
GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

# ARTIFICIAL INTELLIGENCE FOR UNIVERSAL ENERGY ACCESS

Autor: Carlos Mazuecos Reíllo

Director: Pablo Dueñas Martínez

Co-Director: Mario Castro Ponce

Co-Director: Andrés González García

Madrid

Junio de 2025

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Artificial Intelligence for Universal Energy Access en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2024/25 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Carlos Mazuecos Reíllo Fecha: 19/Junio/2025

Autorizada la entrega del proyecto

Fdo.: Pablo Dueñas Martínez Fecha: 19/Junio/2025

Quería empezar agradeciendo a mi madre y a mi padre, por haberme apoyado siempre en todas las decisiones que he tomado en mi vida y por haberme dado la mejor educación posible. A mi abuela y a mi abuelo por haber celebrado los logros siempre incluso más que yo y por haberme dado todo el cariño que tienen. A Flavio y Paloma por haberme aguantado en todas y por la piña que formamos los tres. Y me quiero acordar también del resto de la familia por haberme acompañado también en este camino.

A mi tutor Pablo, por haber estado siempre atento y dispuesto a ayudar cuando lo necesitaba. Por haberme escuchado y entendido ya fuera con el TFG u otras cosas que me han ido pasando estos dos años. A mi jefe de estudios David, porque nadie sabe el esfuerzo y las ganas que le ha llevado hacer esta carrera que tanto me ha gustado.

A Azoter0s porque aunque ahora no nos veamos todos los días, sé que están presentes siempre que les necesito. Que me han hecho disfrutar mucho de la vida fuera de la universidad y también me han ayudado a ser una mejor versión de mí mismo.

A Jorge porque más allá de haber sido mi compañero de proyectos los 4 años, ha sido un gran amigo y un apoyo cuando lo he necesitado. A todo iMAT porque, además de haber aprendido mucho de ellos, han hecho que este camino sea muy divertido. Y por último a Emma, por haber estado absolutamente siempre, tanto en las buenas, como para escucharme todos los lamentos de la carrera y levantarme en las malas. Porque sé que de la carrera me llevo a una mejor amiga que va a estar incondicionalmente siempre.

# ARTIFICIAL INTELLIGENCE FOR UNIVERSAL ENERGY ACCESS

**Author: Mazuecos Reíllo, Carlos.**
Director: Dueñas Martínez, Pablo.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

**Summary**

This project improves the scalability of the Reference Electrification Model (REM) by replacing its brute-force clustering step with a faster and technically feasible algorithm. The new method iteratively merges customer groups using spatial proximity and technical constraints, balancing cost and feasibility. The solution was developed on a dataset of 10,000 customers and tested on a dataset of more than 250,000 customers and achieved notable performance gains. Compared to the original REM output, the new approach reduced the cost per client while also cutting the total runtime.

**Key words:** REM, clustering, PRIM, minimum spanning tree (MST), Voltage drop

## 1. Introduction

Access to electricity plays a central role in improving quality of life and economic opportunity. However, many remote and rural areas still lack reliable energy access. Planning electrification in these settings is a complex task that requires balancing technical constraints, infrastructure costs, and social needs. Traditional planning tools often struggle with scalability, making it difficult to adapt quickly to new data or changes in policy.

One of the most detailed planning tools available is the Reference Electrification Model (REM), which simulates the cost-optimal electrification strategy for each individual building. It decides whether each location should be connected to the main grid, served by a mini-grid, or powered through a standalone system. Although REM offers high-resolution results, its computational demands are extremely high. The core of this challenge lies in its clustering algorithm, which evaluates countless groupings of potential customers to find the most efficient network layout. This brute-force approach often requires several days of computation to produce a single output.

This project proposes an improved approach that improves the technical accuracy of REM while drastically reducing the time needed to generate a solution. By redesigning the clustering phase with artificial intelligence techniques and efficient heuristics, this new method enables faster, scalable, and flexible electrification planning.

## 2. Project Definition

The project focuses on reengineering the clustering phase of the REM, which is responsible for grouping buildings into electrification units. The original approach attempts to evaluate all possible combinations of customer groupings and is extremely slow. Instead, this new method begins with each customer as an independent unit and incrementally merges them into larger clusters if doing so leads to a lower overall system cost.

The merging decision is based on three factors: spatial proximity, electrical feasibility, and economic gain. For each proposed merge, the algorithm verifies that the chosen transformer can handle the combined demand and that voltage drop limits are respected

along the cable layout. If these conditions are met and the merge leads to a reduction in total infrastructure cost, it is accepted.

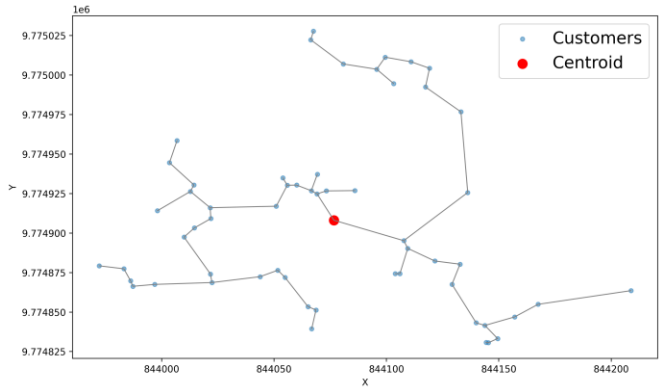
The overall process is iterative: clusters are selected randomly to avoid repetitive patterns, and the algorithm continues merging until no further cost-saving opportunities remain. This local, greedy approach avoids the need to explore every possibility, drastically reducing runtime while maintaining practical feasibility.

## 3. System Description

Each cluster in the system is defined by its customer's energy demand, its geographical center, and the internal network layout that connects all customers to a shared transformer. The layout is constructed using Prim's algorithm, which builds a minimum spanning tree that minimizes total cable length while ensuring every customer is connected. The transformer, located at the root of the layout, distributes power to all customers in the cluster.

The algorithm checks that the transformer can supply the total cluster demand and calculates the voltage drop along the network. It then searches for a cable type that keeps the drop within limits, and among the feasible options, selects the cheapest one to ensure cost-efficiency.

To further optimize the process, the algorithm only considers a small number of nearest neighbors for merging, reducing computational overhead while maintaining quality results. This neighborhood search is performed using spatial indexing, allowing the method to scale to national datasets with hundreds of thousands of customers.




*Electric Grid Solution*

## 4. Results

The algorithm was first validated on a development-scale dataset of 10,000 clients, where it successfully generated technically viable electrification plans while reducing computation times compared to the original REM implementation. This smaller test case allowed for fine-tuning of merging heuristics, voltage drop validation, and cable layout strategies. The real test, however, came with the national-scale dataset containing 250,000 customers.

| Metric | Optimized Clustering Algorithm | Original Solution |
|---|---|---|
| Total cost (USD) | $72.49 million | $69.39 million |
| Connected clients | 250,401 | 249,101 |
| Unconnected clients | 0 | 1,400 |
| Cost per client (USD) | $289.51 | $277.10 |
| Transformers used | 3,222 | 3,176 |
| Most common transformer | 160 kW (2,791 units) | 160 kW (1,787 units) |
| LV_10mm² cable usage (km) | 4,562 (98.9%) | 3,840 (84.0%) |
| Execution time | 3h 17min | 7h 5min |

*Statistics Comparison Versus Dataset 250,000 Customers*

In this scenario, the proposed algorithm was able to cluster and connect every single user, achieving full coverage, unlike the baseline REM model, which left more than 1,400 clients unserved due to feasibility constraints. This comprehensive coverage was achieved while ensuring compliance with transformer loading limits and voltage drop regulations. Furthermore, the infrastructure design showed high levels of standardization: 86.7% of the deployed transformers were 160 kW units (which has the cheapest cost per kW), and almost 99% of the cable length consisted of LV-10mm², the cheapest cable in the catalog.

