POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Mapping Electricity Access Disparities in Mozambique: An AI-Driven Satellite Imagery Approach

TESI DI LAUREA MAGISTRALE IN
ENERGY ENGINEERING
INGEGNERIA ENERGETICA

Author: **Francesco Paolo Antinori**

Student ID: 10983291
Advisor: Manuel Heitor
Advisor: Nicolò Stevanato
Academic Year: 2025-26

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Manuel Heitor at Instituto Superior Técnico for his invaluable advice and guidance throughout this research. I am also deeply grateful to him for welcoming me into the IN+ Center for Innovation, Technology and Policy Research, where I had the opportunity to grow both as a researcher and as a person.

I would also like to thank my co-supervisor Professor Nicolò Stevanato and Alessandro Onori at Politecnico di Milano for their constant availability and support, and for the methodological rigour they brought to every discussion.

Abstract

This study develops an artificial intelligence and high-resolution satellite imagery approach to map and analyze electricity access disparities in Mozambique over the period 2015–2024.

The main research question is: what factors explain electricity access disparities in Mozambique? The derived applied question is: where should the government intervene to maximise electricity access expansion?

The study integrates nighttime satellite imagery with socioeconomic, infrastructural, and conflict variables in a dataset of approximately 38.5 million observations at 1 km^2 spatial resolution, covering the entire Mozambican territory over 2015–2024. The study trains a Random Forest algorithm on 2015–2021 observations and evaluates its predictive capacity on the 2022–2024 period, which remains entirely unseen during training. Model interpretability is ensured through SHAP analysis, an interpretable AI tool quantifying each variable's contribution to predictions, applied at both national and provincial level. Results are translated into an operational planning tool through a gap analysis identifying areas with unrealised structural electrification potential.

The model achieves an R^2 of 0.785 on the out-of-time test set. SHAP analysis reveals that electricity access is primarily a socioeconomic phenomenon: individual wealth and population density jointly explain approximately 74% of total feature importance. Drivers vary geographically in a systematic manner: Maputo, Cabo Delgado, and Niassa present anomalous profiles interpretable respectively as effects of urban maturity, armed conflict, and extreme geographic marginalisation. The gap analysis identifies 26,874 priority pixels organised into three intervention clusters: grid extension (11–21 km), connection subsidies for Maputo, and off-grid solutions for Tete and Niassa.

A uniform national electrification strategy would be methodologically inefficient. Province-differentiated approaches are necessary to achieve universal access targets by 2030. The framework is annually updatable and represents a decision support tool directly usable by EDM and international organisations operating in Mozambique's energy sector.

Key-words: electricity, energy access, machine learning, SHAP, Mozambique.

Abstract in italiano

Questo studio sviluppa un approccio basato sull'intelligenza artificiale e sull'analisi di immagini satellitari ad alta risoluzione per mappare e analizzare le disparità di accesso all'elettricità in Mozambico nel periodo 2015–2024.

La domanda di ricerca principale è: quali fattori spiegano le disparità di accesso all'elettricità in Mozambico? La domanda applicativa derivata è: dove dovrebbe intervenire il governo per massimizzare l'espansione dell'accesso elettrico?

Lo studio integra dati di luminosità notturna da satellite con variabili socioeconomiche, infrastrutturali e di conflitto in un dataset di circa 38.5 milioni di osservazioni a risoluzione di 1 km², coprendo l'intero territorio mozambicano per il periodo 2015–2024. Lo studio addestra un algoritmo Random Forest sulle osservazioni 2015–2021 e ne valuta la capacità predittiva sul periodo 2022–2024, rimasto completamente escluso dalla fase di training. L'interpretabilità del modello è assicurata tramite analisi SHAP, uno strumento di intelligenza artificiale interpretabile che quantifica il contributo di ciascuna variabile alle predizioni, applicata sia a livello nazionale che per ciascuna delle dieci province. I risultati sono infine tradotti in uno strumento operativo attraverso una gap analysis che identifica le aree con potenziale strutturale non ancora realizzato.

Il modello raggiunge un R^2 di 0.785 sul test set out-of-time. L'analisi SHAP rivela che l'accesso elettrico è primariamente un fenomeno socioeconomico: reddito individuale e densità demografica spiegano circa il 74% dell'importanza totale delle variabili. I driver variano geograficamente in modo sistematico: Maputo, Cabo Delgado e Niassa presentano profili anomali interpretabili rispettivamente come effetto della maturità urbana, del conflitto armato e della marginalizzazione geografica estrema. La gap analysis identifica 26.874 pixel prioritari organizzati in tre cluster di intervento: estensione della rete (11–21 km), sussidi alla connessione per Maputo, e soluzioni off-grid per Tete e Niassa.

Una strategia nazionale uniforme sarebbe metodologicamente inefficiente. Approcci differenziati per provincia sono necessari per raggiungere gli obiettivi di accesso universale al 2030. Il framework è aggiornabile annualmente e rappresenta uno strumento di supporto decisionale direttamente utilizzabile da EDM e dalle organizzazioni internazionali operanti nel settore energetico del paese.

Parole chiave: elettricità, accesso all'energia, machine learning, SHAP, Mozambico

Glossary

ACLED (Armed Conflict Location and Event Data Project) - A global dataset that records the dates, locations, actors, and types of armed conflict events. In this study, ACLED data is used to construct a pixel-level measure of conflict exposure in Mozambique.

Builtup - The fraction of built-up area within a 1 km² pixel, derived from the Global Human Settlement Layer (GHSL). Used as a proxy for settlement density.

dist_grid_km - Euclidean distance in kilometres from a pixel to the nearest electricity transmission or distribution line. A proxy for the physical cost of grid connection.

EDM (Electricidade de Moçambique) - The national electricity utility of Mozambique, responsible for generation, transmission, distribution, and commercialisation of electricity.

Gap analysis - An analytical method that compares observed electricity access levels with model-predicted potential, identifying pixels where structural conditions are favourable for electrification but access has not yet been achieved.

GIS (Geographic Information System) - A system for capturing, storing, analysing, and visualising geographically referenced data.

HDI (Human Development Index) - A composite index developed by UNDP that measures average achievement in three dimensions of human development: health (life expectancy), education (years of schooling), and standard of living (income per capita). In this study, HDI is measured at district level.

HDF5 - A file format used by NASA for distributing satellite data, including the VIIRS Black Marble night-time light product.

LHF (Low-Hanging Fruit) - Pixels that are inhabited, currently dark (no observed electricity access), predicted by the model to have electrification potential, and located outside active conflict zones. LHF pixels represent priority areas for short-term electrification intervention.

log_ntl - The log-transformed night-time light intensity at pixel level, defined as log(NTL + 1). Used as the dependent variable of the predictive model.

log_pop - The log-transformed population estimate per pixel, derived from the WorldPop dataset.

Mini-grid - A small-scale electricity generation and distribution system serving a local community, typically powered by solar photovoltaic panels. An alternative to conventional grid extension in remote areas.

NDVI (Normalized Difference Vegetation Index) - A satellite-derived index measuring vegetation density. Used in this study as a proxy for rurality and land use intensity.

NTL (Night-Time Light) - The intensity of artificial light at night as detected by satellite sensors. Used as a proxy for electricity access in this study.

OnSSET (Open Source Spatial Electrification Tool) - A geospatial modelling tool that estimates least-cost electrification strategies — grid extension, mini-grids, or standalone systems — for different geographic contexts. Developed by KTH Royal Institute of Technology.

Out-of-time validation - A model validation scheme in which the model is trained on historical data and evaluated on future data not seen during training. More rigorous than random split validation for policy applications.

Pixel - The basic spatial unit of analysis in this study, corresponding to a 1 km² grid cell covering the territory of Mozambique.

R^2 (R-squared) - A statistical measure indicating the proportion of variance in the dependent variable explained by the model. An R^2 of 0.785 means the model explains 78.5% of the variance in the test set.

Random Forest - A machine learning algorithm that builds a large number of decision trees and combines their predictions. Particularly suited for complex, non-linear relationships between variables and robust to multicollinearity.

RMSE (Root Mean Square Error) - A measure of the average prediction error of a model, expressed in the same units as the dependent variable.

RWI (Relative Wealth Index) - A sub-national estimate of relative household wealth developed by Meta and CIESIN, derived from mobile connectivity data and satellite imagery. Available at high spatial resolution for all low- and middle-income countries.

SHAP (SHapley Additive exPlanations) - An interpretable artificial intelligence method that decomposes each model prediction into the individual contributions of each input variable. Based on cooperative game theory. Used in this study to identify the drivers of electricity access at both national and provincial level.

Temporal split - The validation scheme used in this study, in which the model is trained on data from 2015–2021 and tested on data from 2022–2024.

VIIRS (Visible Infrared Imaging Radiometer Suite) - A sensor aboard the NASA/NOAA Suomi NPP and NOAA-20 satellites that measures visible and infrared light. The source of the night-time light data used in this study.

VNP46A4 - The specific NASA Black Marble annual composite night-time light product derived from VIIRS, used as the dependent variable in this study.

WorldPop - A high-resolution gridded population dataset produced by the University of Southampton, providing annual population estimates at 100-metre resolution globally.

Summary

1. Introduction

1.1. Energy Access in Sub-Saharan Africa: A Data Problem

Access to electricity is one of the most important enablers of human development. It powers hospitals, schools, and businesses; extends productive hours beyond sunset; enables the use of modern information and communication technologies; and is a prerequisite for nearly every form of economic activity in the modern world. In general, access to electricity is one of the most effective drivers of a community's socioeconomic development. Yet, despite decades of international commitment to universal energy access, formalized in Sustainable Development Goal 7, which calls for affordable, reliable, sustainable, and modern energy for all by 2030, electricity remains inaccessible to hundreds of millions of people in the developing world. The vast majority of people who lack access live in sub-Saharan Africa (SSA), where in 2022 approximately 600 million people (about half the continent's population) did not have access to the electricity grid(1).

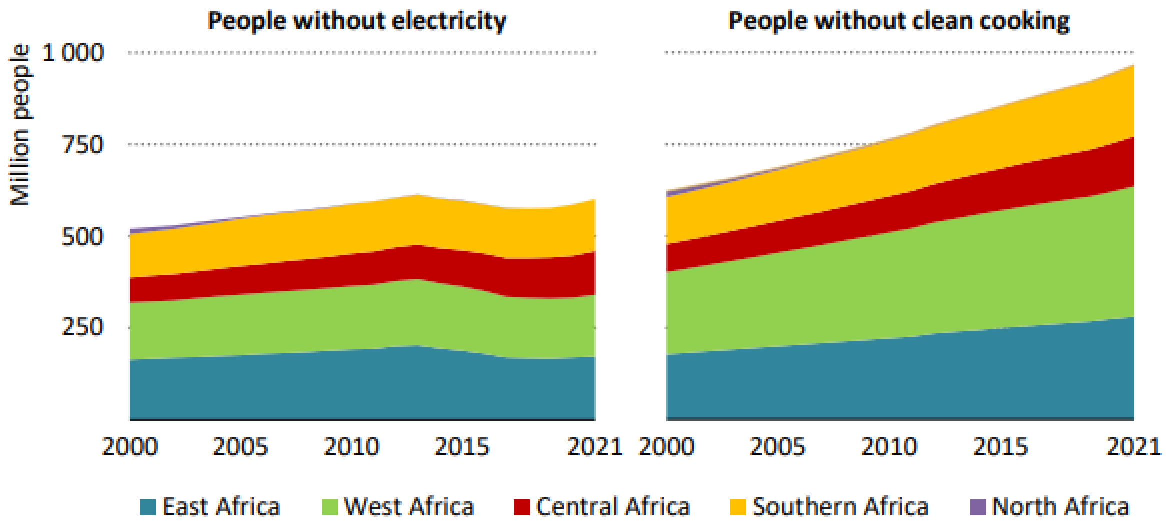


Figure 1- Population without access to modern energy services in Africa (IEA, WHO)

The challenge of expanding access to electricity in sub-Saharan Africa (SSA) is not merely technical or financial. It is also, fundamentally, a data problem. Effective energy planning requires knowing who has access and who does not, where unserved populations are concentrated, what factors explain patterns of access and exclusion, and where investments are most likely to have an impact. In most SSA countries, however, the data infrastructure needed to answer these questions is severely limited. National household sample surveys (the traditional tool for measuring energy access) are conducted infrequently, often at intervals of five or ten years, and are subject to sampling limitations that make subnational analysis difficult. Administrative data on grid coverage are often incomplete, outdated, or not publicly available. The consequences are concrete: without reliable data, both the design and evaluation of electrification policies operate on uncertain foundations. Falchetta et al. (2020)(2) document how this blind spot has allowed deep inequalities to go undetected, as electrification progress in SSA has been largely confined to urban and peri-urban areas, leaving the most remote rural populations systematically excluded.

This gap has drawn increasing attention from the scientific community, which has increasingly turned to satellite remote sensing as a scalable and economically viable alternative. Night-time lights (NTL) data (images of the Earth's surface at night captured by satellites) offer a particularly promising proxy for access to electricity. Areas with access to artificial lighting appear as bright zones in NTL images, while unelectrified areas remain dark. Unlike sample surveys, NTL data are available annually, at high spatial resolution, free of charge, and with global coverage. These properties make them particularly well-suited for monitoring the progress of electrification over time and space.

Night-time light data did not become a standard research tool overnight. Their credibility as a proxy for economic activity and energy access rests on a decade of converging empirical evidence. Henderson and Storeygard(3) were among the first to establish the link rigorously, showing that NTL intensity tracks GDP growth at the subnational level with sufficient precision to substitute for conventional economic statistics where these are absent or unreliable. Doll and Pachauri (2010)(4) extended this logic to the energy access problem specifically, demonstrating that satellite imagery of the night-time surface could be used to estimate the share of rural populations without electricity in developing countries, one of the earliest systematic attempts to measure energy exclusion from space. The question of which NTL product to use was later addressed by Gibson et al. (2021)(5), whose comparative assessment confirmed the VIIRS Black Marble product as the most appropriate choice for subnational analyses in low-income country contexts.

Figure 2- Artist's rendering of the Suomi National Polar-orbiting Partnership (Suomi-NPP) satellite in orbit. Source: NASA

This study relies on the VIIRS Black Marble VNP46A4, a NASA annual composite product derived from the Visible Infrared Imaging Radiometer Suite sensor aboard the Suomi-NPP satellite. It provides annual composites of average nighttime radiance at a native spatial resolution of 500 meters, subsequently aggregated to 1 km² for this study. What distinguishes VNP46A4 from earlier NTL products is the rigour of its pre-processing pipeline. Romàn et al. (2018)(6) provide a detailed account of the corrections applied (atmospheric effects, lunar seasonality, cloud contamination) each of which had been a source of systematic noise in predecessor datasets. Elvidge et al. (2017)(7) provided a comparative assessment of available NTL sensors, confirming VIIRS as the reference product for research applications requiring high-resolution annual data.

Falchetta et al. (2019)(2) took this a step further by developing a high-resolution gridded dataset on electrification in sub-Saharan Africa (SSA) based on NTL data, and demonstrating the feasibility of a systematic analysis of electricity access at the pixel level on a continental scale.

In parallel with the development of satellite data, the literature has seen significant growth in the application of machine learning to the estimation of energy access and economic development. Yeh et al. (2020)(8) demonstrated that deep learning models applied to satellite imagery are capable of predicting indicators of economic well-being in Africa with high accuracy. Ratledge et al. (2022)(9) extended this approach by assessing the impact of electricity access on livelihoods in Africa through machine learning techniques, paving the way for a more sophisticated use of these tools in energy policy evaluation. Chi et al. (2022)(10) developed the Relative Wealth Index (RWI) using mobile connectivity data and machine learning to produce high-resolution estimates of relative wealth for all low, and middle-income countries, one of the central independent variables in this study.

From a methodological standpoint, Belgiu and Drăguț (2016)(11) demonstrated the superiority of the Random Forest over other methods in a variety of remote sensing applications. Hengl et al. (2018)(12) demonstrated that Random Forest is a framework particularly well-suited for spatial and spatiotemporal modeling. Lundberg and Lee (2017)(13) developed the SHAP framework for interpreting machine learning models, a central tool for the analysis of access drivers developed in this study.

Despite these advances, the existing literature has three main gaps that this study aims to address. First, most studies use coarse spatial resolutions and focus on limited time periods, without incorporating socioeconomic variables at the pixel level. Second, studies tend to estimate average drivers of access at the national level, implicitly assuming homogeneous relationships across the territory, an assumption that this study explicitly tests through provincial SHAP analysis. Third, most studies do not translate analytical results into operational recommendations for decision makers. These gaps motivate the approach developed in this study, discussed in detail in Section 1.3.

