTENTIAL USING INTEGRATED TOOL IN QGIS. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-5(November):115–119, nov 2018.

[18] Henry A. Adornado and Masao Yoshida. GIS-BASED WATERSHED ANALYSIS AND SURFACE RUN-OFF ESTIMATION USING CURVE NUMBER (CN) VALUE. *Journal of Environmental Hydrology*, 18(Paper 9):1–14, 2010.

[19] Gah-Muti Salvanus Yevalla, Dadjeu Nguemeu Seidou, Bang Vu Ngoc, Tran Van Hoi, and Tabod Charles Tabod. Hydrological Studies for the Assessment of Run-of-River Hydropower Potential and Generation over the Wouri-Nkam River using GIS and Remote Sensing Techniques. *Aquademia: Water, Environment and Technology*, 2(1):1–7, 2018.

[20] Ketul Shah, A. T. Motiani, Indra Prakash, and Khalid Mehmood. Application of SCS-CN Method forEstimation of Runoff Using GIS. *International Journal of Advance Engineering and Research Development*, (April), 2017.

[21] Scs Usda. Urban Hydrology for Small Watersheds. Technical Report Technical Release 55 (TR-55), 1986.

[22] Global Geospatial Potential EvapoTranspiration & Aridity Index Methodology and Dataset Description.

[23] Kyoung Jae Lim, Bernard A. Engel, Suresh Muthukrishnan, and Jon Harbor. Effects of initial abstraction and urbanization on estimated runoff using CN technology. *Journal of the American Water Resources Association*, 42(3):629–643, 2006.

[24] H. E. Driver and A. L. Kroeber. Quantitative expression of cultural relationships. Quantitati:211–256, 1932.

[25] Joseph Zubin. A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, 33(4):508–516, 1938.

[26] Robert Choate Tryon. Cluster analysis; correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality. 1939.

[27] Raymond B. Cattell. The description of personality: basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, 38(4):476–506, 1943.

[28] Vladimir Estivill-Castro. Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75, 2007.

[29] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, jan 2007.

[30] Olaf R P Bininda-Emonds and George B Schaller. Phylogenetic Position of the Giant Panda:. *Giant Pandas*, pages 11–35, 2004.

[31] José J. López, José A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and José E. Ruiz. Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research*, 81(2):716–724, feb 2011.

[32] Datanovia. center-based clustering.

[33] Xiaowei and others Ester, Martin and Kriegel, Hans-Peter and Sander, Jörg and Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96:226–231, 1996.

[34] Lei Gong, Toshiyuki Yamamoto, and Takayuki Morikawa. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. In *Transportation Research Procedia*, volume 32, pages 146–154. Elsevier, jan 2018.

[35] Shan Zeng, Rui Huang, Haibing Wang, and Zhen Kang. Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models. *Neurocomputing*, 171:673–684, jan 2016.

[36] E.A.B.S.G. Williamson. *Lists, Decisions and Graphs*. S. Gill Williamson.

[37] Minimum Spanning Tree, Optimal Substructure, and Greedy Choice Property. Lecture 12 : Greedy Algorithms and Minimum Spanning Tree. pages 1–7, 2015.

[38] Camil Demetrescu, Irene Finocchi, and Giuseppe Italiano. *Algoritmi e Strutture Dati*. McGraw-Hill, 2004.

[39] Bang Ye Wu. Steiner Minimal Trees â. pages 1–6, 2004.

[40] Gabriel Robins and Alexander Zelikovsky. Minimum Steiner Tree Construction*. *Handbook of Algorithms for Physical Design Automation*, 2010.

[41] Magda Moner-Girona, Daniel Puig, Yacob Mulugetta, Ioannis Kougias, Jafaru AbdulRahman, and Sándor Szabó. Next generation interactive tool as a backbone for universal access to electricity. *Wiley Interdisciplinary Reviews: Energy and Environment*, 7(6):1–12, 2018.

[42] Douglas Ellman. The Reference Electrification Model: A Computer Model for Planning Rural Electricity Access. *Physics, Princeton University*, (2009):109, 2015.

[43] Carlos Mateo Domingo. RNM: Reference Network Model.

[44] Carlos Mateo Domingo, Tomás Gómez San Román, Álvaro Sánchez-Miralles, Jesús Pascual Peco González, and Antonio Candela Martínez. A reference network model for large-scale distribution planning with automatic street map generation. *IEEE Transactions on Power Systems*, 26(1):190–197, feb 2011.

[45] http://www.ecowrex.org/acp-eu.

[46] http://offgrid.energydata.info/#/?_k=85ph3x.

[47] Energydata.

[48] KTH. OnSSSET datasets.

[49] http://gaia.geosci.unc.edu/rivers/.

[50] Konstantinos M. Andreadis, Guy J.P. Schumann, and Tamlin Pavelsky. A simple global river bankfull width and depth database. *Water Resources Research*, 49(10):7164–7168, 2013.

[51] Conference Paper, Giuseppe Borruso, and Gabriella Schoier. Computational Science and Its Applications â ICCSA 2013. 7971(June 2013), 2013.

[52] Networkx. `https://networkx.github.io/documentation/latest/reference/algorithms/generated/networkx.algorithms.approximation.steinertree.steiner{_}tree.html`.

[53] Scipy. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.minimum{_}spanning{_}tree.html?highlight=minimum{%}20spanning{%}20tree{#}scipy.sparse.csgraph.minimum{_}spanning{_}tree`.

[54] Networkx. `https://networkx.github.io/documentation/latest/reference/algorithms/generated/networkx.algorithms.shortest{_}paths.weighted.dijkstra{_}path.html`.

[55] https://www.renewables.ninja/.

[56] FUNAE. http://gestoenergy.com/project/renewable-energy-atlas-of-mozambique/, 2013.

[57] Esto Integrato, Delle Condizioni, Tecniche Ed, Economiche Per, Connessione Alle, Reti Con, Obbligo Di, Connessione Di, Terzi Degli, and Impianti D I Produzione. TESTO INTEGRATO DELLECONDIZIONI TECNICHEED ECONOMICHE PER LA CONNESSIONE ALLE RETI CON OBBLIGO DI CONNESSIONE DI TERZI DEGLI IMPIANTI DI PRODUZIONE (TESTO INTEGRATO DELLECONNESSIONI ATTIVE âTICA). 2012:1–101, 2018.