The total runtime was reduced from approximately seven hours in the original REM to just three hours and seventeen minutes using the new approach, representing a performance improvement of over 50%. This consistency across the system significantly simplifies procurement, logistics, and future maintenance, while also reducing overall network complexity.

## 5. Conclusions

The proposed algorithm delivers significant improvements in both technical feasibility and practical scalability, making it a robust alternative to the original REM clustering approach. Its localized merging strategy, guided by spatial heuristics and electrical constraints, replaces brute-force enumeration with a streamlined process capable of handling hundreds of thousands of customers efficiently.

The results demonstrate that full client coverage can be achieved without sacrificing performance or violating technical thresholds. Moreover, the high standardization in component selection reduces costs and simplifies the rollout of rural electrification infrastructure. By drastically shortening computation times, this method makes high-resolution planning tools like REM suitable for national-scale deployments, enabling fast simulations and agile adaptation to policy or demographic changes.

Looking forward, this work opens the path toward integrating AI models such as encoder-decoder architectures or Graph Neural Networks, which could learn from generated layouts and provide instant configuration recommendations. Such future systems could combine spatial data, electrical simulation, and real-time optimization, moving rural electrification planning into a new era of intelligence and speed.

## 6. References

[1] Akinwale, A. (2022, February 15). Prim algorithm approach to improving local access network in rural areas. ijcte.org. https://www.academia.edu/71618530/Prim_Algorithm_Approach_to_Improving_Local_Access_Netwoe thrk_in_Rural_Areas
[2] Anzola, John & Espada, Jordán & Tarazona, Giovanny & Gonzalez Crespo, Ruben. (2018). A Clustering WSN Routing Protocol Based on k-d Tree Algorithm. Sensors. 18. 1-25. 10.3390/s18092899.
[3] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), 509–517. https://doi.org/10.1145/361002.361007
[4] Ciller, P., Ellman, D., Vergara, C., Gonzalez-Garcia, A., Lee, S. J., Drouin, C., ... & Perez-Arriaga, I. (2019). Optimal electrification planning incorporating on-and off-grid technologies: the Reference Electrification Model (REM). Proceedings of the IEEE, 107(9), 1872-1905.
[5] Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. Nature Reviews Methods Primers, 4(1), 17.
[6] Fredman, M. L., & Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM, 34(3), 596–615. https://doi.org/10.1145/28869.28874
[7] Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4).
[8] González García, A. (2024). A comprehensive decision support framework for the provision of universal access to modern power services in developing countries. PhD thesis dissertation, Universidad Pontificia Comillas, Madrid (Spain)
[9] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50.
[10] LHer, G., Osborne, A., Schweikert, A., Ramstein, C., Stoll, B., & Deinert, M. (2023, October 17). Potential of photovoltaics and energy storage to address lack of electricity access. arXiv.org. https://arxiv.org/abs/2310.11615
[11] Likas , A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2
[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231.
[13] Sahlbert, A., Korkovelos, A., Kabongo, C., Trujillo, C., Khavari, B., & Fuso Nerini, F. (2023). Attention to detail – exploring effects on technology selection in geospatial electrification modelling. https://doi.org/https://doi.org/10.21203/rs.3.rs3043251/v1
[14] Salehi, A., & Davulcu, H. (2019). Graph attention auto-encoders. arXiv preprint arXiv:1905.10715.
[15] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80.

# INTELIGENCIA ARTIFICIAL PARA EL ACCESO UNIVERSAL A LA ENERGÍA

**Autor: Mazuecos Reíllo, Carlos.**
Director: Dueñas Martínez, Pablo.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

**Resumen**

Este proyecto mejora la escalabilidad del Reference Electrification Model (REM) al sustituir su fase de agrupamiento por fuerza bruta por un algoritmo más rápido y técnicamente viable. El nuevo método fusiona iterativamente grupos de clientes utilizando criterios de proximidad espacial y restricciones técnicas, equilibrando coste y viabilidad. La solución se desarrolló sobre un conjunto de datos de 10,000 clientes y se probó en otro de más de 250,000 clientes, logrando mejoras notables en rendimiento. En comparación con la salida original de REM, el nuevo enfoque redujo tanto el coste por cliente como el tiempo total de ejecución.

**Palabras clave:** REM, agrupamiento, PRIM, árbol de recubrimiento mínimo (MST), caída de tensión

## 1. Introducción

El acceso a la electricidad desempeña un papel central en la mejora de la calidad de vida y las oportunidades económicas. Sin embargo, muchas zonas rurales y remotas aún carecen de un acceso fiable a la energía. Planificar la electrificación en estos contextos es una tarea compleja que requiere equilibrar restricciones técnicas, costes de infraestructura y necesidades sociales. Las herramientas tradicionales de planificación suelen tener dificultades de escalabilidad, lo que dificulta su adaptación ágil a nuevos datos o cambios de política.

Una de las herramientas de planificación más detalladas es el Reference Electrification Model (REM), que simula la estrategia de electrificación más rentable para cada edificio individual. Decide si cada ubicación debe conectarse a la red principal, servirse mediante una mini-red o alimentarse a través de un sistema autónomo. Aunque REM ofrece resultados de alta resolución, sus exigencias computacionales son extremadamente elevadas. El núcleo de este desafío radica en su algoritmo de agrupamiento, que evalúa un número enorme de combinaciones posibles de clientes para encontrar la disposición de red más eficiente. Este enfoque por fuerza bruta puede requerir varios días de cálculo para generar una única salida.

Este proyecto propone un enfoque mejorado que aumenta la precisión técnica del REM y reduce drásticamente el tiempo necesario para obtener una solución. Al rediseñar la fase de agrupamiento con técnicas de inteligencia artificial y heurísticas eficientes, este nuevo método permite una planificación de electrificación más rápida, escalable y flexible.

## 2. Definición del Proyecto

El proyecto se centra en rediseñar la fase de agrupamiento del REM, responsable de agrupar edificios en unidades de electrificación. El enfoque original intenta evaluar todas las combinaciones posibles de agrupamientos de clientes y es extremadamente lento. En su lugar, este nuevo método comienza tratando cada cliente como una unidad

independiente y los fusiona progresivamente en grupos más grandes si ello conduce a una reducción del coste total del sistema.

La decisión de fusionar se basa en tres factores: proximidad espacial, viabilidad eléctrica y ganancia económica. Para cada fusión propuesta, el algoritmo verifica que el transformador seleccionado pueda manejar la demanda combinada y que se respeten los límites de caída de tensión a lo largo del trazado del cable. Si estas condiciones se cumplen y la fusión reduce el coste de infraestructura, se acepta.