1.2. Case Study: Mozambique

Mozambique serves as an ideal case study for analyzing disparities in access to electricity in sub-Saharan Africa, for three interrelated reasons: the severity of the energy deficit, the extreme spatial heterogeneity, and the presence of conflict dynamics that create unique conditions for analysis.

1.2.1. The energy deficit

Mozambique is one of the countries with the lowest rates of electricity access in the world. According to IEA data(14), in 2022, 52.6% of the Mozambican population lacked access to electricity, one of the highest rates on the African continent and significantly higher than the Sub-Saharan African average. However, this aggregate figure masks an even more pronounced disparity between urban and rural areas: while access exceeds 70% in major cities, in rural areas, where the majority of the population lives, it drops to less than 30%. Mozambique therefore faces a twofold problem: an overall quantitative deficit, and a deeply unequal distribution of the limited access that does exist.

Figure 3- The Cahora Bassa Dam on the Zambezi River, Mozambique, with an installed capacity of approximately 2,075 MW the largest hydroelectric facility in the country and one of the largest in Africa

This situation is paradoxical given the country's energy resources. Mozambique possesses significant reserves of natural gas in the Rovuma Basin, in Cabo Delgado Province. It also has one of the highest hydroelectric potentials on the continent, with the Cabora Bassa dam on the Zambezi River representing one of the largest hydroelectric plants in Africa. Yet, the majority of its population lives in darkness. This contradiction is one of the factors that makes the Mozambican case scientifically and politically significant.

.

1.2.2. Spatial Heterogenity

The second factor that makes Mozambique an exceptional case study is its extreme spatial diversity. Mozambique stretches over 2,500 km along the Indian Ocean coast, and the differences across that distance are not marginal.

The south, anchored by Maputo Province and the capital, has a relatively developed electricity grid, higher population density, and income levels well above the national average. The northern provinces of Niassa, Cabo Delgado, and Nampula present a substantially different picture: extreme poverty, dispersed settlements, and rural electricity access that in many districts approaches zero. This geographic heterogeneity is not merely a demographic fact but reflects centuries of unequal development, with the south historically more integrated into the commercial and institutional networks of Southern Africa and the north more isolated.

This diversity makes Mozambique a particularly suitable setting for testing whether the drivers of access vary geographically, which is precisely one of the questions this study aims to answer through provincial SHAP analysis.

1.2.3. The conflict in Cabo Delgado

The third distinctive feature of the Mozambican case is the ongoing armed conflict in Cabo Delgado province, which began in 2017 and was still ongoing at the time this study was written. The conflict, waged by local armed groups with documented ties to international jihadist networks, has caused the displacement of over one million people, the destruction of infrastructure in an area already among the least developed in the country, and the disruption of major natural gas development projects that were expected to transform the province's economy(15).

For this study, the Cabo Delgado conflict represents not only a contextual element but an analytical one. The existing literature documents that armed conflict has a systematic negative impact on access to electricity(16): it destroys existing infrastructure, prevents new investments, and creates structural discontinuities in the relationships between socioeconomic variables and energy access.

This study adopts a socio-technical systems engineering approach to a critical humanitarian problem (electricity access) and does not engage with the full political, institutional, and diplomatic complexity of the Mozambican context. Mozambique's energy deficit is embedded in a multi-layered reality that encompasses weak institutional capacity, endemic corruption, ethnic and religious tensions, external donor dependency, and the largely insufficient response of multilateral institutions to the country's structural challenges. Acknowledging this complexity does not fall within the scope of the present analysis, which focuses on the quantitative and spatial dimensions of electricity access disparities. The results and recommendations of this study should therefore be read as contributions to evidence-based energy planning, not as a comprehensive assessment of the socio-political conditions that ultimately determine the feasibility of any intervention.

1.3. Research Question and Objectives

This study aims to address one main research question and one derived policy question.

The main research question is: What factors explain disparities in access to electricity in Mozambique during the period 2015–2024?

The policy question is: Where should the Mozambican government focus its interventions in the coming years to maximize the expansion of electricity access?

The two questions are hierarchically ordered: the second derives from the first. Identifying the structural drivers of electricity access is the analytical prerequisite for identifying the areas where an intervention is most likely to succeed. The answer to the applied question, developed through the gap analysis and in the policy implications, is therefore empirically grounded in the results of the main question, not on a priori assumptions about where it is most appropriate to intervene.

1.3.1. Specific objectives

To address these questions, the study pursues five specific objectives.

The first objective is to construct a pixel-year panel dataset (a structured dataset in which each observation corresponds to a specific 1 km² geographic unit (pixel) in a specific year, allowing both spatial and temporal patterns to be analysed simultaneously) that integrates VIIRS NTL data with socioeconomic, infrastructural, environmental, and conflict variables for Mozambique over the period 2015–2024. This dataset, with approximately 38.5 million observations at 1 km^2 resolution, serves as the empirical foundation of the study and constitutes a methodological contribution in itself, given its unprecedented spatial and temporal granularity in the Mozambican context.

The second objective is to develop and validate a high-resolution predictive model of electricity access at pixel level, using an artificial intelligence algorithm, specifically a Random Forest Regressor, with an out-of-time validation scheme that tests the model's ability to generalise to future unobserved periods. The choice of the temporal split ,with training on 2015–2021 and testing on 2022–2024, is motivated by the need to evaluate the model's predictive capability in a realistic policy context, where forecasts must necessarily refer to future periods not observed during training.

The third objective is to quantify the relative importance of the drivers of electricity access using SHAP (SHapley Additive exPlanations) analysis(13), both at the global level and disaggregated by province. This objective directly addresses the main research question and yields the study's most original scientific contribution: the demonstration that access drivers vary geographically in a systematic and interpretable manner, with direct implications for the design of region-specific policies.

The fourth objective is to conduct a gap analysis that identifies the pixels where the model predicts a higher level of access than that observed, i.e., the areas where structural conditions are favorable for electrification but access has not yet been achieved. These pixels represent the priority areas for intervention in the short term, and their aggregation at the provincial level produces an actionable list of priorities that decision-makers can use directly.

The fifth objective is to translate the analytical results into policy recommendations differentiated by clusters of provinces, each with distinct intervention strategies: grid extension for areas with limited distance from the grid, off-grid and mini-grid solutions for remote areas, and integrated approaches for conflict-affected areas.

This study contributes to the existing literature in three ways, addressing the gaps identified in the previous section.

From a methodological standpoint, it introduces the temporal split as a validation scheme for predictive models of electricity access based on NTL data (a more rigorous approach than the random split commonly used in the which does not assess temporal generalization ability, a limitation documented in several energy access studies that rely exclusively on cross-sectional or random-split validation schemes(2,9). It also introduces provincial SHAP analysis as a tool to identify the geographic heterogeneity of drivers, going beyond the simple estimation of global feature importance to produce interpretable results at the subnational level.

On the empirical front, it constructs the most granular dataset ever used to analyze electricity access in Mozambique (previous studies have analysed NTL-based electricity access in SSA at coarser spatial resolutions or over shorter time periods(2,17)), without integrating socioeconomic variables at pixel level or conducting subnational driver disaggregation and uses it to systematically document access disparities and their drivers at the provincial level. The results confirm the dominance of socioeconomic variables over purely infrastructural variables, with important implications for the debate on the infrastructure-first narrative in electrification policies in SSA.

In practical terms, it produces a tool that can be directly applied to energy planning: a prioritized list of "low-hanging fruit" areas where structural conditions are favorable for electrification and safety constraints allow for infrastructure projects. This output transforms the present study from an academic exercise into a concrete decision-support tool for the Mozambican government and for international organizations operating in the country's energy sector.

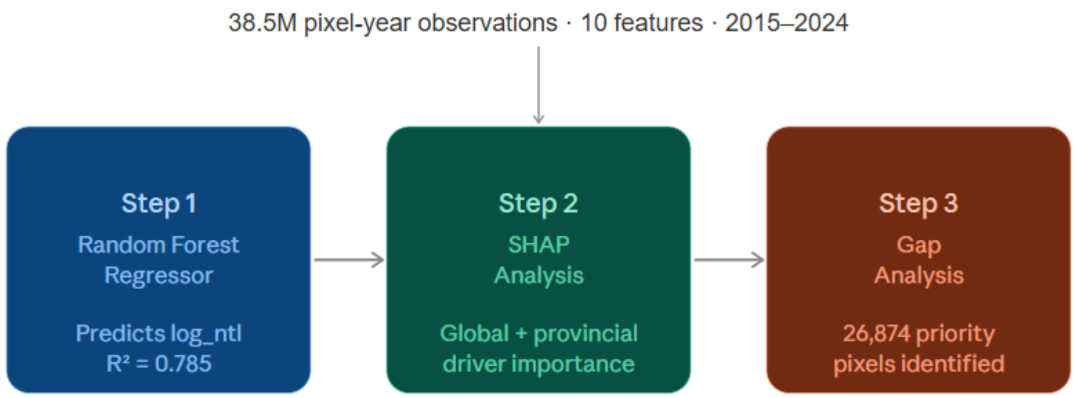


Figura 4 - Analytical framework of the study

1.4. Structure of the thesis

This thesis is organized into six chapters, each of which contributes in a distinct way to answering the main research question.

Chapter 1 (this introduction) presents the global and regional context of the issue of access to electricity in sub-Saharan Africa, explains the choice of Mozambique as a case study, formulates the research question and specific objectives, and describes the study's original contribution relative to the existing literature.

Chapter 2 describes the Mozambican case study in depth. It presents the context of the energy sector and electrification, analyzes the country's spatial heterogeneity, and discusses key events from 2015 to 2024, particularly the conflict in Cabo Delgado and major electrification policies, which provide the interpretive context necessary to understand the analytical results in subsequent chapters.

Chapter 3 describes the construction of the dataset. It outlines the sources of the data used, the data processing and cleaning procedures, the construction of the dependent variable and independent variables, the handling of ACLED conflict data, and the final panel structure of the dataset.

Chapter 4 presents the exploratory data analysis. It includes descriptive statistics, analysis of temporal patterns, analysis of spatial patterns, and analysis of the correlation structure among variables. This section provides a basic understanding of the data distribution and the relationships between variables prior to the modeling analysis.

Chapter 5 is the analytical core of the thesis. It presents the Random Forest predictive model (methodology, feature engineering, and validation scheme), the model's performance results at the global and provincial levels, the SHAP analysis of access drivers at the global and provincial levels, and the gap analysis with the identification of priority "low-hanging fruit" areas for intervention.

Chapter 6 discusses the results in relation to the research question, presents differentiated policy implications by province cluster, outlines the methodological and substantive limitations of the study, and suggests directions for future research.

2. Study Area: Mozambique

Figure 5 - Mozambique map by province

2.1. Energy Sector and Electrification Context

Mozambique is a country of approximately 34 million people located along the eastern coast of southern Africa. Despite significant energy resources, it has one of the lowest levels of per capita energy consumption in the world(14). In 2022, only 48% of the population had access to electricity, with a marked disparity between urban (76%) and rural (15%) areas(14). However, this aggregate figure masks a further complexity: formal access to the grid does not necessarily equate to reliable access. Salite et al. (2021)(18) document that Mozambique ranks 114th out of 137 countries in terms of electricity supply quality, with frequent outages and voltage fluctuations that damage household appliances and limit the actual value of the connection for families.

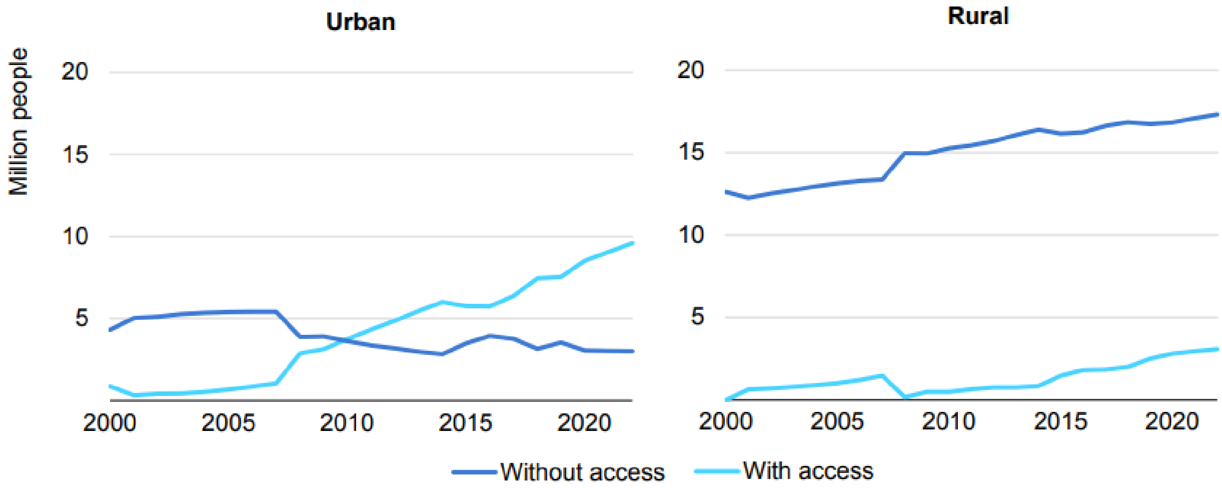


Figure 6 - Urban and rural electricity access in Mozambique

Mozambique presents one of the continent's most striking energy paradoxes: it is rich in resources, yet the majority of the population remains without access to electricity. Its hydroelectric potential is estimated at over 18 GW in total, of which only 2.1 GW has been developed(14). The Cahora Bassa Dam (HCB) on the Zambezi River, with approximately 2,075 MW of installed capacity, is the country's main power generation asset. However, over two-thirds of its output is exported to South Africa under agreements dating back to the colonial era, significantly limiting the availability of energy for the domestic market(18). In terms of natural gas, the country has reserves estimated at 3.766 trillion cubic meters recoverable in the Rovuma Basin(14), among the largest in Africa, but tax revenues from major LNG projects are not expected until at least the mid-2030s. Solar potential is also significant, with an average irradiance of 2,100 kWh/m²/year(14), but remains largely untapped.

The electricity sector is dominated by EDM (Electricidade de Moçambique), the vertically integrated state-owned utility that acts as the sole buyer, system operator, and sole supplier in grid-connected areas. Electricity rates in Mozambique are among the highest in Southern Africa (10 US cents/kWh) but, paradoxically, are not cost-reflective(18). The average energy expenditure of connected households accounts for between 39.6% and 59.4% of the minimum monthly wage(14), a form of energy poverty that affects even those who have formal access to the grid.

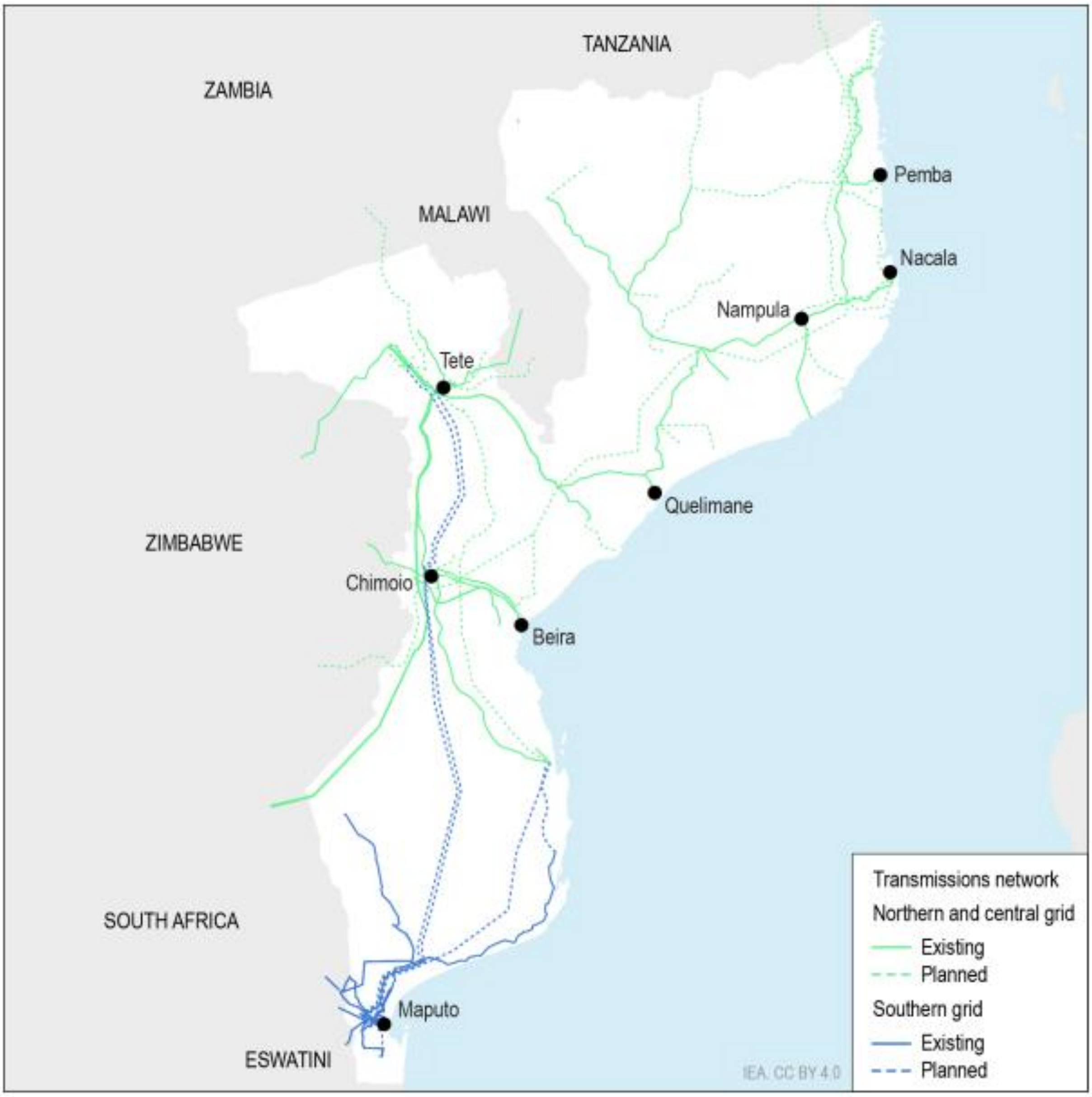


Figure 7 - Mozambique's planned and existing transmission network, 2023 (IEA)

On the policy front, in 2023 the government approved the Energy Transition Strategy (ETE), which sets a goal of universal access to electricity by 2030. The strategy outlines six priority programs for the electricity sector(14). On the generation front, it calls for the addition of 2–4 GW of new hydropower capacity by 2031. For renewables, the target is 1–2 GW of solar and 200–500 MW of wind power by 2030. On the infrastructure front, the ETE plans to connect the country's two separate transmission systems. For off-grid access (particularly relevant for remote rural areas), the strategy calls for mini-grids and solar home systems to contribute 30–35% toward universal access by 2030(14). Between 2017 and 2022, the number of connections increased by 50% (from 8.2 to 12.7 million), but population growth continues to erode progress in terms of access rates(14).

2.2. Spatial Heterogeneity

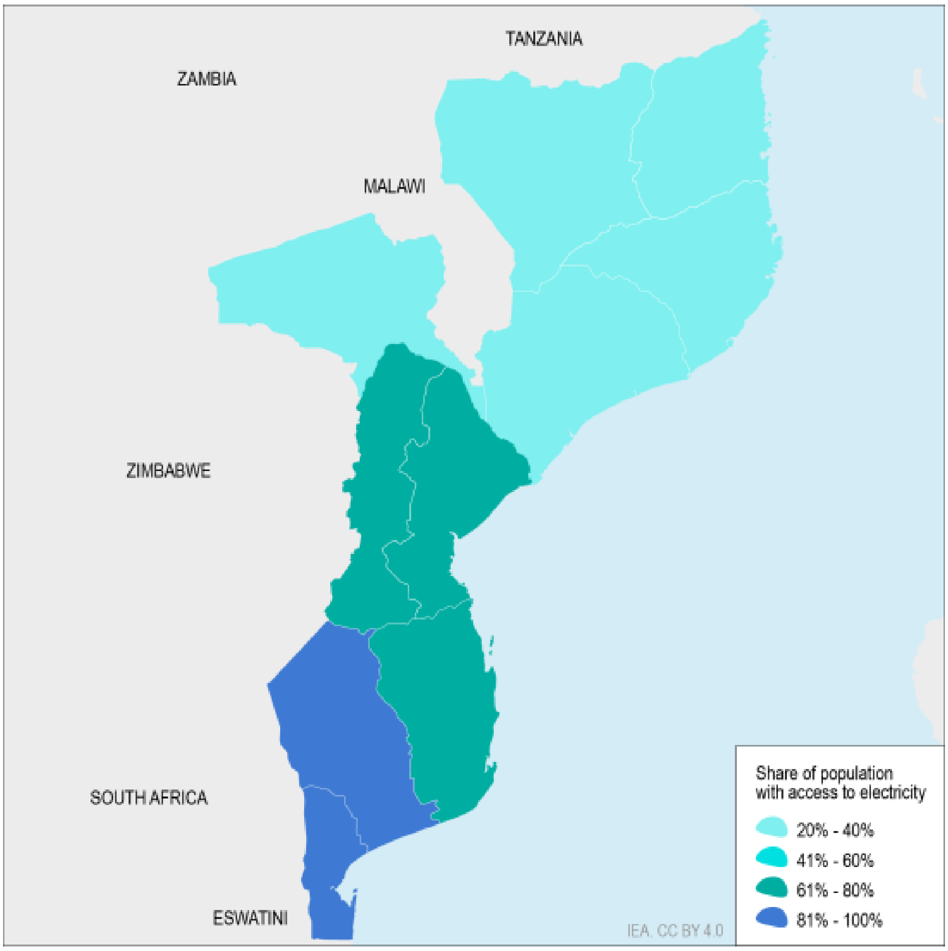


Figure 8 - Share of population with access to electricity by province in Mozambique, 2022 (IEA)

Mozambique stretches over 2,500 km from north to south along the Indian Ocean coast, with radically different geographical, demographic, economic, and institutional characteristics across its ten provinces. This spatial heterogeneity is not a neutral background factor, but rather one of the main drivers of the disparities in access that this study aims to analyze.

2.2.1. The North-South Divide

The most significant geographical divide is between the south and the north. The south (the provinces of Maputo, Gaza, and Inhambane) is historically more developed, with more extensive infrastructure and higher income levels. Maputo has an electrification rate of 98% among registered residential customers(18), the highest in the country, and accounts for the majority of public and private investment in the energy sector. Proximity to generation sources and connections to the South African grid ensure a relatively more stable supply.

The north (the provinces of Niassa, Cabo Delgado, Nampula, and Zambezia) presents a radically different picture. These provinces are characterized by extreme poverty, low population density, and virtually non-existent infrastructure in rural areas. The geographical distance from generation sources is enormous: Nampula is located 731 km from HCB(14). Nampula, despite being the economic hub of the north, has an urban electrification rate of 89% but relies on a floating power plant (Karpowership) anchored in the port of Nacala as a backup source(14), a temporary solution that reflects the region's infrastructure fragility.

The central region (the provinces of Sofala, Manica, Tete, and Zambezia) occupies a central position. Sofala is home to Beira, the country's second-largest port and a major regional logistics hub, with an urban electrification rate of 88%(18). Tete is the center of Mozambique's extractive industry (coal and hydroelectric power), with energy dynamics influenced by the presence of large-scale industrial projects.

.

2.2.2. The urban-rural divide

The urban-rural divide overlaps with the north-south geographical divide, amplifying it. For the 70% of Mozambique's population living in rural areas, the rate of access to electricity remains at 15%(14), and in many northern areas it drops to near-zero levels. The dispersed settlement patterns of rural communities, connected by low-density road networks, create significant additional costs for grid expansion(18). This dual heterogeneity, geographic and urban-rural, is why this study employs a pixel-level analysis at a 1 km^2 resolution: only high spatial granularity allows for capturing the intra-provincial disparities that data aggregated at the national or provincial level conceal.

Furthermore, the country does not have a unified national power grid: there is a central-northern grid, where most of the hydroelectric resources are located, and a separate southern grid that relies significantly on imports from South Africa(14).

2.3. Key Events 2015-2024

The period from 2015 to 2024 was marked by a series of events that significantly influenced the dynamics of electricity access in Mozambique and provide the necessary context for understanding the results of the analysis.

2.3.1. The conflict in Cabo Delgado (since 2017)

Since 2017, the province of Cabo Delgado has been the scene of an armed jihadist insurgency led by the group known locally as Al-Shabaab, an Arabic term meaning "the youth," reflecting the movement's predominantly young membership(15). The group's roots date back to the mid-2000s, when fundamentalist cells began organizing in the Mocímboa da Praia region. The shift to armed violence occurred in October 2017, resulting in over 3,000 civilian casualties and at least 800,000 displaced persons(15). In 2019, the group pledged allegiance to the Islamic State, integrating into the Islamic State of Central Africa Province (ISCAP)(15).

The causes of the insurgency are manifold. Socioeconomically, Cabo Delgado is one of Mozambique's poorest provinces, with 86% of the population living below the poverty line(15), high unemployment rates, and a lack of prospects for young people. The discovery of large natural gas fields between 2009 and 2010 generated unmet expectations: international investments created extractive enclaves without significant benefits for the local population, exacerbating inequalities (15).

Economically, the conflict led to the suspension in 2021 of TotalEnergies' Mozambique LNG project, with significant repercussions for the province's development prospects. For this study, the conflict is a central analytical element: as discussed later, the model performs significantly worse in this province, a result we interpret as indirect evidence of the disruptive impact of the conflict on the structural relationships between socioeconomic variables and access to electricity.

2.3.2. Cyclones Idai and Kenneth (2019)

In March 2019, Cyclone Idai struck the city of Beira and the central provinces of Mozambique, causing extensive damage to the electrical infrastructure. Beira lost the use of one of its two main transmission lines, which remained out of service for months(18). EDM had to rebuild 800 km of medium-voltage lines, 600 km of low-voltage lines, and 47 transformers with support from the UN, the EU, and the World Bank(14). Cyclone Kenneth, which struck Cabo Delgado a few weeks later, further exacerbated the situation in an area already destabilized by conflict. These extreme weather events, whose frequency is expected to increase with climate change, illustrate the vulnerability of Mozambique's energy infrastructure to exogenous shocks and the difficulty of maintaining and expanding access in a context of growing environmental instability.

3. Dataset Construction

3.1. Building a Data Pyramid: Datasets and Data Processing

This study relies on the integration of various heterogeneous data sources into a single spatially and temporally consistent analytical dataset. To understand the process, it is helpful to conceptualize it as a data pyramid, a hierarchical process in which raw satellite and administrative inputs are progressively transformed into variables and then used in models. The pyramid metaphor captures the essential characteristics of the approach: each successive level involves a reduction in raw data volume and an increase in analytical structure, progressing from the original raster files at the base to model-ready panel observations at the apex. Similar frameworks have been adopted in studies combining satellite data and socioeconomic variables for the analysis of electricity access in sub-Saharan Africa ((2);(9)).

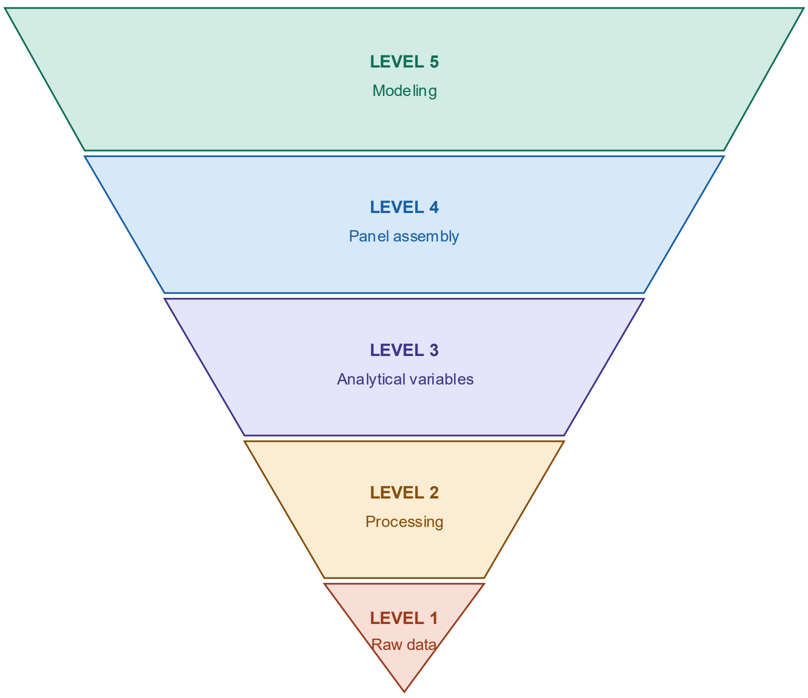


Figure 9- The data pyramid framework adopted in this study, illustrating the five-level hierarchical processing pipeline from raw satellite and administrative inputs (Level 1) to model-ready panel observations and modeling output (Level 5)

The pyramid is divided into five distinct levels. The first level comprises the raw data: HDF5 files downloaded from NASA Earthdata for the VIIRS VNP46A4 Black Marble product, monthly MODIS MODI13A3 rasters, WorldPop R2025A files in GeoTIFF format, ACLED data in CSV format, shapefiles of the power grid and services downloaded from OpenStreetMap via the Overpass API, SHDI tables from the Global Data Lab, and RWI rasters from Meta/CIESIN. The second level involves processing this raw data: converting HDF5 files to GeoTIFF, mosaicking tiles to cover the entire national territory, cropping to the Mozambican border, sampling NTL values at the centroid of each pixel in the 1-km analytical grid, calculating Euclidean distance variables, and spatially aggregating ACLED events using 50-km buffers. The third level involves the construction of analytical variables: log_ntl, log_pop, annual NDVI averages, the conflict exposure indices conflict_violence and instability, and static distance variables. The fourth level corresponds to the assembly of the panel: the merging of all variables onto the pixel × year grid, thereby producing a final dataset of approximately 3.86 million pixels over 10 years. The fifth level is modeling: Random Forest regression on log_ntl, spatial validation via Leave-Location-Out cross-validation with provinces as geographic blocks, analysis of variable importance using SHAP, and spatial mapping of predictions.

The unit of analysis is the pixel, defined at a spatial resolution of 1 km. This spatial resolution has been chosen to guarantee consistency across all variables in the dataset. While VIIRS Black Marble NTL data are available at 500 metres resolution, several key independent variables present in the data set, in particular HDI and RWI, are not available at such resolution in data-scarce context as Mozambique. Instead, the 1 km resolution represents the finest scale at which all variables in the dataset can be considered jointly reliable.

3.2. Dependent Variable: Night-Time Lights

The dependent variable is the NTL (night-time lights) value, extracted by the VIIRS Black Marble VNP46A4 product, developed by NASA. This product provides free annual composites of average nighttime radiance at a native resolution of 500 meters, in HDF5 format. This product represents the gold standard among freely available NTL products, as it includes corrections for atmospheric, lunar, and seasonal effects.

Figura 10-Global night-time light composite from NASA's Black Marble product (VIIRS VNP46), 2016.

Before describing the processing pipeline, it is helpful to briefly introduce the satellite system and the sensor from which the data originate.

The data are acquired by the Visible Infrared Imaging Radiometer Suite (VIIRS), a multispectral radiometer aboard the Suomi National Polar-orbiting Partnership (Suomi-NPP) satellite, launched in October 2011 as part of a joint mission between NASA and NOAA. The Day/Night Band (DNB) is a panchromatic band sensitive to low-intensity light emissions in the 500–900 nm spectral range, with a native spatial resolution of approximately 750 meters at nadir. As documented by Elvidge et al. (2017)(7), VIIRS offers three key improvements over its predecessor, the DMSP-OLS (Defense Meteorological Satellite Program Operational Linescan System): the absence of pixel saturation in densely lit urban areas; substantially finer spatial resolution; and a wider dynamic range that allows for the detection of low-radiance emissions characteristic of partially electrified rural settlements. These properties make VIIRS the preferred data source for mapping electrification in contexts such as Mozambique, where the signal of interest lies predominantly in the low and medium-low radiance range (Gibson et al., 2021)(5).

The Black Marble (VNP46) product suite, developed by Román et al. (2018)(6), represents the state of the art among freely available nightlight products. The annual composite product VNP46A4 is generated from the moonlight-corrected daily nightlight product VNP46A2 by selecting high-quality, cloud-free observations and applying corrections for atmospheric effects, terrain, vegetation, snow cover, the bidirectional reflectance distribution function (BRDF) of the moon, and diffuse light. The result is a gap-filled annual composite of average nighttime radiance at a native resolution of 15 arcseconds, corresponding to approximately 500 meters at the equator, provided in HDF5 format. This product is available from January 2012 to the present and is accessible for free via NASA's Earthdata portal.

The raw HDF5 files were downloaded for each year of the 2015–2024 period. For each year, the tiles were converted from HDF5 to GeoTIFF files to allow for better visualization in QGIS software, then assembled into a mosaic to cover the entire area of interest, and finally cropped to Mozambique's national border using a vector mask derived from GADM administrative boundaries.