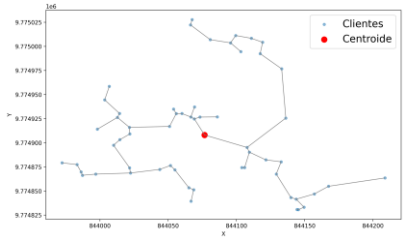
El proceso global es iterativo: los clústeres se seleccionan aleatoriamente para evitar patrones repetitivos, y el algoritmo continúa fusionando hasta que no queden más oportunidades de ahorro de costes. Este enfoque local y codicioso evita la necesidad de explorar todas las posibilidades, reduciendo drásticamente el tiempo de cálculo sin comprometer la viabilidad técnica.

## 3. Descripción del Sistema

Cada clúster en el sistema se define por la demanda energética de sus clientes, su centro geográfico y el trazado interno de red que conecta a todos los clientes con un transformador compartido. El trazado se construye utilizando el algoritmo de Prim, que genera un árbol de recubrimiento mínimo que minimiza la longitud total del cableado y asegura que todos los clientes estén conectados.

El algoritmo comprueba que el transformador pueda abastecer la demanda total del clúster y calcula la caída de tensión a lo largo de la red. Luego busca un tipo de cable que mantenga la caída dentro de los límites y, entre las opciones viables, selecciona el más barato para garantizar la rentabilidad.

Para optimizar aún más el proceso, el algoritmo solo considera un número reducido de vecinos más cercanos al momento de fusionar, lo que reduce la carga computacional manteniendo resultados de alta calidad. Esta búsqueda por vecindad se realiza mediante indexación espacial, lo que permite que el método escale a conjuntos de datos nacionales con cientos de miles de clientes.




*Cluster Layout Using Prim's Algorithm: Displays the minimum spanning tree connecting all customers to the transformer with minimal cable lengt*

## 4. Resultados

El algoritmo fue validado inicialmente en un conjunto de datos de desarrollo con 10,000 clientes, donde logró generar planes de electrificación técnicamente viables y redujo los tiempos de cálculo en comparación con la implementación original del REM. Este caso de prueba más pequeño permitió ajustar las heurísticas de fusión, la validación de la

caída de tensión y las estrategias de trazado de cables. Sin embargo, la verdadera prueba vino con el conjunto de datos a escala nacional, que contenía 250,000 clientes.

| Métrica | Algoritmo de Clustering Optimizado | Solución Original |
|---|---|---|
| Costo total (USD) | $72.49 millones | $69.39 millones |
| Clientes conectados | 250,401 | 249,101 |
| Clientes no conectados | 0 | 1,400 |
| Costo por cliente (USD) | $289.51 | $277.10 |
| Transformadores usados | 3,222 | 3,176 |
| Transformador más común | 160 kW (2,791 unidades) | 160 kW (1,787 unidades) |
| Uso de cable LV_10mm² (km) | 4,562 (98.9%) | 3,840 (84.0%) |
| Tiempo de ejecución | 3h 17min | 7h 5min |

*Comparación entre Solución Original y Propuesta por el Algoritmo*

En este escenario, el algoritmo propuesto logró agrupar y conectar a todos los usuarios, alcanzando una cobertura total, a diferencia del modelo REM original, que dejó sin servicio a más de 1,400 clientes debido a restricciones de viabilidad. Esta cobertura completa se logró cumpliendo con los límites de carga de los transformadores y las regulaciones de caída de tensión. Además, el diseño de infraestructura mostró altos niveles de estandarización: el 86.7% de los transformadores desplegados eran unidades de 160 kW (el coste más barato por kW), y casi el 99% de la longitud de cable utilizada fue LV-10mm², el más barato del catálogo.

El tiempo total de ejecución se redujo de aproximadamente siete horas en el REM original a solo tres horas y diecisiete minutos con el nuevo enfoque, lo que representa una mejora de rendimiento de más del 50%. Esta consistencia en el sistema simplifica notablemente la adquisición de componentes, la logística y el mantenimiento futuro, al tiempo que reduce la complejidad de la red.

## 5. Conclusiones

El algoritmo propuesto aporta mejoras significativas tanto en viabilidad técnica como en escalabilidad práctica, posicionándose como una alternativa robusta al enfoque de agrupamiento original del REM. Su estrategia de fusiones localizadas, guiada por heurísticas espaciales y restricciones eléctricas, sustituye la enumeración por fuerza bruta con un proceso optimizado capaz de manejar cientos de miles de clientes de forma eficiente.

Los resultados demuestran que es posible alcanzar una cobertura total de clientes sin sacrificar el rendimiento ni violar límites técnicos. Además, la alta estandarización en la selección de componentes reduce costes y simplifica el despliegue de infraestructuras de electrificación rural. Al acortar drásticamente los tiempos de cálculo, este método hace que herramientas de planificación de alta resolución como REM sean aptas para implementaciones a escala nacional, permitiendo simulaciones rápidas y adaptación ágil ante cambios demográficos o políticos.

De cara al futuro, este trabajo abre la puerta a la integración de modelos de IA como arquitecturas encoder-decoder o Graph Neural Networks, que podrían aprender de los diseños generados y ofrecer recomendaciones instantáneas de configuración. Estos sistemas futuros podrían combinar datos espaciales, simulación eléctrica y optimización en tiempo real, llevando la planificación de electrificación rural a una nueva era de inteligencia y velocidad.

## 6. Bibliografía


[1] Akinwale, A. (2022, February 15). Prim algorithm approach to improving local access network in rural areas. ijcte.org. https://www.academia.edu/71618530/Prim_Algorithm_Approach_to_Improving_Local_Access_Netwoe thrk_in_Rural_Areas

[2] Anzola, John & Espada, Jordán & Tarazona, Giovanny & Gonzalez Crespo, Ruben. (2018). A Clustering WSN Routing Protocol Based on k-d Tree Algorithm. Sensors. 18. 1-25. 10.3390/s18092899.

[3] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), 509–517. https://doi.org/10.1145/361002.361007

[4] Ciller, P., Ellman, D., Vergara, C., Gonzalez-Garcia, A., Lee, S. J., Drouin, C., ... & Perez-Arriaga, I. (2019). Optimal electrification planning incorporating on-and off-grid technologies: the Reference Electrification Model (REM). Proceedings of the IEEE, 107(9), 1872-1905.

[5] Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. Nature Reviews Methods Primers, 4(1), 17.

[6] Fredman, M. L., & Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM, 34(3), 596–615. https://doi.org/10.1145/28869.28874

[7] Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4).

[8] González García, A. (2024). A comprehensive decision support framework for the provision of universal access to modern power services in developing countries. PhD thesis dissertation, Universidad Pontificia Comillas, Madrid (Spain)

[9] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50.

[10] LHer, G., Osborne, A., Schweikert, A., Ramstein, C., Stoll, B., & Deinert, M. (2023, October 17). Potential of photovoltaics and energy storage to address lack of electricity access. arXiv.org. https://arxiv.org/abs/2310.11615

[11] Likas , A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231.

[13] Sahlbert, A., Korkovelos, A., Kabongo, C., Trujillo, C., Khavari, B., & Fuso Nerini, F. (2023). Attention to detail – exploring effects on technology selection in geospatial electrification modelling. https://doi.org/https://doi.org/10.21203/rs.3.rs3043251/v1

[14] Salehi, A., & Davulcu, H. (2019). Graph attention auto-encoders. arXiv preprint arXiv:1905.10715.

[15] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80.