The resulting raster, with a native resolution of approximately 500 m, was then resampled to the panel's 1 km grid by calculating the average radiance value of all 500 m pixels falling within each 1 km cell. The entire pipeline was executed in python and systematically replicated for all years in the panel, ensuring consistency in the processing procedure over time.

Then, these data were used to represent electrification in different ways. The list of the ways used is reported below.

1. *ntl*
   This represents the average annual radiance value for each pixel, expressed in $nW/cm^2/sr$. It represents the effective measure of the night-time light emitted by a single pixel in 1 year. Otherwise, this variable has limitations due to the fact that for the majority of the pixels this value is close to zero, as Mozambique is a country with low electrification and is predominantly rural, while it is very high for a few pixels, representing cities. This asymmetric distribution make the variable *ntl* difficult to use in regression models.

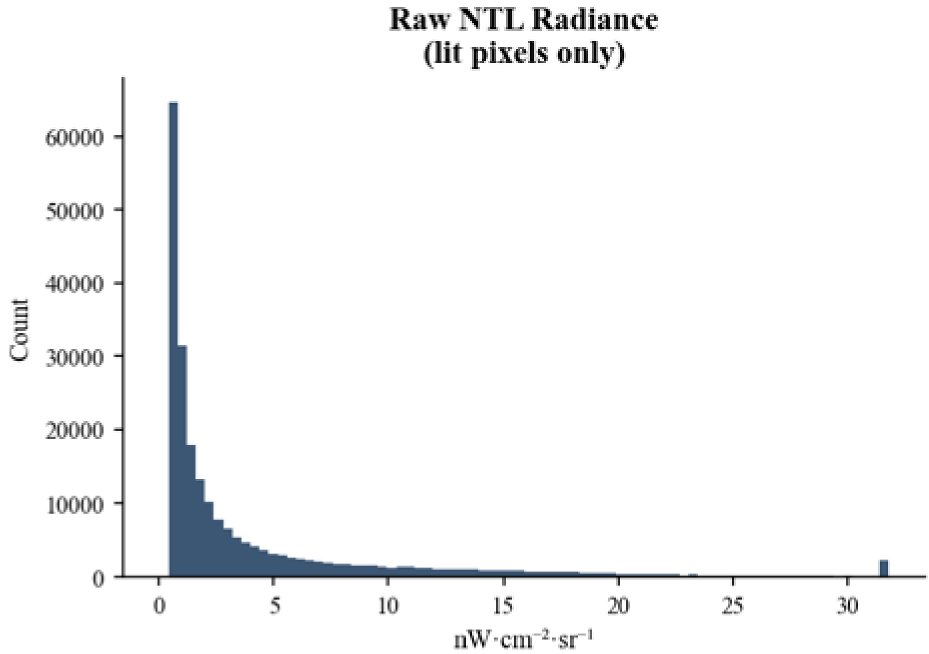


Figure 11-Distribution of raw NTL radiance among lit pixels (ntl_raw > 0), full panel 2015–2024 (N = 226,888).

Figure 11 illustrates this distribution. The y-axis is displayed on a logarithmic scale to make both the mass of near-zero values and the sparse tail visible simultaneously; on a linear scale, the tail would be completely invisible. The histogram confirms the expected pattern for a low-electrification context: most pixel-year observations cluster very close to zero radiance, while a small number of observations form a long right tail, with values exceeding 20 nW/cm^2/sr and peaking above 30 nW/cm^2/sr, corresponding to the brightest urban pixels, primarily located in the Maputo metropolitan area.

2. *log_ntl*

This variable is mathematically represented by the value log(NTL + 1). It was introduced in order to overcome the problem of the variable *ntl*. On the one hand, the logarithmic transformation allows the right tail of the distribution to be compressed, reducing the influence of extreme urban values; on the other hand, it allows the regression coefficients to be interpreted in percentage terms, so that a unit increase in an independent variable is associated with a proportional change in *ntl* intensity. This approach is a standard method used in NTL literature to address asymmetric radiance distributions(3,8). The +1 in front of the *ntl* variable ensures that pixels with radiance equal to 0 receive a value equal to zero and are not eliminated from the analysis, due to the property of logarithms.

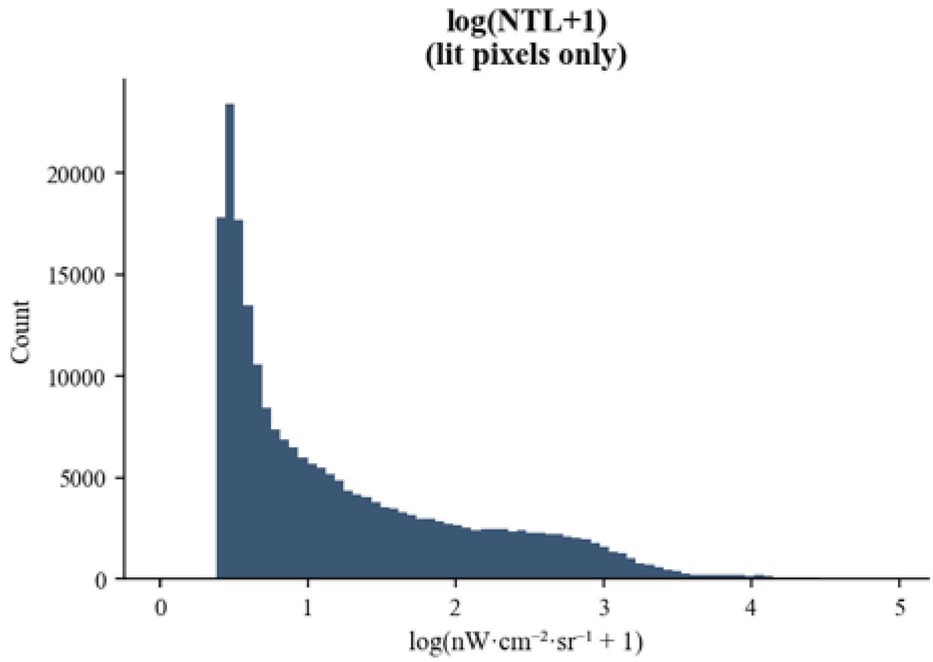


Figura 12-Distribution of log_ntl among lit pixels (log_ntl > 0), full panel 2015–2024 (N = 226,888).

Figure 12 shows the distribution of the transformed variable log (NTL + 1) for lit pixels. Compared to the raw NTL histogram, the right tail is substantially compressed and the mass of observations is more smoothly spread over a wider range of values, although the distribution remains right-skewed. This transformation yields a variable that is considerably better behaved from a modelling perspective, reducing the leverage of extreme urban outliers while preserving meaningful variation across rural and peri-urban pixels.

3. *growth_ntl*

The variable is defined by the following equation:

$$\Delta NTL_{i,t} = \ln\left(NTL_{i,t} + 1\right) - \ln\left(NTL_{i,t-1} + 1\right)$$

This approximates the year-on-year percentage growth rate of NTL intensity. It is used in exploratory analyses and as a robustness check, but is not adopted as a main dependent variable in the model.

There are several sources of uncertainty associated with the VIIRS NTL data. First, although the Black Marble product includes atmospheric corrections, cloud cover in tropical regions can reduce the actual number of valid observations contributing to the annual composites. Second, a known artifact of the Black Marble correction algorithm produces slightly negative radiance values in some pixels (minimum observed in this dataset: −0.029 nW/cm²/sr); these values are treated as zero in the construction of log_ntl, in line with standard practice(6,7). Third, and more fundamentally, NTL intensity is a proxy for electricity access and not a direct measure of it: it cannot distinguish between different end uses of electricity, different consumption levels, or the presence of off-grid sources such as solar lanterns and diesel generators that may not generate sufficient radiance to be detected at 500-meter resolution. As shown by Gibson and Boe-Gibson (2021)(19), VIIRS NTL data explain a high proportion of the cross-sectional variation in economic activity but a significantly smaller proportion of the temporal variation, suggesting caution in interpreting changes in radiance over time as one-to-one variations in electrification or income levels.

3.3. Independent Variables

This section describes the independent variables included in the dataset, their sources, and the rationale for their inclusion. They are intended to capture the main structural and dynamic determinants of spatial heterogeneity in electricity demand across Mozambique. Their selection reflects a twofold criterion: nationwide availability and resolution compatible with the panel's 1-km grid, and theoretical relevance with respect to known mechanisms that determine electricity supply and demand in rural and peri-urban areas of low-income countries(2,4). The independent variables are as follows:

1. *log_pop*
   The population per pixel is included in the dataset as a fundamental control variable and as a structural determinant of electrification. The data were obtained from WorldPop, an interdisciplinary applied research group focusing primarily on spatial demographic data. The data, which are part of the *Individual countries 2015–2030 (1 km resolution) R2025A v1* product (WorldPop, 2025)(20), were downloaded for each year in the panel at 1 km resolution and then aligned to the

1 km grid used in this study. In order to ensure consistency with the dependent variable and reduce the skewness of the distribution, population enters the models as **log_pop = log(pop + 1)**. The expected relationship between this variable and the dependent variable is positive: all else equal, more densely populated pixels are likely to exhibit higher levels and faster growth of electrification. It should be noted, however, that WorldPop population estimates carry their own uncertainty, particularly in contexts characterised by outdated or low-quality census data and in rural areas or informal settlements, where the spatial distribution of the population is more difficult to capture(20).

2. *HDI (SHDI)*

   L'Human Development Index (HDI) is a composite index of life expectancy, education and per capita income indicators, developed by the economist Mahbub ul-Haq and then used by the United Nations Development Programme's Human Development Report Offiice. The Global Data Lab combined it with other indicators to create the Subnational Human Development Index (SHDI), which is the same index but different for each province in Mozambique(21). The SHDI is therefore more useful than the HDI for our dataset, as it is able to observe territorial differences. The data are assigned at the district level and attributed to all pixels falling within the corresponding district; the variable therefore presents cross-district spatial variation but limited temporal variation, given the slow pace at which the HDI evolves in the short term. However, geographical resolution is also a limitation. The expected relationship with the NTL is positive: districts with higher levels of development tend to show a greater increase in electrification.

3. *RWI*

   The Relative Wealth Index (RWI) is an estimate of relative wealth at the sub-district level, produced by Meta in collaboration with the Center for International Earth Science Information Network (CIESIN) of the Columbia University(10). The RWI is derived from a model that integrates mobile connectivity data, daytime satellite imagery, and auxiliary geospatial variables, with a native resolution of approximately 2.4 km. Unlike the HDI, the RWI is not a composite index of multidimensional development but a measure of relative wealth between pixels within a country (a pixel with a positive RWI is relatively richer than the national average, one with a negative RWI is relatively poorer). The variable is static, referring approximately to the year 2020, and does not capture temporal variations. It is included in the model as a complement to the HDI, capturing spatial heterogeneity in wealth at a finer scale than the district level. A limitation of the RWI is that the underlying model relies on mobile connectivity data that may be sparse in rural areas and conflict-affected zones, potentially introducing a systematic bias precisely in the areas of greatest interest for electrification policies.

4. NDVI
   The Normalized Difference Vegetation Index (NDVI) is included as a control variable capturing land cover and vegetation density at the pixel level. NDVI is derived from the MODIS MOD13A3 product(22), which provides monthly composites of surface reflectance at 1 kilometre resolution. In this study, NDVI serves two theoretical roles. First, it controls for land cover heterogeneity, forested and agricultural pixels are systematically different from urban and peri-urban pixels in ways that affect both electricity demand and the socioeconomic variables of interest. Second, it captures seasonal and inter-annual variability in agricultural productivity, which in a predominantly rural economy like Mozambique may influence household income and therefore energy affordability. The expected sign of the NDVI coefficient is ambiguous a priori: denser vegetation is associated with rural areas and lower electrification, but also with agricultural productivity and income.

5. *dist_grid_km*
   The distance from the nearest power grid, expressed in kilometers, is included as a measure of infrastructure accessibility. Data on power lines was downloaded from *Open Infrastructure Map* via the Overpass API, which allows the OSM database to be queried with custom geographic queries. The query extracts all power lines and cables within the bounding box of Mozambique. As we saw in section 3.1, Mozambique has an electricity grid concentrated in large urban areas and undeveloped in rural areas, as in most developing countries. It is derived from the digitization of the transmission and distribution lines available for Mozambique and calculated as the Euclidean distance from the centroid of each pixel to the nearest grid section. The expected relationship is negative: pixels further away from the network are less likely to be connected and therefore have slower electrification.

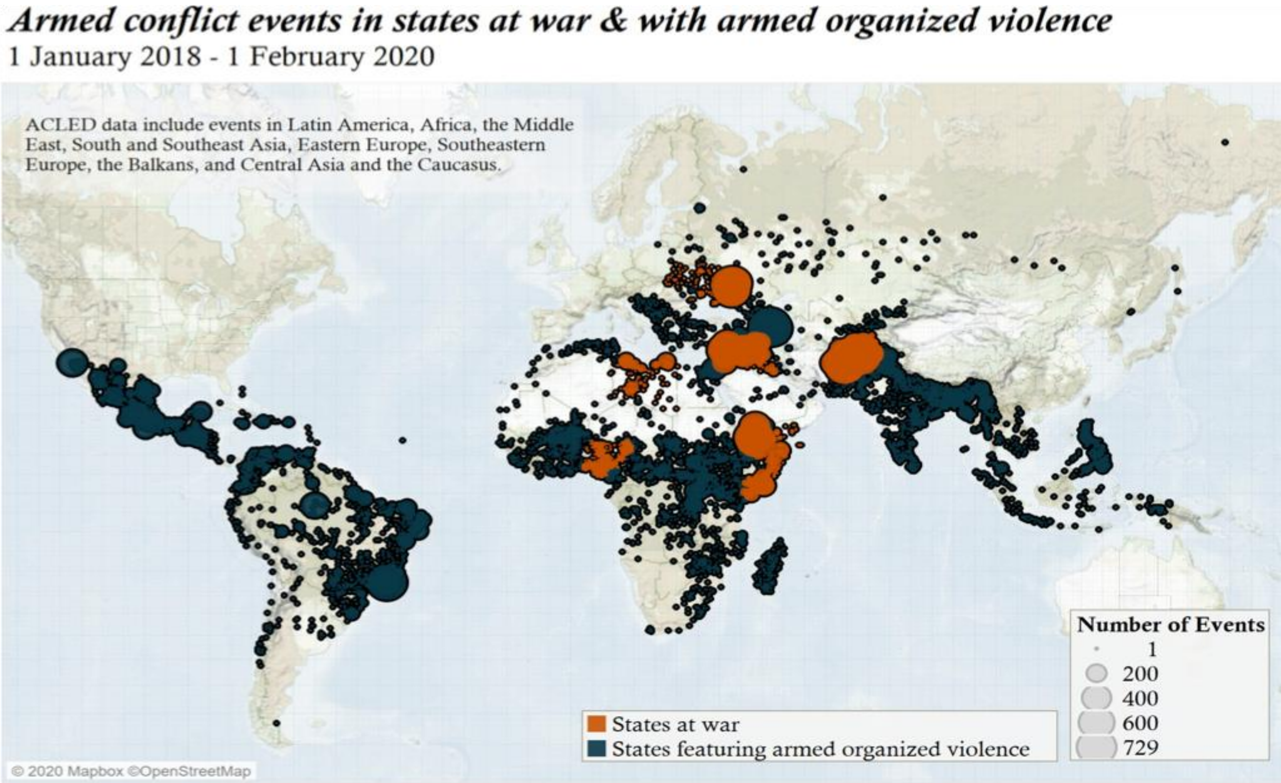
6. *dist_hospital_km, dist_school_km, dist_market_km*
   Distances from essential services (nearest hospitals, schools, and markets) are included as proxies for infrastructure provision and local development levels. Pixels closer to these services tend to be located in more densely populated and better-served areas, with a higher probability of access to the electricity grid. All three variables are calculated as the Euclidean distance in kilometers from the centroid of each pixel to the nearest point of interest. As with *dist_grid_km*, these variables are static (they do not capture the expansion of services over the study period).

7. *Conflict_violence* and *Instability*
   Two conflict variables are included in the dataset (*conflict_violence* and *instability*) derived from the ACLED dataset. Their construction and the spatial aggregation procedure adopted are described in detail in Section 3.4.

8. builtup_m2
   The built-up area variable measures the total surface area of artificial surfaces like buildings and constructions, within each 1km pixel, expressed in square metres. It is derived from the Global Human Settlement Layer (GHSL) and it provides a direct measure of urbanisation intensity at the pixel level.
   Despite its intuitive appeal as a predictor of electrification, the variable *builtup_m2* is excluded from the main model specifications due to its near-perfect collinearity with *log_pop*. The two variables capture essentially the same underlying phenomenon (urban concentration) and their simultaneous inclusion would introduce severe multicollinearity, inflating standard errors and making the individual coefficients unreliable. Since *log_pop* is available for all years and provides a cleaner measure of demographic pressure, it is retained in the main specifications while *builtup_m2* is excluded.

3.4. ACLED Conflict Data Processing

Mozambique has experienced significant episodes of conflict and political instability over the study period, most notably the Cabo Delgado insurgency which escalated sharply from 2017 onwards. Given the potential for conflict to disrupt economic activity, damage infrastructure, and displace populations, its inclusion as an explanatory variable in the model is theoretically motivated. At the same time, political instability in the form of protests and riots tends to be geographically concentrated in areas with higher population density and greater economic activity, suggesting a potentially different, and possibly opposite, relationship with electricity access compared to direct violence. For these reasons, conflict and instability are treated as distinct constructs and operationalised through separate variables derived from the ACLED dataset.



Figue 13- Global distribution of armed conflict events recorded by ACLED, January 2018 – February 2020. Circle size reflects the number of events per location. Source: ACLED / Mapbox / OpenStreetMap

Data on conflicts and political instability was taken from ACLED (Armed Conflict Location and Event Data Project) dataset, a georeferenced source that records the date, location, type, and number of fatalities from individual incidents of conflict and political violence on a global scale(23). The ACLED dataset classifies events into six categories: battles, violence against civilians, explosions/remote violence, riots, protests, and strategic developments. For the purposes of this study, the six categories have been aggregated into two distinct variables, excluding one category. The variable *conflict_violence* includes battles, violence against civilians, and explosions/remote violence, capturing direct physical violence and its capacity for immediate economic disruption. The variable *instability* includes riots and protests, capturing political and social dissent, a phenomenon that tends to be concentrated in peri-urban areas. Strategic developments, which record logistical movements and repositioning of armed actors rather than episodes of actual violence, are excluded from both variables as they are not considered to have an impact on economic activities and electrification patterns. For each pixel and each year, all ACLED events belonging to the included categories that fall within a 50 km buffer from the pixel's centroid are identified. The value assigned to the pixel is a weighted sum of these events, where the weight of each event is defined as

$$w = \frac{1}{1 + distance\ [km]}$$

ensuring that geographically close events contribute more to the exposure score than events at the edge of the buffer. Pixels for which no events are recorded in the buffer receive a value of zero. The choice of a 50 km buffer reflects a balance between the need to capture the spatial spillover effects of conflict and the risk of excessive spatial smoothing. This spatial aggregation approach follows the methodology adopted in AfroGrid(24), a validated framework for the integration of ACLED data with geospatial socioeconomic variables in Africa.

3.5. Panel Structure

The final analytical dataset is a spatiotemporal panel with the pixel (1×1 km grid cell) as the unit of observation and the calendar year as the temporal dimension. The panel covers the period 2015–2024, including all 10 available years, for a total of 10 annual observations per pixel. The geographic coverage includes the entire territory of Mozambique, excluding water bodies and pixels with missing values for at least one variable, for a total of approximately 3,863,232 pixels and approximately 38,632,320 pixel-years.

The panel combines time-varying variables (log_ntl, log_pop, NDVI, conflict_violence) with slowly changing variables such as HDI/SHDI, and purely static variables such as RWI, dist_grid_km, dist_hospital_km, dist_school_km, and dist_market_km. The instability variable was excluded from the final model due to its high correlation with conflict_violence (ϱ = 0.96, Spearman), which would have introduced informational redundancy and distorted the interpretation of SHAP values(13). The year variable was similarly excluded: its correlation with log_ntl on illuminated pixels is close to zero (ϱ = −0.051, Spearman), and its inclusion does not add structural explanatory power but risks capturing general temporal trends that are not directly informative for the research question. This combination of time-varying and static predictors justifies the choice of the Random Forest over fixed-effects panel estimators: while two-way fixed effects would account for all time-invariant variation, including the infrastructure and wealth variables central to the research question, the Random Forest can simultaneously capture both cross-sectional and temporal variation(11,12).

Table 2 provides a comprehensive summary of all variables included in the panel dataset, including their sources, native spatial resolution, temporal coverage, and classification as time-varying or static.

Table 2. Summary of variables included in the panel dataset.

Variable	Source	Native resolution	Temporal coverage	Type	Role
log_ntl	NASA VIIRS Black Marble VNP46A4	500 m	2015–2024 (excl. 2016)	Time-varying	Dependent variable
log_pop	WorldPop R2025A v1	1 km	2015–2024	Time-varying	Socioeconomic
rwi	Meta / CIESIN (Chi et al., 2022)	~2.4 km	~2020 (static)	Static	Socioeconomic
hdi	Global Data Lab SHDI (Smits & Permanyer, 2019)	District level	2015–2024	Slowly varying	Socioeconomic
ndvi	MODIS MOD13A3 (Didan, 2015)	1 km	2015–2024	Time-varying	Environmental
dist_grid_km	OpenStreetMap via Overpass API	Vector	Static	Static	Infrastructure
dist_hospital_km	OpenStreetMap via Overpass API	Vector	Static	Static	Infrastructure
dist_school_km	OpenStreetMap via Overpass API	Vector	Static	Static	Infrastructure
dist_market_km	OpenStreetMap via Overpass API	Vector	Static	Static	Infrastructure
conflict_violence	ACLED (Raleigh et al., 2010)	Event-level	2015–2024	Time-varying	Conflict
builtup	Global Human Settlement Layer (GHSL)	100 m	Static	Static	Geographic context

4. Exploratory Data Analysis

This chapter presents a targeted descriptive analysis of the dataset prior to estimation. The objective is to characterise the distributional properties of the key variables, document the temporal evolution of night-time light intensity over the study period, and illustrate the spatial distribution of electricity access across the Mozambican territory. These analyses provide the empirical foundation for interpreting the modelling results developed in Chapter 5 and situate the dataset within the electrification dynamics documented in the sub-Saharan African literature(2,25).

4.1. Descriptive Statistics

Table 3. Descriptive Statistics — Full Panel (N = 38,632,320) and Lit Pixels Only (N = 226,888)

Variable	Mean	Std	Min	25%	Median	75%	Max	Skewness
Panel A: All pixels (N = 38,632,320)								
ntl_raw	0.026	0.634	0.000	0.000	0.000	0.000	140.09	49.10
log_ntl	0.007	0.115	0.000	0.000	0.000	0.000	4.949	19.79
log_pop	0.111	0.360	0.000	0.000	0.000	0.000	5.605	5.11
ndvi	0.570	0.109	-0.200	0.517	0.576	0.636	0.982	-1.59
rwi	-0.489	0.202	-1.839	-0.618	-0.505	-0.382	1.968	1.25
hdi	0.463	0.042	0.402	0.431	0.443	0.495	0.657	1.25
dist_grid_km	51.86	44.14	0.002	18.22	39.23	73.30	249.32	1.29
dist_school_km	51.18	34.57	0.006	24.26	44.12	71.32	189.26	0.98
dist_hospital_km	64.64	37.88	0.007	35.49	58.57	88.23	208.98	0.77
dist_market_km	77.84	46.58	0.012	41.88	69.78	105.90	222.38	0.69
builtup	28.28	224.73	0.000	0.000	0.000	0.000	9025.0	13.17
conflict_violence	0.993	8.642	0.000	0.000	0.000	0.000	140.0	12.85
instability	2.053	14.184	0.000	0.000	0.000	0.000	244.0	12.07
Panel B: Lit pixels only (N = 226,888)								
ntl_raw	4.413	7.009	0.011	0.747	1.586	4.830	140.09	4.07
log_ntl	1.251	0.835	0.011	0.558	0.950	1.763	4.949	1.04
log_pop	2.064	1.498	0.000	0.712	2.088	3.315	5.605	0.04
ndvi	0.443	0.101	-0.142	0.378	0.443	0.510	0.844	-0.22
rwi	0.338	0.504	-1.018	-0.035	0.272	0.650	1.968	0.51
hdi	0.517	0.077	0.402	0.443	0.500	0.603	0.657	0.47
dist_grid_km	11.45	18.12	0.009	2.206	5.328	11.594	194.96	3.57
dist_school_km	12.63	19.32	0.011	1.611	4.061	13.639	158.08	2.36
dist_hospital_km	17.22	24.12	0.007	2.696	6.064	22.194	177.62	2.19
dist_market_km	20.49	27.81	0.017	2.911	7.450	27.114	210.70	2.11

Variable	Mean	Std	Min	25%	Median	75%	Max	Skewness
builtup	1445.6	1608.5	0.000	0.000	758.0	2703.0	9025.0	0.84
conflict_violence	2.002	3.398	0.000	0.000	0.000	3.000	140.0	5.46
instability	11.29	22.28	0.000	0.000	0.000	16.00	244.0	3.15

Table 3 reports descriptive statistics for all variables in the dataset, computed on the full panel of approximately 38,632,320 pixel-year observations (Panel A) and separately for the subset of lit pixels only (Panel B, N = 226,888). For each variable, mean, standard deviation, minimum, median, maximum, and skewness coefficient are reported. The skewness coefficient is particularly informative for assessing the need for logarithmic transformations and anticipating potential issues in the modelling phase.

First, the extreme asymmetry of the dependent variable ntl_raw (skewness = 49.1 in the full panel) confirms the necessity of the log transformation adopted in all model specifications. After log-transformation, the distribution of log_ntl among lit pixels is substantially more symmetric (skewness = 1.04), with mean 1.251 and median 0.950.

Second, the contrast between Panel A and Panel B reveals a consistent pattern: electrified pixels are systematically wealthier (mean RWI rises from −0.489 to +0.338), more densely populated (mean log_pop rises from 0.111 to 2.064), closer to the electricity grid (mean dist_grid_km drops from 51.9 to 11.5 km), and closer to all service infrastructure. This descriptive pattern anticipates the SHAP findings of Chapter 5, where RWI and log_pop emerge as the dominant predictors of electricity access.

4.2. Temporal patterns

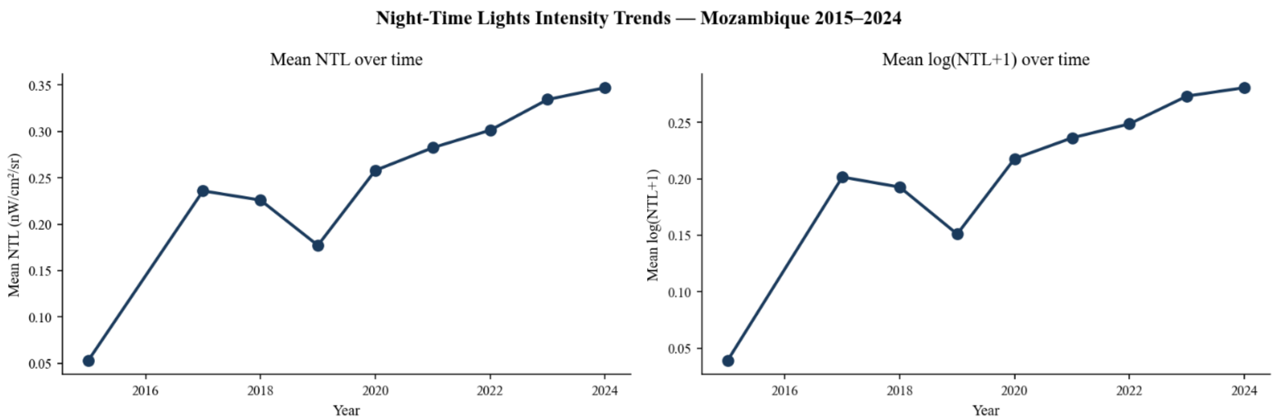


Figura 15-Evolution of mean NTL intensity (left axis) and mean log_ntl (right axis) over the study period 2015–2024. Values computed on the full panel of inhabited pixels.

Figure 15 shows the evolution of mean NTL intensity over the study period. Three aspects merit attention.

First, the 2015 values are anomalously low across both metrics. Direct inspection of the raw VNP46A4 raster files confirms that these values are genuine and present in the original NASA data prior to any processing step. The 2015 observations are retained in the panel and treated as informative data points, with the awareness that their contribution to the model may be partially influenced by this satellite limitation. This anomaly does not affect the temporal split validation scheme, as 2015 falls within the training period.

Second, from 2017 onwards there is a clear and sustained upward trend in mean NTL intensity, consistent with the gradual expansion of electricity access documented by national statistics and with the increase in grid connections from 8.2 to 12.7 million between 2017 and 2022 reported by the IEA(14).

Third, a negative dip is visible in 2019 followed by a sharp recovery in 2020. This pattern is consistent with the impact of Cyclone Idai, which struck central Mozambique in March 2019, causing widespread infrastructure damage including the destruction of 800 km of medium-voltage lines and large-scale population displacement. The recovery in 2020 reflects the reconstruction efforts documented in Section 2.3.

4.3. Spatial Patterns

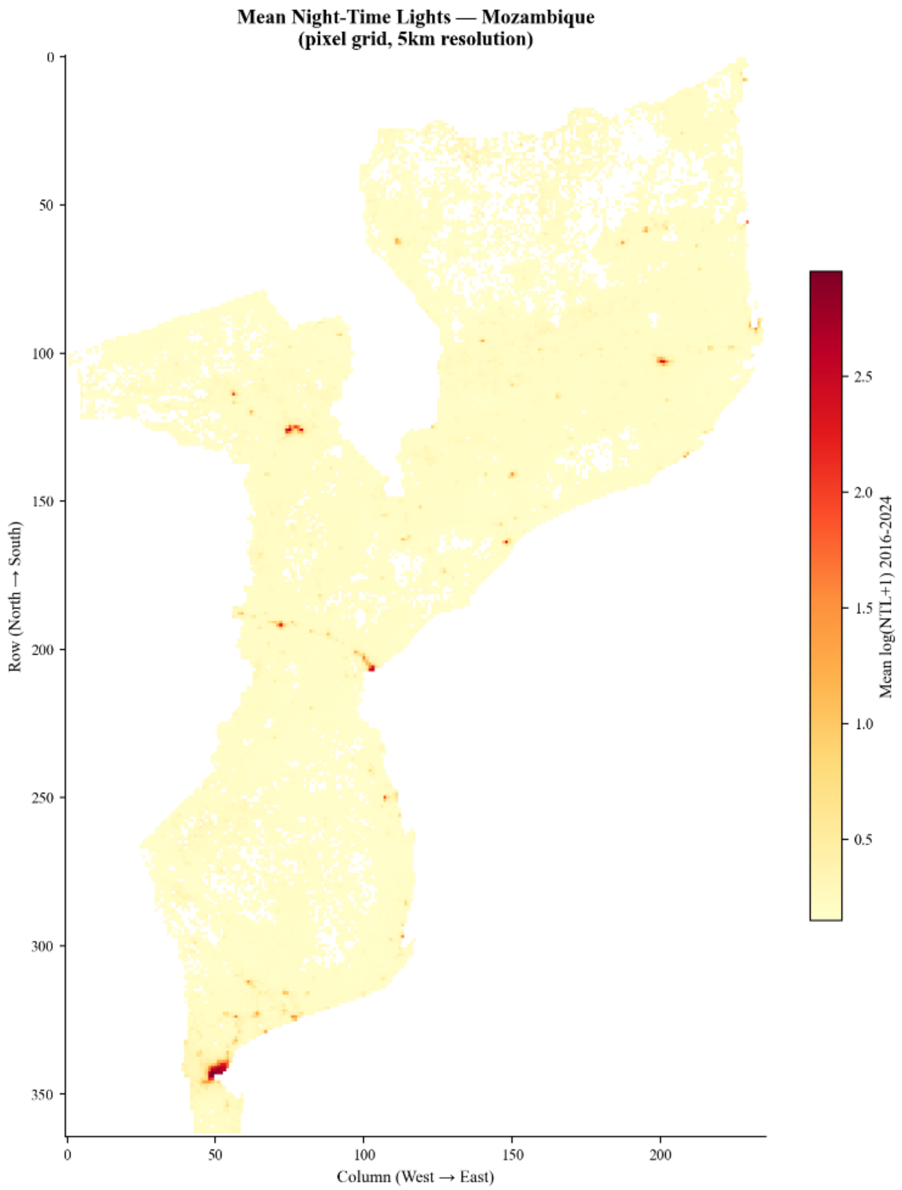


Figura 16-Spatial distribution of mean log_ntl across Mozambique, averaged over the full study period 2015–2024. Brighter pixels indicate higher average NTL intensity

Figure 16 shows the mean log_ntl for each pixel averaged over the full study period. The map reveals a pattern of extreme spatial concentration that motivates the provincial disaggregation developed in Section 5.4.2.