# *Table of Contents*

# *List of figures*

# *List of tables*

# 1. INTRODUCTION

## 1.1. CONTEXT AND MOTIVATION

Electricity access is essential for development, as it supports basic services like healthcare, education, and communication, and enables economic activities such as lighting, refrigeration, and digital tools. Without it, communities face serious limitations and miss many opportunities to improve their quality of life (LHer et al. 2023).

Despite progress in recent decades, rural and remote areas often remain unelectrified due to isolation, low population density, and limited infrastructure. Electrifying these regions involves complex trade-offs between cost, technical feasibility, and diverse user needs, while also requiring adaptability to changes in demand or policy.

The Reference Electrification Model (REM) is a detailed planning tool that helps determine the least-cost option for each user—grid extension, mini-grid, or standalone system—based on spatial and demand data. However, its detailed clustering process is computationally intensive and can take several days or even weeks for large datasets, limiting its responsiveness.

Improving REM's performance, particularly in the clustering phase which is the bottleneck of the computational time of the algorithm, is key to enabling faster and more flexible planning. This phase is based on a brute-force approach that evaluates all possible customer groupings, making it extremely time-consuming. A more efficient model would support real-time decision-making, scenario testing, and broader stakeholder engagement. These improvements directly contribute to the achievement of several Sustainable Development Goals, especially SDG 7 (affordable and clean energy), SDG 1 (no poverty), and SDG 10 (reduced inequalities), by accelerating access to modern energy in underserved areas.

## 1.2. OBJECTIVES

The main objective of this project is to improve the efficiency and scalability of the REM clustering phase using artificial intelligence (AI) techniques. This involves designing and testing new clustering strategies that can reduce execution time while preserving or improving the quality of the resulting electrification plans.

The work focuses on three specific goals: first, to implement and compare different clustering algorithms; second, to define and evaluate meaningful metrics that assess the quality of these clustering in both technical and economic terms; and third, to apply the improved approach to a large-scale dataset not used during training, in order to test the model's generalization ability. This final step is essential to validate the robustness of the proposed solution in real-world settings.

## 1.3. PLANNING AND ECONOMIC FEASIBILITY

The project was implemented using open-source tools and executed on the Instituto de Investigación Tecnológica (IIT) high-performance computing infrastructure at Universidad Pontificia Comillas, ensuring low development cost and high scalability. Testing was structured in two phases: an initial prototype with 10,000 customers to verify performance and tune parameters, followed by deployment on a national dataset with over 250,000 customers.

The proposed method significantly reduces runtime by replacing REM's brute-force logic with spatially constrained heuristics and local merging decisions. This improves responsiveness and lowers the computational burden, enabling broader use of REM in real-time or iterative planning contexts. Moreover, the final output achieves a lower cost per client and per kW than the original configuration, confirming the practical viability of the approach in economic terms.

## 1.4. STRUCTURE OF THE THESIS

This thesis is structured as follows. Section 2 reviews the state of the art, focusing on the main characteristics and limitations of the REM and its

clustering component within the broader context of rural electrification tools. Section 3 presents the proposed methodology, including the technical details of the algorithm. Section 4 describes the experimental framework, covering the datasets employed, the configuration of the simulations, and the specific testing phases. Section 5 discusses the results, comparing different configurations in terms of runtime, infrastructure cost, and network design. Finally, Section 6 presents the main conclusions and outlines future research directions that build on the findings of this work.

## 2. STATE OF THE ART

The REM is the most detailed and technically rigorous tool for electrification planning at the building level (Ciller et al., 2019). Its key innovation is that it models every individual household or facility and determines the least-cost way to electrify it, whether through grid extension, a local mini-grid, or a standalone system. At the heart of REM lies its clustering phase, which is responsible for grouping nearby customers into electrification units. This phase is where most of REM's computational burden occurs.

REM uses a fully brute-force approach: it explicitly evaluates all possible combinations of customer groupings to identify the configuration that minimizes total cost. For each possible cluster, it recalculates infrastructure costs, including cable lengths, transformer selection, and load distribution. It does not rely on approximations, learned rules, or pre-filtering. Every possible merge is tested, and every potential configuration is compared. This means that for even a modest number of buildings, the number of combinations explodes, making the process extremely time-consuming. The algorithm's merging logic is greedy, but each decision is backed by a full recalculation of technical and economic feasibility, meaning that no shortcuts are taken. While this brute-force nature ensures optimality in many local decisions, it makes REM fundamentally hard to scale (González, 2024).

In contrast, geospatial tools such as OnSSET adopt more scalable, high-level approaches. OnSSET uses settlement-level demand data combined with heuristic rules and regression-based cost models to choose between grid extension, mini-grids, or standalone solar systems. This makes it suitable for national-scale scenario analysis, though with significantly less technical detail. As discussed in recent studies, OnSSET's modeling outcomes are highly sensitive to user-defined parameters and input assumptions, which can substantially affect the resulting technology mix (Sahlbert et al. 2023). Its speed and simplicity come at the cost of resolution and electrical accuracy, especially when compared to building-level models like REM.

## 3. METHODOLOGY

This algorithm addresses the challenge of the REM of planning rural electrification systems by proposing an iterative process that combines spatial clustering, cost minimization, and technical feasibility. The main objective is to group nearby customers into electrification units that can be connected to a single transformer, while minimizing infrastructure cost and ensuring that the resulting design complies with electrical and operational constraints.

The process begins by treating each customer as an independent unit, each equipped with its own transformer. Each of these units is defined by its location (*Coordinate X* and *Coordinate Y*), demand (*kW*), and the corresponding infrastructure costs (*$*). From this initial state, the algorithm enters a loop where it searches for opportunities to merge small clusters into larger ones if doing so reduces the overall system cost. Merging decisions are based on cost-benefit evaluations that consider the sum of cable expenses, transformer costs, and technical constraints like allowable voltage drop.

When merging clusters, the algorithm first checks whether there exists a transformer capable of handling the combined demand, in kW, of all customers within the proposed cluster. Only if a suitable transformer is available for the combined

demand does the algorithm proceed to evaluate the internal layout and technical feasibility of the new grouping.

If a transformer meets the requirements, the algorithm determines the optimal location for the shared transformer. This location is computed as the demand-weighted centroid of the customers' coordinates, specifically, the average of the *Coordinate X* and *Coordinate Y* positions, each weighted by the respective customer's *kW* demand.

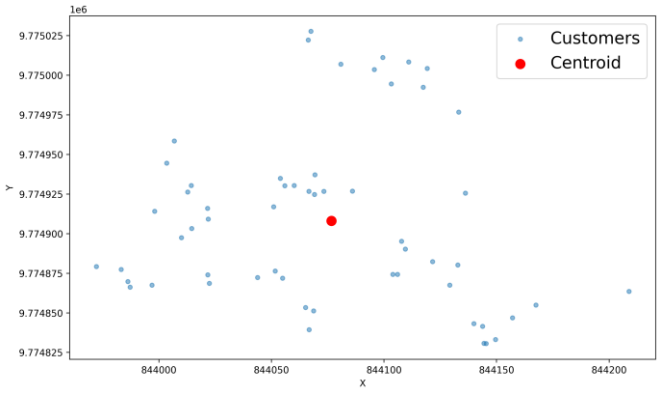



*Figure 1 - Transformer Location via Demand-Weighted Centroid: Shows how the transformer's position is calculated based on the weighted average of customer coordinates using their power demand*

Then, the method continues with the construction of the network layout within each cluster. To approximate how customers would be physically connected to the transformer, the algorithm generates a minimum spanning tree (MST) using Prim algorithm. This ensures that the internal layout of each cluster is realistic, reflecting the most efficient way to lay out cables between the transformer and the connected customers. The figure below illustrates a typical layout produced by the algorithm.