The brightest area is the Maputo metropolitan area in the south, the dominant economic and administrative centre of the country. Secondary concentrations are visible in Beira on the central coast, Nampula in the north, Tete in the west, and Quelimane on the northern coast. A corridor of scattered low-intensity light runs northward from Maputo along the coastal road, reflecting the linear settlement pattern associated with transport infrastructure.

The northern provinces (Cabo Delgado, Niassa, and parts of Nampula) show uniformly low NTL values throughout the study period, reflecting the combination of low population density, limited grid coverage, and the progressive deterioration of security conditions associated with the Cabo Delgado insurgency from 2017 onwards. This spatial pattern is consistent with what has been documented by Falchetta et al. (2019)(2) for sub-Saharan Africa as a whole, and reflects the dual structure of electrified urban areas versus unelectrified rural areas that characterises the Mozambican energy sector(25).

The extreme geographic concentration visible in Figure 16, with the vast majority of the territory recording near-zero NTL values and light concentrated in a small number of discrete urban nodes, is the spatial expression of the dataset imbalance discussed before: 99.4% of pixel-year observations record log_ntl = 0. This pattern directly motivates the balanced sampling strategy adopted in the SHAP analysis and the definition of the gap analysis threshold in Chapter 5.

5. Predictive Model

5.1. Random Forest Regressor — Methodology

This study uses a Random Forest Regressor to model the spatial and temporal determinants of electricity access in Mozambique. The algorithm builds a large ensemble of decision trees and returns their averaged prediction, which reduces the instability that characterises individual trees. Three features of this method make it well-suited to the present problem(26). The first is its ability to capture non-linear predictor relationships without prior specification, a relevant property given that electrification dynamics vary considerably across geographic and socioeconomic contexts. Second, it is robust to multicollinearity among features, an important property given the structural correlations among the socioeconomic variables in the dataset documented in Section 4.1. Third, it provides variable importance measures directly from the training process, a property that is central to the interpretive objectives of this study and that motivates the subsequent SHAP analysis. Alternative ensemble methods such as Gradient Boosting or XGBoost were considered but not adopted, as the Random Forest provides equivalent interpretability within the SHAP framework while requiring less sensitivity to hyperparameter tuning in high-dimensional spatial panel datasets(26).

The dependent variable is the annual composite NTL intensity at the pixel level, expressed as log-transformed radiance (log_ntl = log(NTL + 1)). A continuous formulation was preferred over a binary electrification indicator for two reasons. First, NTL intensity carries information about the degree of electrification, not merely its presence: a pixel with log_ntl = 0.8 differs substantially from one with log_ntl = 0.1 in terms of energy consumption and economic activity. Second, regression on a continuous target preserves more information and avoids the arbitrary selection of a binary threshold, which would introduce an additional source of uncertainty with no empirical basis in the Mozambican context.

The model was trained on a pixel-year panel dataset covering the full territory of Mozambique over the period 2015–2024, comprising approximately 38.5 million observations, each representing a 1 km^2 pixel in a given year. Hyperparameters were set as reported in Table 1:

Table 1. Random Forest Hyperparameters

n_estimators	max_depth	min_samples_leaf
300	15	50

These values balance predictive capacity and overfitting risk. Each hyperparameter reflects a specific concern. n_estimators = 300 is sufficient for stable ensemble predictions; returns diminish sharply beyond this point. max_depth = 15 constrains tree complexity, reducing the risk that the model learns patterns specific to the 2015–2021 training period rather than structural relationships that persist into the test years. min_samples_leaf = 50 addresses the dataset's most pressing structural problem: with 99.4% of observations recording log_ntl = 0, unrestricted leaf splitting would allow the model to fit spurious relationships in the small minority of electrified pixels.

5.2. Feature Engineering and Train/Test Split

The model is trained on a pixel-year panel covering Mozambique at 1 km^2 resolution over 2015–2024, approximately 38.5 million observations in total. Independent variables fall into four categories, following the structure introduced in Section 3.1.

The socioeconomic group includes log_pop, RWI, and district-level HDI. These variables collectively describe the economic and demographic profile of each pixel: how many people live there, how wealthy they are relative to the national distribution, and how developed their district is by composite human development standards.

Infrastructure proximity is captured through distances to the nearest grid line (dist_grid_km), hospital (dist_hospital_km), school (dist_school_km), and market (dist_market_km).

Geographic and environmental context is captured by NDVI (Normalized Difference Vegetation Index, a proxy for rurality and vegetation cover) and builtup (built-up area fraction per pixel, derived from the Global Human Settlement Layer). NDVI indirectly reflects the degree of urbanisation and land use intensity, while builtup provides a direct measure of settlement density.

The conflict exposure variable is conflict_violence, derived from the ACLED dataset, which records the weighted count of armed violence events within a 50 km buffer of each pixel per year. This variable is included to capture the disruptive effect of conflict on electricity infrastructure expansion, particularly relevant for northern Mozambique where the Cabo Delgado insurgency has significantly slowed infrastructural development since 2017.

A structural feature of the dataset that warrants explicit attention is the severe imbalance of the target variable: 99.4% of observations record log_ntl = 0, corresponding to completely dark pixels. This reflects the reality of electrification in Mozambique, as documented by the IEA, in 2022 more than half of the population (52.6%) lacked access to electricity(14). This imbalance has two important methodological implications.

First, a model that always predicts zero would achieve a spuriously high R^2 simply by virtue of the target distribution. For this reason, performance metrics are reported separately for the subset of lit pixels (log_ntl > 0), denoted R^2 test (lit), in addition to the full dataset. The R^2 on lit pixels only is substantially lower (0.487 vs. 0.785 in the temporal split), reflecting the model's reduced ability to capture variance in already-electrified pixels, where dynamics are more contingent and less structurally determined.

Second, SHAP values were computed on a balanced sample of lit and dark pixels rather than a random sample. A random sample, being predominantly composed of dark pixels, would have biased the feature importance analysis toward the drivers of darkness rather than the drivers of access. The balanced sample ensures that the SHAP analysis reflects the determinants of electricity access across both electrified and unelectrified contexts.

Two distinct validation schemes were adopted, each serving a methodologically distinct role.

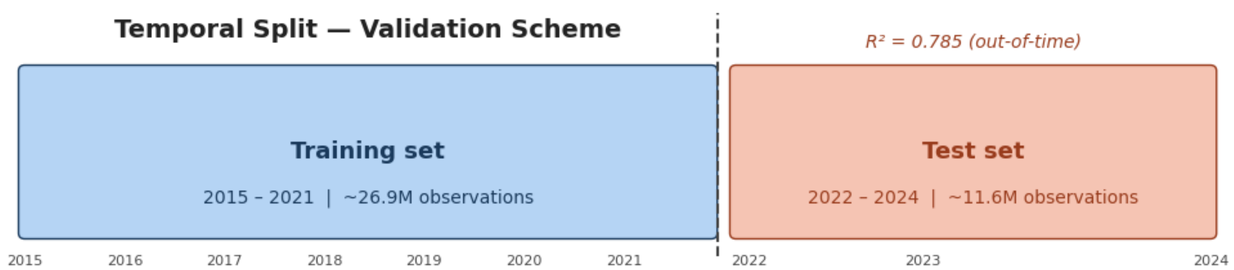


Figura 18-Temporal validation scheme. The dataset is divided into a training set (2015–2021) and an out-of-time test set (2022–2024). The random split robustness benchmark uses an 80/20 random division across all years.

The temporal split divides the dataset into a training set (years 2015–2021, approximately 26.9 million observations) and a test set (years 2022–2024, approximately 11.6 million observations). This scheme simulates an out-of-time forecasting context: the model is trained on historical observations and evaluated on its ability to generalise to future periods not observed during training. This is the methodologically most rigorous assessment of predictive capacity in a real policy context, where forecasts must necessarily refer to future unobserved periods.

The random split randomly divides the dataset into training (80%, approximately 30.8 million observations) and test (20%, approximately 7.7 million observations), preserving the temporal distribution across both subsets. This approach maximises predictive performance but does not assess temporal generalisation ability. The random split is therefore used as a robustness benchmark: if both models produce similar feature importance hierarchies, the results can be considered robust to the choice of validation scheme.

The R^2 gap between the two schemes (0.785 in the temporal split vs. 0.881 in the random split) is expected and methodologically informative. It quantifies the cost of temporal generalisation: approximately 10 percentage points of R^2 are lost when the model is asked to predict future years rather than years mixed into the training set. This gap does not represent a model deficiency, but an honest measure of the intrinsic difficulty of out-of-time prediction in a spatiotemporal panel context.

5.3. Model Performance and Validation

5.3.1. Temporal Split vs Random Split

The performance results of the two models are reported in Table 2:

Table 2. Random Forest Model Performance: Temporal Split vs. Random Split

Metric	Temporal Split	Random Split
R^2 train	0.894	0.890
R^2 test (all pixels)	0.785	0.881
R^2 test (lit pixels only)	0.487	0.670
RMSE test (all)	0.060	0.044
RMSE test (lit)	0.596	0.533

The temporal split achieves an R^2 of 0.785 on the test set, compared to 0.881 for the random split. This gap of approximately 10 percentage points is expected and does not represent a model deficiency; it is an honest measure of the cost of temporal generalisation.

In the random split, training and test sets contain observations from all years in the dataset. This scheme maximises predictive performance but introduces a form of temporal data leakage: the model has effectively "seen the future" during training, in the sense that it has been exposed to the average characteristics of the test years. The result is an optimistic estimate of predictive capacity that does not reflect real-world forecasting conditions.

In the temporal split, by contrast, the model has never been exposed to any data from the 2022–2024 period during training. It must generalise not only to new pixels, but to an entirely new temporal period. The temporal split is therefore the only scheme that correctly assesses predictive capacity in a real operational context, where forecasts must necessarily refer to future unobserved periods. For this reason, all results and interpretations in the following sections are based on the temporal split model, with the random split serving exclusively as a robustness benchmark.

A particularly relevant finding is the stability of model performance across individual years in the temporal split test set, as reported in Table 3:

Table 3. Temporal Split Model Performance by Year (Test Set: 2022–2024)

Year	R^2	RMSE
2022	0.796	0.056
2023	0.775	0.062
2024	0.785	0.062

The variation in R^2 across the three test years is minimal (0.796 in 2022, 0.775 in 2023, 0.785 in 2024) and shows no degradation trend. This is an important signal of structural stability: the relationships between socioeconomic features and electricity access have not changed significantly between the training period (2015–2021) and the test period (2022–2024). Had a major structural shift occurred (for instance, a large-scale electrification policy that altered the relationship between income and access, or an infrastructure expansion that reduced the explanatory weight of grid distance) we would expect a progressive deterioration in out-of-time performance. The absence of such degradation suggests that the fundamental drivers of electrification in Mozambique remained substantially stable throughout the period analysed

The R^2 on lit pixels only is substantially lower in both models: 0.487 in the temporal split and 0.670 in the random split. This result warrants careful discussion, as it has both methodological and substantive implications.

From a methodological standpoint, the high overall R^2 (0.785) is partly an artefact of dataset imbalance. With 99.4% of observations recording log_ntl = 0, the model achieves apparently strong performance simply by predicting the mass of dark pixel observations accurately. The R^2 test (lit) removes this structural advantage and evaluates model performance precisely where it is most challenging, on the pixels where real variance exists and needs to be explained. It is therefore a more demanding and more informative diagnostic than the overall R^2.

From a substantive standpoint, the value of 0.487 is not surprising but is informative. Lit pixels are those where electricity access already exists, and their dynamics are inherently more complex than those of dark pixels. A dark pixel is dark for structurally stable reasons: it is far from the grid, sparsely populated, and economically marginalised. A lit pixel, by contrast, can vary in NTL intensity for more contingent reasons: fluctuations in usage intensity, changes in consumption habits, supply interruptions, or seasonal patterns. These dynamics are less systematically captured by the structural variables included in the model, which are designed to explain the presence or absence of access rather than its intensity among already-electrified locations.

This result has a direct implication for the gap analysis developed in Section 5.5: because the model is less precise on already-lit pixels, the gap analysis was conducted exclusively on inhabited but dark pixels, where the model is most reliable and where the identification of underelectrified areas has the greatest operational value for energy planning.

5.3.1.1. Robustness check: model without RWI

A potential methodological concern with the inclusion of RWI as an independent variable relates to its construction. The Relative Wealth Index developed by Chi et al. (2022)(10) was built using a rich set of proxy variables derived from Facebook connectivity data, satellite imagery, and, critically, night-time light data as one of its input features. This introduces a risk of circular reasoning: if NTL data was used to construct RWI, and RWI is then used to predict NTL, part of RWI's apparent predictive power may reflect this shared information content rather than a genuine independent socioeconomic signal. In the worst case, the model would be partially predicting NTL from NTL, inflating the apparent importance of RWI and overstating the contribution of individual wealth to electricity access.

This concern was raised in the supervisory review of this study and warrants explicit empirical treatment. To assess its severity, the model was re-estimated excluding RWI from the feature set, with all other specifications (hyperparameters, validation scheme, SHAP sampling strategy) held constant. The results are reported in Table X

Table X. Robustness Check: Model Performance With and Without RWI

Metric	With RWI	Without RWI	Difference
R^2 train	0.894	0.868	−0.026
R^2 test (all pixels)	0.785	0.754	−0.030
R^2 test (lit pixels only)	0.487	0.376	−0.111
RMSE test	0.060	0.064	+0.004

The first and most important finding is that removing RWI reduces out-of-time R^2 by only 3 percentage points, from 0.785 to 0.754. This is a modest decline relative to the scale of the concern: if RWI's predictive power were primarily driven by circular NTL information, its removal would produce a much sharper performance drop. The limited magnitude of the decline suggests that RWI was contributing genuine socioeconomic information (individual wealth estimated from mobile connectivity patterns and other non-NTL proxies) rather than merely recycling the dependent variable back into the model.

The second finding concerns the redistribution of feature importance in the absence of RWI. In the model without RWI, log_pop becomes the dominant predictor with a mean SHAP value of 0.309, absorbing a large share of the importance previously attributed to RWI. HDI increases substantially from 0.047 to 0.121, suggesting that in the absence of the individual-level wealth signal, the model compensates by relying more heavily on aggregate district-level human development. Infrastructure distance variables also gain modest weight: dist_hospital_km rises from 0.027 to 0.062, dist_school_km from 0.028 to 0.046, and dist_grid_km from 0.026 to 0.039. This redistribution is consistent with the expected behaviour of a Random Forest under multicollinearity: when a correlated variable is removed, its importance is absorbed by the remaining variables that carry overlapping information.

The third and most substantively important finding is that the main conclusion of this study is fully preserved in the absence of RWI. Electricity access in Mozambique remains primarily a socioeconomic phenomenon: log_pop alone accounts for approximately 44% of total feature importance in the no-RWI model, and socioeconomic variables collectively (log_pop and HDI) account for approximately 62%. Infrastructure distance variables collectively account for less than 25%. The narrative does not change: it is not proximity to the grid that determines who has access, but the economic and demographic characteristics of the population.

The provincial R^2 values also remain broadly stable without RWI, though with some notable changes. Tete experiences the sharpest decline (from 0.800 to 0.604), consistent with the province's pattern of extractive industry-driven electrification where individual wealth is a particularly strong discriminant. Maputo, by contrast, shows minimal change (from 0.802 to 0.814), consistent with the provincial SHAP finding that HDI, not RWI, is the dominant driver in that province. Cabo Delgado remains the lowest-performing province (0.523), confirming that its anomalous performance is attributable to conflict dynamics rather than to the inclusion of RWI.

Taken together, these results support a qualified interpretation of the RWI findings in this study. The concern about circular reasoning is legitimate and cannot be entirely dismissed: a precise decomposition of how much of RWI's SHAP importance derives from its NTL component versus its non-NTL components would require access to the internal feature weights of the RWI model itself, which are not publicly available at the level of detail required. What the robustness check does establish is that the endogeneity concern does not materially alter the substantive conclusions of the study: the 3-percentage-point decline in

R² is consistent with RWI contributing genuine independent information, the main finding on the primacy of socioeconomic drivers is preserved, and the provincial heterogeneity patterns remain interpretable. Accordingly, RWI is retained in the main model, and the no-RWI specification is reported here as a robustness benchmark.

5.3.2. Provincial Performance

Model performance varies substantially across provinces, as reported in Table 4:

Table 4. Province-Level Random Forest Performance (Temporal Split, Test Set: 2022–2024)

Province	R^2 test
Sofala	0.818
Maputo	0.802
Tete	0.800
Nampula	0.766
Niassa	0.765
Gaza	0.739
Manica	0.723
Zambezia	0.720
Inhambane	0.694
Cabo Delgado	0.609

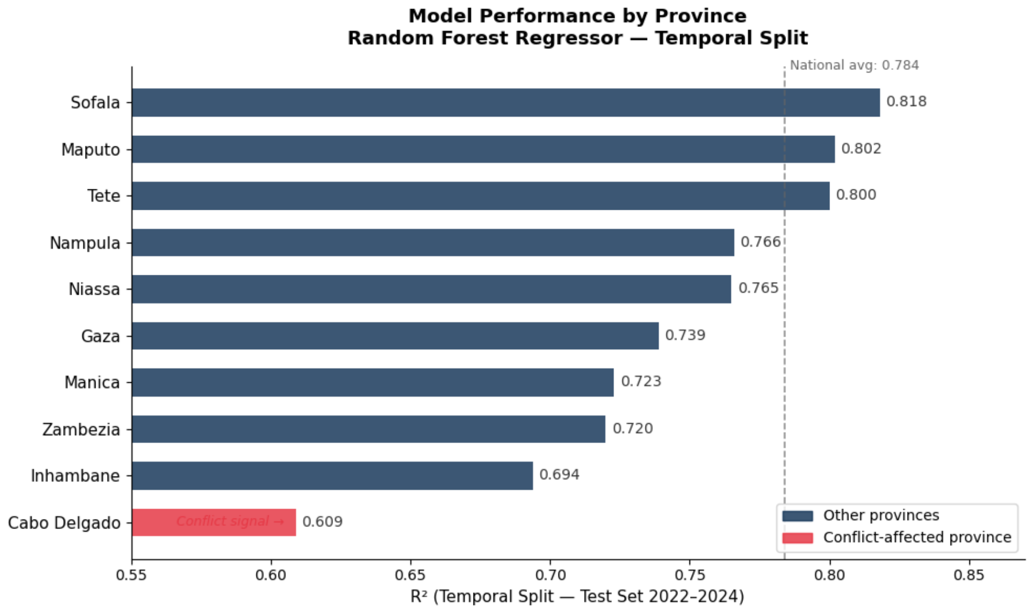


Figura 19-Random Forest model performance by province (temporal split, test set 2022–2024). Cabo Delgado is highlighted as a conflict-affected province. The dashed line indicates the national average R^2 of 0.784.

The range of R^2 values across the ten provinces spans from 0.609 to 0.818, a difference of nearly 20 percentage points that cannot be attributed to random variation. This heterogeneity indicates that the model does not perform uniformly across the national territory, and that province-specific dynamics are captured with varying degrees of accuracy. Understanding why the model performs better in some provinces than others is itself analytically informative, as it reveals where the structural relationships between socioeconomic variables and electricity access are most and least stable..

The three highest-performing provinces (Sofala (R^2 = 0.818), Maputo (0.802), and Tete (0.800)) share a common characteristic: their electricity access dynamics are dominated by relatively stable and predictable structural forces. Sofala is home to Beira, the country's second-largest port and industrial hub, where access patterns are shaped by consistent commercial and infrastructural demand. Maputo is the country's only fully urbanised province, with a mature grid and access dynamics that are strongly and stably correlated with income and population density. Tete's economy is anchored by large-scale extractive industries (coal mining and hydroelectric power) that generate predictable patterns of electrification around industrial corridors. In all three cases, the structural variables in the model (RWI, log_pop, dist_grid_km) capture the dominant drivers of access with high fidelity.

The middle-performing provinces (Nampula (0.766), Niassa (0.765), Gaza (0.739), Manica (0.723), and Zambezia (0.720)) represent the heterogeneous interior of Mozambique, where access dynamics are shaped by a more complex mix of rural poverty, geographic remoteness, and partial grid coverage. The model performs adequately in these contexts but captures less of the total variance, likely because the relationships between structural predictors and access are more spatially heterogeneous within these large and diverse provinces.

Inhambane (0.694) occupies an intermediate position, reflecting its mixed coastal and inland character: coastal areas with tourism-driven economic activity present different access dynamics from the more remote inland districts, creating within-province heterogeneity that the model partially captures but does not fully resolve.

Cabo Delgado stands out as a clear negative outlier, with an R^2 of 0.609, approximately 17 percentage points below the best-performing province. This province has been the theatre of an armed jihadist insurgency since 2017, concentrated in the northern coastal areas around Mocímboa da Praia and Palma. The existing literature documents that armed conflict has a systematic negative impact on electricity consumption, production, and access rates in sub-Saharan Africa, operating through multiple channels(16,27): direct destruction of infrastructure, displacement of population, disruption of supply chains, and the withdrawal of private and public investment from conflict-affected areas. In Cabo Delgado, these dynamics have fundamentally altered the structural relationships between socioeconomic variables and electricity access that the model learns from the training data. A pixel with high RWI and proximity to the grid in 2015 may have lost access entirely by 2020 due to conflict-related infrastructure destruction, a discontinuity that the model's structural variables cannot capture. The low R^2 in Cabo Delgado is therefore not a model failure in the conventional sense, but indirect empirical evidence of the conflict's disruptive impact on the normal determinants of electrification. This interpretation is developed further in the SHAP provincial analysis in Section 5.4.2 and in the discussion of limitations in Section 6.3.

5.4. SHAP Analysis – Drivers of Access

5.4.1. Global Feature Importance

SHAP (SHapley Additive exPlanations) values quantify the marginal contribution of each feature to model predictions in an additive and theoretically grounded manner(13). Unlike traditional importance measures based on Gini impurity, which tend to overestimate the importance of high-cardinality variables, SHAP values provide a prediction decomposition that is both consistent and locally accurate: the sum of all SHAP values for a given observation equals the difference between the model's prediction and the global mean prediction, guaranteeing full additivity.

The results of the global SHAP analysis are reported in Table 5.

Table 5. Global SHAP Feature Importance (Temporal Split Model)

Feature	Mean SHAP	Relative Importance
RWI	0.338	49.8%
log_pop	0.162	23.9%
HDI	0.047	6.9%
NDVI	0.029	4.3%
dist_school_km	0.028	4.1%
dist_hospital_km	0.027	4.0%
dist_grid_km	0.026	3.8%

Feature	Mean SHAP	Relative Importance
dist_market_km	0.010	1.5%
conflict_violence	0.009	1.3%
builtup	0.003	0.4%

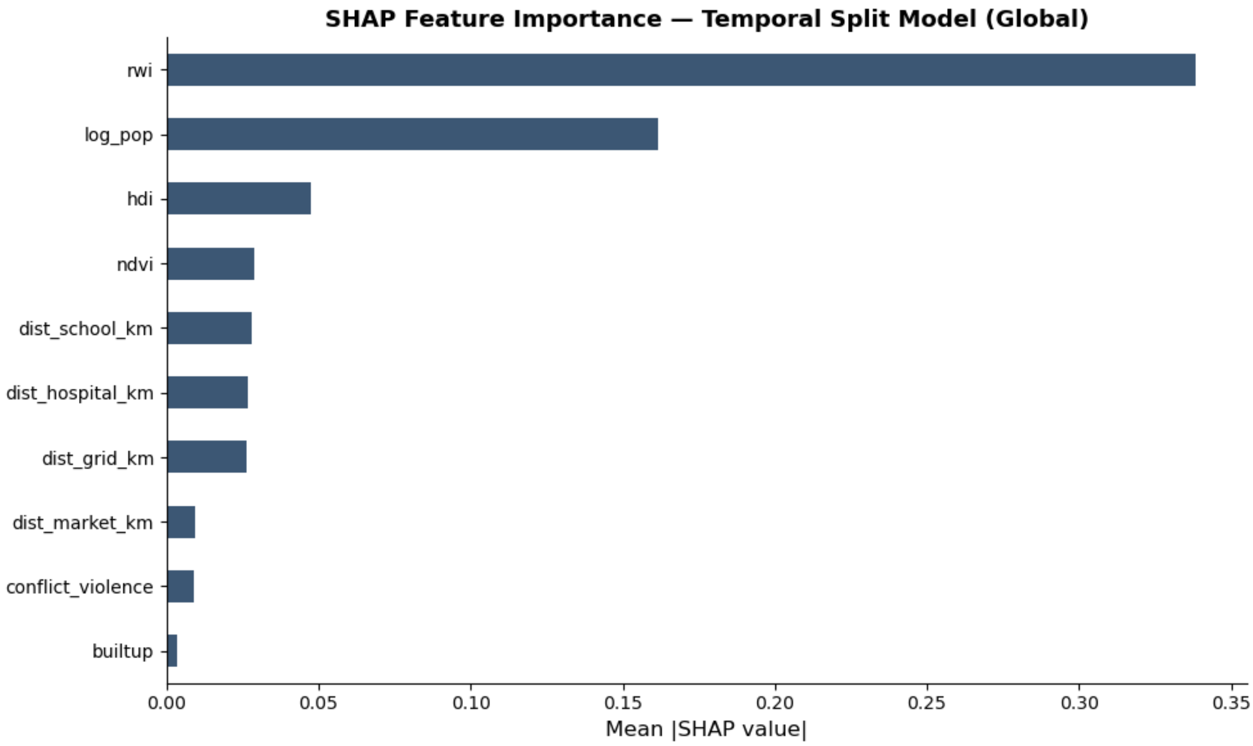


Figura 20- Global SHAP feature importance (temporal split model). Values represent mean absolute SHAP contributions computed on a balanced sample of 5,000 lit and dark pixel-year observations from the training set.

The most striking finding is the clear dominance of RWI (mean SHAP = 0.338, relative importance 49.8%) and log_pop (0.162, 23.9%), which together account for approximately 74% of total feature importance. This result has a substantive implication of the first order: electricity access in Mozambique is primarily a socioeconomic phenomenon. It is not infrastructure that determines who has access, but income and population density.

This finding stands in apparent tension with the infrastructure-first narrative that dominates policy debates on electrification in sub-Saharan Africa, where grid distance is frequently presented as the binding constraint. In this study's dataset, dist_grid_km has a SHAP importance of only 0.026, lower than both dist_school_km (0.028) and dist_hospital_km (0.027). This does not imply that grid distance is causally irrelevant. Rather, it reflects a structural collinearity: pixels far from the grid tend also to be poor and sparsely populated, and it is these socioeconomic characteristics, not distance per se, that the model identifies as the primary determinants of electrification status. Controlling for income and population, grid distance adds relatively little marginal explanatory power. The policy implication is significant: interventions that focus exclusively on grid extension without addressing the income constraints that prevent connection even where the grid is physically present may prove less effective than anticipated.

HDI shows an aggregate SHAP importance of 0.047, higher than any individual infrastructure distance variable but substantially lower than RWI. This result does not contradict the extensive literature documenting the strong relationship between human development and energy access at the cross-country level(4,28). Rather, it reflects a scale and resolution effect: HDI in this dataset is assigned at the district level and therefore captures between-district variation in development, smoothing out the within-district heterogeneity that is central to understanding who within a community has access and who does not. RWI, by contrast, is available at sub-district resolution and captures individual relative wealth at a spatial granularity closer to the household level. In a context where electricity access is determined at the household level, by whether a specific family can afford the connection cost, a finer-grained measure of individual wealth is expected to outperform a coarser district-level development index in a pixel-level predictive model. This is a resolution and aggregation effect, not a substantive claim that wealth matters more than human development in general.

NDVI (0.029) ranks fourth, reflecting its dual role as a proxy for rurality and land cover heterogeneity. Its relatively modest importance suggests that, once socioeconomic variables are controlled for, vegetation density adds limited independent explanatory power at the national level, though its role may vary provincially, as discussed in Section 5.4.2.

The conflict_violence variable shows a global SHAP importance of 0.009, which may appear marginal but requires careful interpretation. Armed conflict in Mozambique is geographically concentrated in Cabo Delgado, which represents a small fraction of the total dataset. SHAP measures the marginal contribution of a variable to the variation in predictions within the sample: if conflict_violence has low within-sample variance, because the vast majority of pixel-year observations record zero conflict exposure, its global SHAP value will be low regardless of its actual causal impact on electricity access in affected areas. The global importance of conflict_violence therefore underestimates its localised relevance. Its true weight becomes visible only in the provincial SHAP disaggregation, where the anomalous profile of Cabo Delgado reveals the extent to which conflict disrupts the normal structural relationships between socioeconomic variables and access.

5.4.2. Provincial Heterogeneity

The provincial SHAP analysis is the most original scientific contribution of this study. By computing SHAP values separately for each of Mozambique's ten provinces, it is possible to test empirically whether the drivers of electrification are homogeneous across the national territory or whether they vary geographically in a systematic and interpretable manner. The results are reported in Table 6, where values represent the mean SHAP per feature and province, computed on a balanced lit/dark sample of 1,500 observations per province from the 2022–2024 test set.

Table 6. Province-Level Mean SHAP Values by Feature (Temporal Split Model)

Province	log_pop	hdi	rwi	ndvi	dist_grid	dist_hospital	dist_school	conflict
Cabo Delgado	0.156	0.009	0.157	0.020	0.017	0.020	0.009	0.002
Gaza	0.120	0.012	0.131	0.015	0.017	0.014	0.010	0.002
Inhambane	0.088	0.012	0.142	0.015	0.021	0.016	0.013	0.002
Manica	0.114	0.011	0.151	0.015	0.013	0.013	0.013	0.002
Maputo	0.125	0.241	0.342	0.017	0.034	0.013	0.027	0.013
Nampula	0.145	0.009	0.167	0.022	0.021	0.019	0.010	0.004
Niassa	0.164	0.011	0.208	0.021	0.022	0.045	0.013	0.003
Sofala	0.129	0.012	0.238	0.019	0.017	0.019	0.019	0.003
Tete	0.101	0.009	0.222	0.030	0.023	0.032	0.024	0.003
Zambezia	0.147	0.013	0.189	0.023	0.021	0.019	0.012	0.003

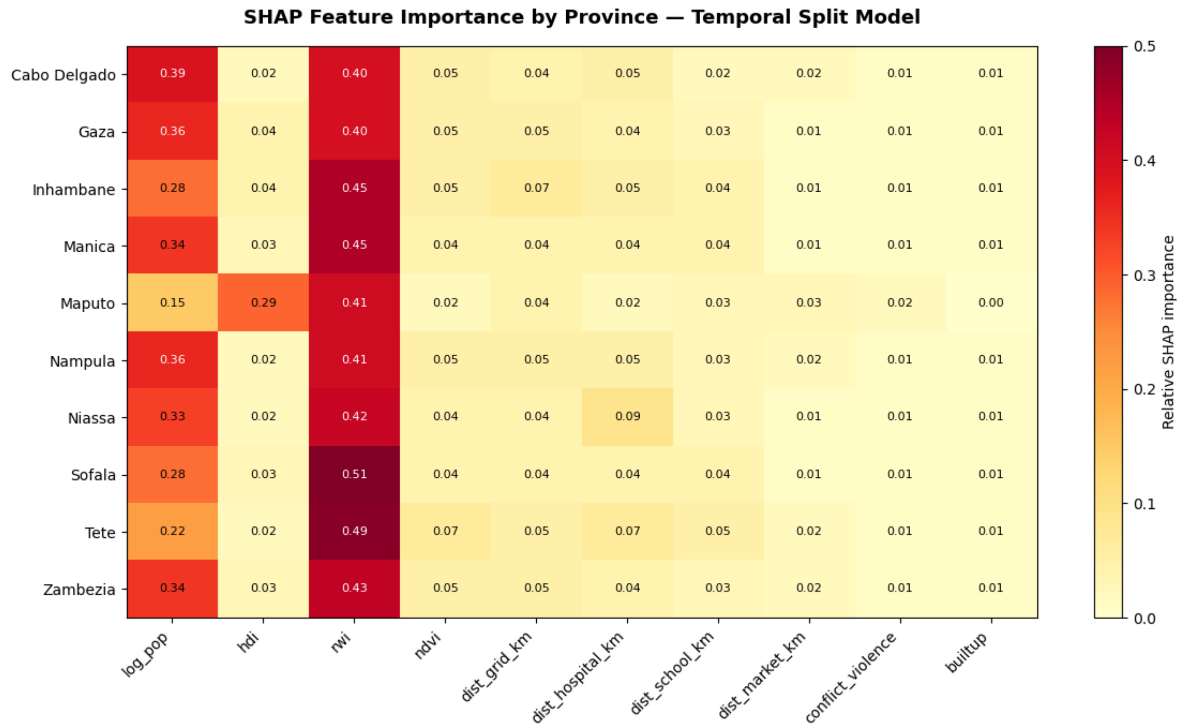


Figura 21- Province-level SHAP feature importance heatmap (temporal split model). Values represent normalised relative importance within each province. Colour intensity reflects the relative contribution of each feature to provincial predictions.