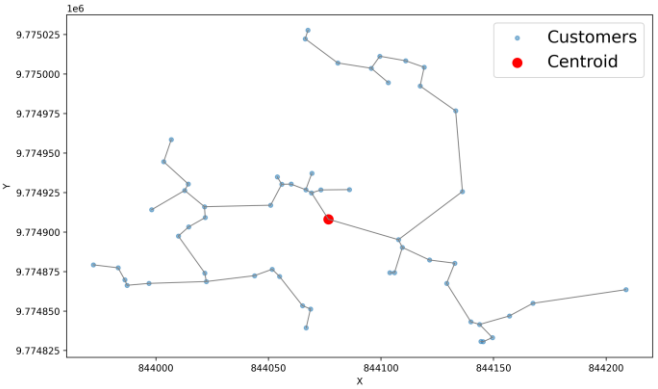



*Figure 2 - Cluster Layout Using Prim's Algorithm: Displays the minimum spanning tree connecting all customers to the transformer with minimal cable length*

Then, the technical feasibility of the configuration is assessed based on voltage drop constraints. The voltage drop along the feeder is calculated segment-by-segment, considering that loads are not evenly distributed but instead connected at discrete points. For each segment $i$, the current $I_i$ flowing through it is computed as a function of the total downstream demand. This current is then used to estimate the voltage drop:

$$\Delta V_i = I_i \cdot (r' \cdot \cos\varphi + x' \cdot \sin\varphi) \cdot \Delta L_i$$

Where $r'$ and $x'$ are the resistance and reactance per kilometer of the cable, $\varphi$ is the power factor angle ($\cos\varphi$ is 0.9) and $\Delta L_i$ is the length of the segment in kilometers. Summing the contributions of all segments gives the total voltage drop from the transformer to the furthest customer:

$$V_{total} = \sum_{i=1}^{n} \Delta V_i$$

The result is then normalized by the nominal system voltage (428 V in low-voltage systems) and compared against the maximum allowable drop:

$$\frac{V_{total}}{V_{nom}} \leq \text{maximum allowable drop}$$

The result is then normalized by the nominal system voltage and compared against a maximum allowable voltage drop. This threshold was not

predefined but instead inferred by analyzing the original solution, where the calculated voltage drops suggested that the system was operating within a range of over- and under-voltage of 7%. This value was therefore adopted as the benchmark for evaluating the feasibility of new configurations.

At this point, the algorithm evaluates the available cable types. From a catalog of candidates, it filters those that satisfy the voltage drop constraint. Among the viable options, the algorithm selects the cable with the lowest cost per kilometer. This cost-efficiency often comes with a trade-off in electrical characteristics: cables with lower cost per unit length generally have higher resistance and reactance values, which can increase voltage drop. Conversely, cables with lower $r'$ and $x'$ values tend to be more expensive but are better suited for longer or higher-load segments due to their reduced impedance. The algorithm thus balances technical constraints and economic optimization by selecting the cheapest cable that still guarantees compliance with the voltage drop requirement.

The algorithm begins by selecting a random cluster and identifying the $n$ nearest neighbors using a K-D tree (Anzola et al., 2018; Bentley, J. L., 1975), a spatial indexing structure optimized for fast nearest-neighbor retrieval in multidimensional space. This approach avoids brute-force distance calculations and significantly improves efficiency when operating on large-scale datasets. Once the neighboring clusters are found, the algorithm generates all possible combinations among them and evaluates each potential merge.

Each grouping is assessed based on two criteria: technical feasibility—ensuring transformer capacity and voltage drop limits are not exceeded—and overall infrastructure cost, which includes transformer and cable expenses. If a merge results in a lower total cost while satisfying all technical constraints, it is accepted and the clustering configuration is updated.

Empirical results showed that setting $n = 4$ yielded the best trade-off between computational efficiency and merge quality. This choice minimized execution time at scale while consistently producing lower-cost network designs, thanks to more effective cluster combinations.

This process is repeated iteratively. The clusters are processed in randomized order to prevent repetitive behavior and improve convergence. The algorithm stops when no further merges can be found that reduce the total cost of the system.

### *3.1. PRIM*

Prim's algorithm (Akinwale, 2022) is a well-established method for constructing a MST, which in the context of electrification planning serves to connect a central transformer to all surrounding customer locations using the shortest total length of cable. By assuming that cable cost is proportional to Euclidean distance, the algorithm produces a network layout that minimizes infrastructure expenses while ensuring every customer is connected.

The process starts by treating the transformer as the root of the network and iteratively grows the tree by identifying and connecting the closest customer location that has not yet been incorporated. At every iteration, the algorithm maintains a distinction between locations that are already part of the tree and those that are still pending. For each unconnected customer, it evaluates the shortest connection available to any node already in the network. The customer with the smallest such distance is selected, and a direct connection is established. This step-by-step, greedy approach guarantees that the tree expands in the most cost-efficient way at every stage.

Throughout this process, a record is kept of how each customer is connected to the rest of the network. Each point is assigned a unique *parent* node—the one it links to directly—resulting in a clear hierarchical structure rooted at the transformer. This tree structure ensures that the network is continuous and acyclic, which simplifies both the design and later maintenance.

To improve computational efficiency, the algorithm is implemented with a heap-based priority queue that keeps track of the current best candidate for expansion (Fredman & Tarjan, 1987). This reduces the cost of selecting the next node to $O(\log n)$ per operation, compared to $O(n)$ in a naïve implementation. However, this per-operation gain does not change the overall complexity in dense graphs. When every node can connect to nearly every other (as in Euclidean distance graphs), the number of potential edge updates still grows as $O(n^2)$, since each of the $n$ nodes may require distance updates involving up to $n$ neighbors.

Thus, even with a heap, the total runtime remains $O(n^2 \log n)$, but the practical performance is substantially improved. The heap reduces the number of full scans over the remaining nodes and makes local decisions more efficient. This optimization is crucial for handling large-scale problems involving hundreds or thousands of customers, where naïve approaches become computationally intractable.

The MST constructed through Prim's algorithm is especially valuable in applications that require clear and traceable network hierarchies. Since each node has a unique path back to the transformer, the structure supports efficient calculations of total cable requirements, cost estimates, and even technical assessments like load flow or voltage drop simulations. Moreover, it enables planners to simulate electricity distribution from the transformer outward and to visualize how demand aggregates as one moves up the tree. This makes the algorithm particularly useful for rural electrification projects, where infrastructure must be both low-cost and reliable across challenging geographic areas.

### *3.1.1. OTHER APPROACHES*

Other layout strategies were also evaluated to compare their performance and practicality in rural electrification contexts. The main alternatives explored were:

- Pure star: every customer is connected straight back to the transformer; it is trivial to generate and easy to phase in the field, but total trench length grows almost linearly with the number of customers. That translates into higher capital cost and no route redundancy.
- Kruskal: In general, Prim is faster than Kruskal when the graph is dense. Kruskal needs to sort all edges first, which takes $O(E \log E)$ time (Kruskal, 1956). In fully connected graphs, where the number of edges $E$ is approximately $n^2$, this sorting phase dominates the runtime, growing as $O(n^2 \log n)$. Once sorted, Kruskal constructs the MST by iteratively adding the next shortest edge that does not create a cycle, using a disjoint-set data structure to manage connectivity.
  Although Kruskal produces a valid MST, it does not define a root node by default. In this project, a post-processing step was added to reconstruct a transformer-centered tree structure by traversing the MST and assigning a parent to each node based on proximity to the transformer. While this makes Kruskal usable for electrification modeling, it introduces extra computational overhead and complexity.
  Empirically, Prim consistently achieves faster execution times than Kruskal in dense spatial networks. Its ability to avoid global edge sorting and to exploit localized expansions through a heap structure makes it significantly more scalable in practice. The graph below illustrates the runtime differences between Prim and Kruskal. As the number of evaluated neighbors increases, and consequently more calls to the algorithms are triggered, the total computation time grows accordingly. Prim consistently outperforms Kruskal, particularly at scale.

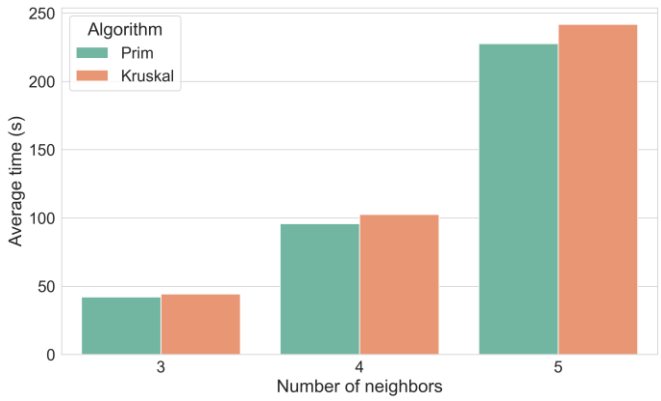



*Figure 3 - Runtime Comparison, Prim vs Kruskal: Compares the execution times of both algorithms for cluster layout generation, highlighting Prim's superior performance on dense graphs.*

### 3.2. OTHER TESTED CLUSTERINGS

During the development of the clustering component for rural electrification planning, various algorithmic strategies were explored in an attempt to replace the computationally intensive brute-force method used in the REM. These early efforts focused on adapting standard unsupervised learning algorithms to the task of grouping customers in a way that was both spatially coherent and technically feasible, particularly with regard to transformer capacity constraints and network layout requirements.

The first method evaluated was DBSCAN (Density-Based Spatial Clustering of Applications with Noise), selected for its ability to identify clusters of arbitrary shape without requiring the number of clusters to be predefined (Ester et al., 1996). This density-based approach initially appeared well-suited for rural environments, where settlement patterns are often irregular and do not conform to uniform spatial distributions. The algorithm's key parameters—$\varepsilon$ (the maximum distance between two points to be considered neighbors) and *min_samples* (the minimum number of points required to form a dense region)—offered some control over clustering behavior. However, in practice, DBSCAN proved difficult to calibrate across diverse geographic conditions. In densely populated rural centers, small $\varepsilon$ values produced reasonable results, but in sparser areas, much larger values were needed to capture any meaningful structure. Increasing $\varepsilon$, however, significantly inflated computational costs, as the algorithm had to compute and compare more pairwise distances. Despite efforts to improve performance through a custom Python implementation and optimization of the region query step using spatial indexing, the clustering output remained unstable. A large number of customers were classified as noise, and cluster boundaries shifted unpredictably across runs. Ultimately, DBSCAN's sensitivity to parameter choice and its high runtime made it unsuitable for reliable use at scale.

K-Means clustering was the second technique tested, selected for its computational efficiency and simplicity (Likas et al., 2003). The number of clusters was estimated by dividing the total demand of all customers by the capacity of the largest available transformer, which provided a rough upper bound. The algorithm then attempted to minimize intra-cluster variance by assigning customers to the nearest centroid and iteratively updating the centroids until convergence. Despite its speed, K-Means suffered from core limitations in this context. The algorithm assumes that clusters are convex and roughly spherical in Euclidean space—an assumption that does not hold in rural environments shaped by terrain, road infrastructure, and spatial constraints. Moreover, K-Means is fundamentally a top-down algorithm, beginning with abstract centroids that may not correspond to physically or electrically meaningful positions. In multiple trials, K-Means created clusters that exceeded transformer capacities or violated network layout constraints. As a result, even though the method completed quickly, its outputs were often technically infeasible and economically suboptimal.

A third method, Fuzzy C-Means (Ghosh & Dubey, 2013), was briefly explored due to its ability to assign soft cluster memberships. This approach assigns each data point a degree of belonging to each cluster rather than forcing a binary assignment. Initially, this flexibility seemed promising, especially for areas where customer locations fell between multiple potential service zones. However,

the soft assignments conflicted with the hard requirements of electrification infrastructure. In this context, each customer must be connected to exactly one transformer, and fractional memberships have no physical or operational meaning. Attempts to convert the soft outputs into hard clusters—typically by assigning each point to the cluster with the highest membership degree—led to uneven and unbalanced groupings. Some clusters exceeded the transformer's capacity limits, while others failed to meet voltage drop constraints. In addition, the increased computational complexity introduced by maintaining membership degrees did not provide sufficient benefits to justify its use.

# 4. EXPERIMENTS

## *4.1. OBJETIVES*

The objective of the experiments is to assess the effectiveness of the clustering algorithm developed as part of this project. The experiments are intended to evaluate three main aspects: technical feasibility, cost efficiency and computational performance.

The first goal is to determine whether the proposed method can generate valid cluster groupings that comply with electrical design constraints, particularly transformer capacity limits and voltage drop thresholds. Ensuring that each customer group is served by a suitable transformer and connected through feasible cable layouts is critical to guarantee that the resulting electrification plan could be implemented in practice.

The second goal is to measure the extent to which the algorithm reduces overall infrastructure costs. These costs include both the cost of the transformers and of the cable required to connect all users within each cluster. A well-performing solution should strike a balance between minimizing the number of transformers deployed and controlling the total cable length, which directly affects the project's financial viability.

Finally, the third goal is to analyze the scalability and runtime behavior of the algorithm. By testing different configurations and input sizes, the experiments provide insight into how quickly the algorithm converges, how it responds to changes in parameters and whether it remains efficient when applied to large-scale national datasets. This aspect is particularly relevant for future real-world applications, where electrification planning tools must operate on datasets with hundreds of thousands of users.

## *4.2. DATASETS*

The experiments were conducted using data from real rural electrification scenarios, more specifically, used in Rwanda. Each customer in the dataset is defined by two spatial coordinates, *Coordinate X* and *Coordinate Y* and a power demand expressed in *kW;* i.e., considering different types of consumers. These points are distributed over rural terrain with varying density patterns, including compact village centers as well as isolated households. This diversity is essential to test the algorithm's ability to adapt its clustering logic to different spatial conditions.

Two datasets were used in the experimental phase. The first one is a small instance of approximately ten thousand customers. This version served as a development benchmark to test the convergence and correctness of the merging algorithm. It enabled early debugging and the calibration of algorithmic parameters such as the neighborhood size and the stopping condition. Once the algorithm was validated at small scale, the second dataset—a full-scale national simulation containing more than 250,000 users—was employed. This large input allowed for robust testing of the algorithm's runtime performance, scalability, and cost-effectiveness under real-world conditions.

The transformers catalog used in the simulations consists of a series of standard transformer sizes, each with an associated cost and kW distribution capacity. The algorithm must choose, for each cluster, the smallest transformer that can safely serve the total demand of the group without violating capacity constraints.

Similarly, the cable catalog includes various options characterized by their electrical resistance, reactance, and cost per kilometer. These values determine not only whether a given cable can be used in a particular cluster, but also how much the connection will cost and what the voltage drop will be. The cable selected must meet both the thermal current requirement and the voltage constraint defined for the system.

### *4.3. CONFIGURATION*

All simulations were executed on a high-performance server from the Instituto de Investigación Tecnológica (IIT), equipped with two processors Intel(R) Xeon(R) Silver 4314, each offering 16 cores at 2.40 GHz, and 256 GB of RAM. During the experiments, the system allocated 8 physical cores and approximately 94 GB of RAM to the Python process, as observed through system monitoring tools. The operating system was Windows Server, which manages resource distribution dynamically depending on concurrent usage. The algorithm was implemented and run using Python 3.12.2, with core packages including NumPy, pandas and SciPy.

## 5. RESULTS

The experimental evaluation is structured in two phases, corresponding to two datasets of increasing complexity and scale: one with 10,000 customers and another with 250,000 customers. This separation allows for a detailed analysis of the algorithm's scalability, cost efficiency, and technical feasibility across different operational contexts.

### *5.1. EVALUATION ON 10,000-CUSTOMER DATASET*

The first objective of the experimental phase was to ensure that the configurations generated by the algorithm were technically feasible. As a starting point, the original REM-based solutions were carefully analyzed to establish a baseline for comparison. This analysis revealed a pair of conditions that compromised their validity. First, some transformers were assigned a total demand that exceeded their rated capacity, making the configurations infeasible from an electrical engineering standpoint. Second, inconsistencies were found in the cable layout, where different types of cables were used along the same distribution line. Such mixing of cable types leads to uneven electrical performance and complicates both installation and maintenance, ultimately increasing operational risk and cost. Due to these issues, only the technically valid portion of the original solution was retained for comparison. This allowed for a fair evaluation of the proposed algorithm, focusing on both its feasibility and its potential to improve cost and scalability, as detailed in the table below where a proposed solution is compared to only a 42% of the whole solution using the 10,000 customers dataset.

| Metric | Optimized Clustering Algorithm | Original Solution |
|---|---|---|
| Number of clusters | 111 | 71 |
| Total number of clients | 10,960 | 4,635 |
| Clients per cluster | 98.74 | 65.28 |
| Total network cost ($) | 2,865,724.13 | 1,640,929.54 |
| Average cost per cluster ($) | 25,817.33 | 23,111.68 |
| Average cost per client ($) | 261.47 | 354.03 |
| Total demand (kW) | 15,675.30 | 7,470.80 |
| Cost per kW ($/kW) | 182.82 | 219.65 |

*Table 1 - General Statistics Comparison Versus Dataset 10,000 Customers*

Table 1 highlights the overall efficiency of the proposed solution in terms of both scale and cost, revealing a network architecture that is not only more expansive but also significantly more economical on a per-unit basis. Although the proposed solution entails a higher total investment in absolute terms, this is justified by its ability to serve a considerably larger user base. Specifically, it supports more than twice the number of clients compared to the original solution and delivers more than double the total installed power capacity in kilowatts. This broader coverage allows the network to spread infrastructure costs over a much larger set of beneficiaries, thereby achieving notable economies of scale.

As a result, the average cost per client in the proposed solution is markedly lower— around $261 versus $354—despite the higher total expenditure.

This reduction demonstrates a more efficient allocation of resources. Similarly, the cost per kilowatt installed also drops significantly, from $219.65 in the original solution to $182.82 in the proposed one.

Beyond cost metrics, the structural characteristics of the network further underscore its superior design. The proposed solution not only increases the total number of clusters but also achieves a higher client density within each cluster. With an average of 98.74 clients per cluster, compared to 65.28 in the original layout, each grouping serves a larger concentration of demand. This increased density is particularly advantageous from an engineering and operational standpoint. It means that each transformer, cable segment, and associated component is used more intensively, enhancing the utilization rate of capital equipment.

Higher client density per cluster translates into reduced per-client infrastructure requirements. Transformers operate closer to their optimal capacity, minimizing underutilization, while cable lengths and distribution paths can be optimized to serve more endpoints within a compact area. This reduction in redundancy and distribution overhead contributes to better load balancing and system efficiency. Moreover, a well-concentrated layout simplifies maintenance and monitoring, as infrastructure is centralized rather than dispersed.

| Transformer Type | Optimized Clustering Algorithm | Original Solution |
|---|---|---|
| 50 kW | 1 | 7 |
| 100 kW | 7 | 23 |
| 160 kW | 103 | 41 |
| **Total** | 111 | 71 |

*Table 2 - Transformers Comparison Versus Dataset 10,000 Customers*

The transformer distribution highlights one of the key advantages of the algorithm: it strongly favors the use of a single transformer type, resulting in a more scalable and cost-effective network configuration. Specifically, the proposed solution selects 160 kW units in 92.7% of cases (103 out of 111), compared to only 57.7% (41 out of 71) in the original solution. This high degree of standardization reduces complexity during procurement and simplifies both deployment and maintenance across rural regions.

The 160 kW transformer is also the most economical per unit of capacity, with a cost of $75 per kW, versus $104 for 100 kW units and $138 for 50 kW units. By concentrating usage on the most cost-efficient option, the Optimized Clustering Algorithm lowers the average cost per installed kilowatt and avoids the fragmentation seen in the original REM-based design. The result is a more homogeneous infrastructure that is easier to scale and replicate in national electrification efforts.

| Cable Type | Optimized Clustering Algorithm | Original Solution |
|---|---|---|
| LV_10mm2 | 110 | 13 |
| LV_25mm2 | 1 | 49 |
| LV_50mm2 | 0 | 6 |
| LV_70mm2 | 0 | 2 |
| LV_95mm2 | 0 | 1 |

*Table 3 - Cables Comparison Versus Dataset 10,000 Customers*

The proposed solution clearly outperforms the original configuration by standardizing almost entirely on the use of LV_10mm² cable, the cheapest option in the catalog at approximately $2.35 per meter. Out of a total of 111 cable segments, 110 (99.1%) use LV_10mm², compared to only 13 out of 71 (18.3%) in the original design. This strong standardization not only lowers costs but also simplifies procurement and implementation. In contrast, the original solution relies heavily on thicker and more expensive cables: 49 segments (69.0%) are LV_25mm², while the remaining segments are spread across LV_50mm², LV_70mm², and LV_95mm², which can reach prices up to $14.51 per meter. This diversification, although potentially beneficial in addressing localized electrical constraints, results in a significantly higher overall cost for the network infrastructure.

What makes the proposed design particularly efficient is its ability to maintain adequate technical performance while relying almost exclusively on the thinnest and least expensive cable. This implies a high degree of optimization in the layout, cluster sizing, and load distribution, effectively eliminating

the need for thicker conductors. The minimal use of LV_25mm² cable—only one segment (0.9%)—and the complete absence of higher cable sizes not only simplify procurement and reduce material costs but also streamline installation and maintenance processes. Ultimately, the proposed solution ensures compliance with voltage drop and thermal limits using minimal resources, yielding a substantially lower total cabling expenditure and demonstrating strong scalability for large-scale rural electrification.

### 5.2. EVALUATION ON 250,000-CUSTOMER DATASET

To assess the algorithm's scalability and robustness under realistic, large-scale conditions, the second phase of the evaluation was conducted using a national dataset of over 250,000 customers. The comparative results are summarized below.

| Metric | Optimized Clustering Algorithm | Original Solution |
|---|---|---|
| Total cost (USD) | $72.49 million | $69.39 million |
| Connected clients | 250,401 | 249,101 |
| Unconnected clients | 0 | 1,400 |
| Cost per client (USD) | $289.51 | $277.10 |
| Transformers used | 3,222 | 3,176 |
| Most common transformer | 160 kW (2,791 units) | 160 kW (1,787 units) |
| LV_10mm² cable usage (km) | 4,562 (98.9%) | 3,840 (84.0%) |
| Execution time | 3h 17min | 7h 5min |

*Table 4 - Statistics Comparison Versus Dataset 250,000 Customers*

Although the original solution appears slightly less costly in absolute terms—$69.39 million versus $72.49 million—this difference is largely explained by a critical omission: 1,400 customers remain unconnected. These clients were excluded under the assumption that they would be electrified separately, likely due to their geographic dispersion or high connection cost. However, internal estimates indicate that reaching these users already requires at least $1.4 million in additional infrastructure. Therefore, if these customers were to be connected, the total cost of the original solution would exceed the reported figure, eliminating its apparent cost advantage and positioning the proposed configuration as more comprehensive and economically competitive.

In contrast, the proposed method connects the full set of 250,401 clients, achieving universal coverage without compromising technical feasibility. It does so with a streamlined infrastructure: 2,791 of the 3,222 (86.7%) transformers deployed are 160 kW units—the most cost-efficient in the catalog—and nearly 99% of the total cable length uses LV-10mm², the cheapest available option. This high degree of standardization simplifies logistics, reduces procurement and installation effort, and minimizes complexity during operation and maintenance. The original design, on the other hand, uses a more diverse set of transformer sizes and cable types, including a substantial share of thicker and more expensive conductors.

The difference in computational efficiency is also substantial. The original REM-based solution required approximately 7 hours and 5 mins to generate the full network layout, following a brute-force evaluation of cluster combinations. In contrast, the proposed algorithm, based on localized merging and spatial heuristics, completes the same task in just 3 hours and 17 minutes. This represents more than a 50% reduction in runtime, making the method far more practical for large-scale deployments, policy scenario testing, or iterative design processes.

## 6. CONCLUSIONS AND FUTURE WORK

This thesis demonstrates that the application of artificial intelligence techniques can substantially improve the scalability, cost-efficiency, and practical usability of high-resolution electrification planning tools. By reengineering the clustering phase of the REM, this project replaces its brute-force, computationally intensive logic with an iterative, spatially guided merging algorithm that respects electrical and economic constraints. The new method successfully improves the technical precision of the REM while reducing runtime and simplifying infrastructure design.

The algorithm was tested on two levels: a development-scale dataset of approximately 10,000

users and a full-scale national simulation with 250,401 customers. In both cases, it generated technically valid electrification plans that fully complied with transformer loading limits and voltage drop thresholds.

The proposed algorithm demonstrated strong performance improvements in layout compactness, cluster sizing, and economic efficiency, consistently producing feasible networks while significantly reducing cost compared to REM's original clustering method. At large scale, the algorithm delivered substantial improvements in computational performance and scalability, reducing total runtime by over 50%. This efficiency gain makes the method practical for use in large-scale electrification efforts, real-time scenario testing, and iterative policy planning. The resulting network also displayed a high degree of standardization in component selection, which facilitates procurement, installation, and long-term maintenance.

The improvements observed in this work are not limited to cost and coverage. By eliminating the need for brute-force enumeration of customer groupings and replacing it with local decisions based on spatial heuristics, the algorithm opens the door to much faster and more scalable electrification planning. This methodological shift allows detailed, building-level models like the REM to be applied at national scale in a fraction of the time.

One of the most promising directions for future research in rural electrification system design is the integration of models capable of learning optimal connection configurations while accounting for technical and economic constraints. In particular, developing neural networks that can infer feasible and cost-effective layouts based on past data could transform the current heuristic-based process into a more adaptive and intelligent framework.

The algorithm proposed in this project relies on a greedy, rule-based approach that iteratively merges customer clusters based on cost-benefit evaluations and electrical feasibility. While effective, this strategy follows a fixed decision path and lacks the flexibility to generalize or adapt rapidly to new scenarios. A promising model architecture would be an encoder-decoder neural network, where the encoder processes geospatial and demand data from the input region and the decoder generates a proposed network configuration that adheres to the constraints (Salehi & Davulcu, 2019). Another compelling direction is the use of Graph Neural Networks, which are especially well-suited for modeling electrical grids (Scarselli et. al, 2008; Corso et. al, 2024).

In summary, this work confirms that artificial intelligence can serve as a catalyst for transforming how rural electrification systems are planned, designed, and optimized. By maintaining the technical depth of the REM while overcoming its computational limitations, the proposed algorithm offers a concrete, feasible, scalable, and economically pathway toward universal energy access. It lays the foundation for future models that are not only more intelligent and adaptive, but also capable of supporting real-time decision-making across entire regions.

## 7. REFERENCES

[1] Akinwale, A. (2022, February 15). Prim algorithm approach to improving local access network in rural areas. ijcte.org. https://www.academia.edu/71618530/Prim_Algorithm_Approach_to_Improving_Local_Access_Netwoe thrk_in_Rural_Areas

[2] Anzola, John & Espada, Jordán & Tarazona, Giovanny & Gonzalez Crespo, Ruben. (2018). A Clustering WSN Routing Protocol Based on k-d Tree Algorithm. Sensors. 18. 1-25. 10.3390/s18092899.

[3] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), 509–517. https://doi.org/10.1145/361002.361007

[4] Ciller, P., Ellman, D., Vergara, C., Gonzalez-Garcia, A., Lee, S. J., Drouin, C., ... & Perez-Arriaga, I. (2019). Optimal electrification planning incorporating on-and off-grid technologies: the Reference Electrification

Model (REM). Proceedings of the IEEE, 107(9), 1872-1905.

[5] Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. Nature Reviews Methods Primers, 4(1), 17.

[6] Fredman, M. L., & Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM, 34(3), 596–615. https://doi.org/10.1145/28869.28874

[7] Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4).

[8] González García, A. (2024). A comprehensive decision support framework for the provision of universal access to modern power services in developing countries. PhD thesis dissertation, Universidad Pontificia Comillas, Madrid (Spain)

[9] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50.

[10] LHer, G., Osborne, A., Schweikert, A., Ramstein, C., Stoll, B., & Deinert, M. (2023, October 17). Potential of photovoltaics and energy storage to address lack of electricity access. arXiv.org. https://arxiv.org/abs/2310.11615

[11] Likas , A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231.

[13] Sahlbert, A., Korkovelos, A., Kabongo, C., Trujillo, C., Khavari, B., & Fuso Nerini, F. (2023). Attention to detail – exploring effects on technology selection in geospatial electrification modelling. https://doi.org/https://doi.org/10.21203/rs.3.rs3043251/v1

[14] Salehi, A., & Davulcu, H. (2019). Graph attention auto-encoders. arXiv preprint arXiv:1905.10715.

[15] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80.