Figure 22 illustrates the spatial distribution of the six key explanatory variables across the Mozambican territory. The maps confirm the extreme geographic heterogeneity that motivates the provincial SHAP disaggregation: RWI, population density, HDI, and infrastructure distances all vary dramatically across provinces, while conflict exposure is concentrated in the north but not absent from the centre and south. This heterogeneity is precisely what the provincial SHAP analysis quantifies in systematic terms, identifying which variables drive electricity access differently in each provincial context.

Three patterns emerge clearly from the data and warrant explicit discussion.

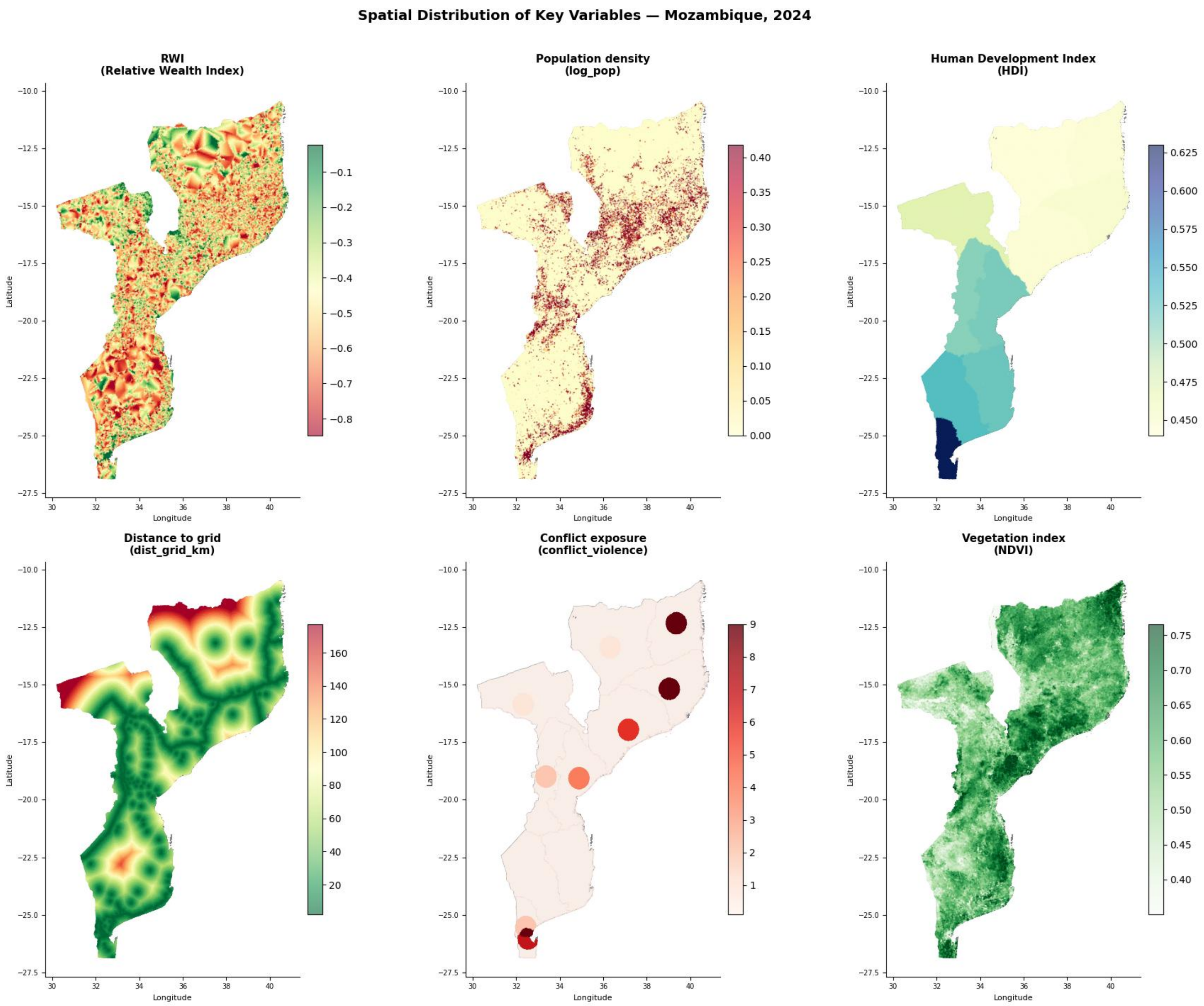


Figure 22-Spatial distribution of key explanatory variables across Mozambique.

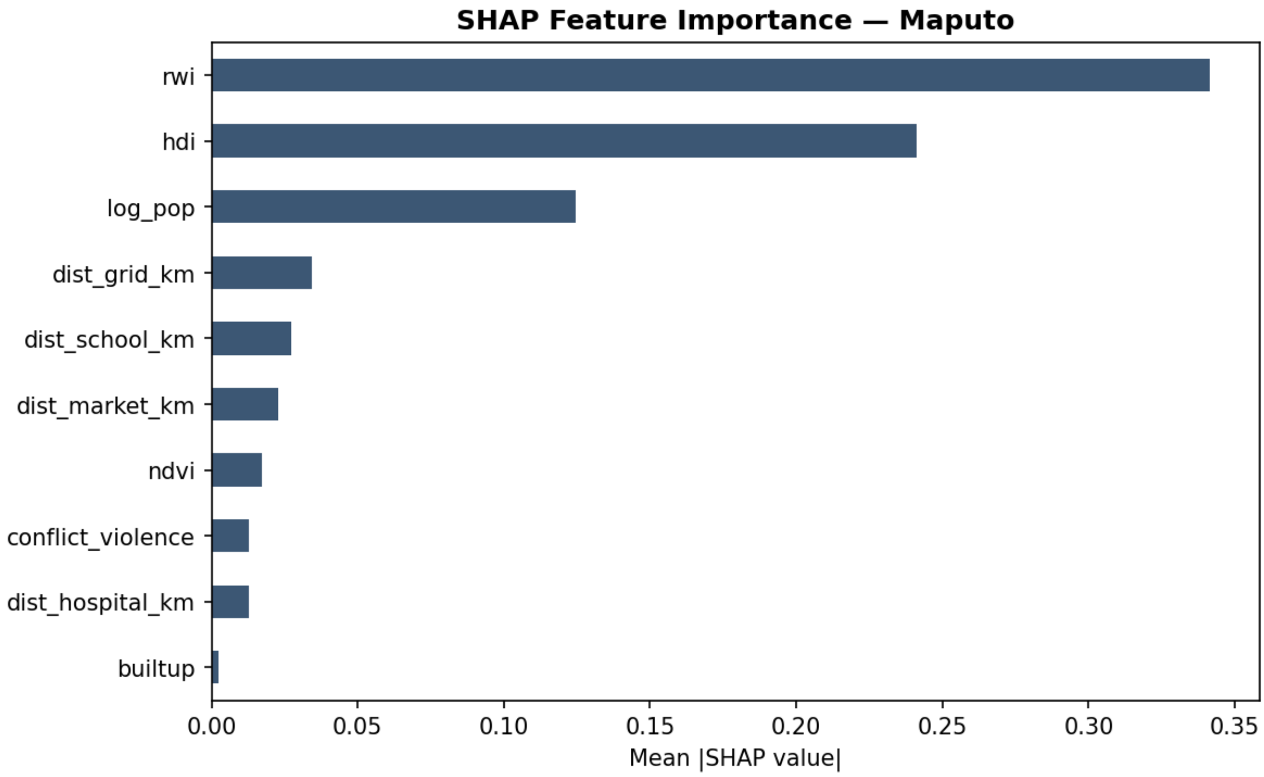


Figura 23-SHAP feature importance for Maputo province. Note the anomalously high contribution of HDI relative to all other provinces.

Maputo presents the most anomalous SHAP profile of all provinces. HDI reaches a mean SHAP value of 0.241, a figure that in no other province exceeds 0.013. This difference is of an order of magnitude and cannot be attributed to sampling variation. Maputo is the only province where aggregate human development carries a weight comparable to individual income in determining electricity access. RWI is also the highest in absolute terms (0.342), confirming that Maputo is the context where both socioeconomic dimensions, individual wealth and aggregate development, are most discriminating. This is consistent with Maputo's character as the country's only fully urbanised and institutionally developed province. In a context where basic infrastructure is already widespread and grid coverage approaches universality, the variable that differentiates who has access from who does not is no longer raw income alone, but the overall level of development of the surrounding environment (quality of public services, institutional density, formal employment) all dimensions captured more by HDI than by RWI.

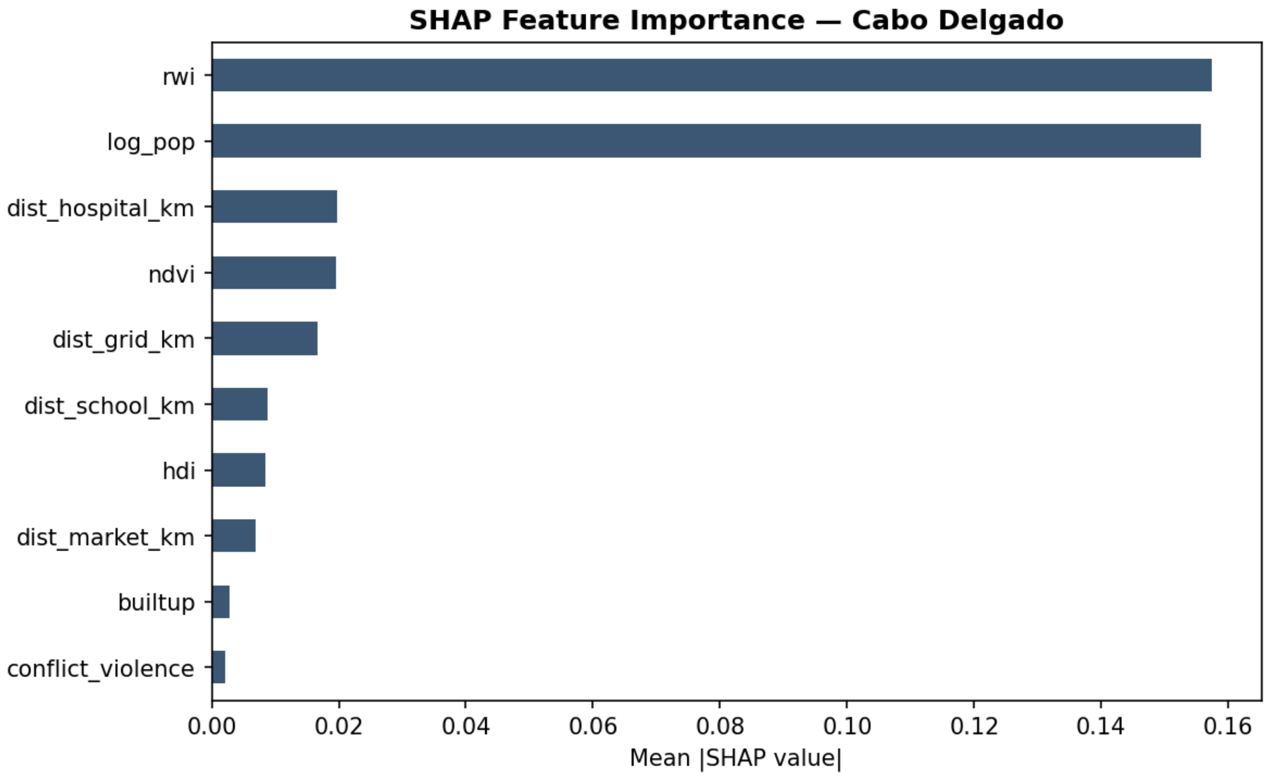


Figura 24-SHAP feature importance for Cabo Delgado province. Note the near-identical weights of RWI and log_pop, consistent with conflict-driven disruption of normal socioeconomic access relationships.

Cabo Delgado is the only province where RWI and log_pop carry nearly identical weights (0.157 vs. 0.156). In all other provinces, RWI exceeds log_pop more clearly, reflecting the national pattern of income dominance documented in the global analysis. This anomalous profile is consistent with the interpretation developed in Section 5.3.2: in a context of prolonged armed conflict, the normal structural relationships between income and electricity access are disrupted. Population concentration, often determined by displacement dynamics rather than economic activity, acquires greater relative weight, while individual wealth loses some of its discriminating power because conflict destroys the infrastructure that would normally translate wealth into access.

A further paradox merits attention: conflict_violence shows a SHAP value of 0.002 in Cabo Delgado, the lowest of all provinces. This might appear counterintuitive given that conflict is widely recognised as the dominant obstacle to electrification in that province. The explanation is methodological. SHAP measures the marginal contribution of a variable to the variation in predictions within the sample. If conflict_violence has low intra-provincial variance ,that is, if conflict is pervasive and approximately uniformly distributed across the province rather than concentrated in specific areas, its SHAP value will be low regardless of its actual causal impact on electricity access. Conflict in Cabo Delgado does not appear as a SHAP driver because it functions as a constant, not a variable. The indirect evidence of its impact is precisely the low provincial R^2 (0.609) documented in Section 5.3.2: the model struggles to explain access in Cabo Delgado because conflict, which cannot be identified as a discriminating intra-provincial variable, is the dominant driver.

Niassa presents dist_hospital_km with a mean SHAP value of 0.045, nearly double the value observed in any other province, where the figure does not exceed 0.032. This pattern is directly observable in the data and does not require causal interpretation. In Niassa, the distance to the nearest hospital contributes to model predictions significantly more strongly than in the rest of the country. The most plausible interpretation is that in contexts of extreme geographic remoteness, distance to health facilities functions as a proxy for general geographic marginalisation: a pixel far from the nearest hospital in Niassa is likely far from everything: markets, schools, roads, and the electricity grid. In more densely served provinces, distance to health facilities and distance to the grid are more loosely correlated, reducing the former's marginal predictive power. In Niassa, the two are so tightly co-located that hospital distance becomes a synthetic indicator of infrastructure access in general.

The three patterns described above demonstrate empirically that the drivers of electricity access vary geographically in a systematic and interpretable manner. The global SHAP analysis, which produces a single national feature importance hierarchy, conceals this heterogeneity. A uniform national electrification policy based on average national drivers would be appropriate for some provinces but potentially misleading for others. Maputo requires attention to aggregate human development beyond individual income. Cabo Delgado requires a separate analytical and policy framework that accounts for conflict dynamics before standard electrification instruments can be applied. Niassa reveals a relationship with health infrastructure that does not emerge in the aggregated national analysis and that points to the need for integrated service delivery rather than isolated energy interventions. These findings motivate the differentiated policy recommendations developed in Section 6.2.

5.5. Gap Analysis - Priority Areas for Intervention

The gap analysis identifies pixels where the model predicts a higher level of electricity access than currently observed. For each pixel in the test set, the gap is defined as::

$$gap = y_{pred} - y_{true}$$

A positive gap indicates that a pixel possesses structural characteristics (income, population density, proximity to infrastructure) compatible with a higher level of electricity access than what is currently observed. These pixels represent areas where access is lower than socioeconomic conditions would lead us to expect, and where the barrier to electrification is therefore not structural but operational: the conditions for access exist, but access has not yet been achieved.

Starting from pixels with a positive gap, a set of actionability criteria is applied to identify priority intervention areas. A pixel is classified as a low-hanging fruit (LHF) if it simultaneously satisfies the following four conditions in 2024:

- it is currently dark: $log\_ntl = 0$
- the model predicts light: $y_{pred} > 0.01$
- it is inhabited: $\log\_pop > 0$
- it is not located in an active conflict zone: $conflict_{violence} = 0$

These criteria translate the concept of unrealised potential into an operational definition that can be directly used for intervention planning. Pixels satisfying all four conditions are areas where structural conditions are favourable for electrification, access has not yet been achieved, and security conditions permit infrastructure works.

The threshold $y_{pred} > 0.01$ deserves explicit justification. A stricter threshold would reduce the number of identified pixels but increase confidence that the model's prediction reflects a genuine structural signal rather than noise. A more permissive threshold would identify more candidates but at the cost of including pixels where the model's prediction is marginal and uncertain. The value of 0.01 was chosen as a conservative operational threshold that filters out near-zero predictions while retaining pixels where the model identifies a meaningful structural advantage.

In the 2024 test set, out of 617,198 inhabited pixels:

- 25,259 are already electrified ($y_{true} > 0$), representing 4.1% of inhabited pixels
- 58,344 are predicted to be electrified by the model ($y_{pred} > 0.01$), representing 9.5%
- 26,874 satisfy all low-hanging fruit criteria, representing 4.35% of inhabited pixels

The gap between the share of pixels predicted to be lit (9.5%) and those actually lit (4.1%) is itself a summary measure of the scale of unrealised electrification potential in Mozambique: the model identifies structural conditions consistent with electrification in more than twice as many pixels as are currently served.

The provincial distribution of LHF pixels is reported in Table 7:

Table 7. Low-Hanging Fruit Pixels by Province (2024 Test Set)

Province	LHF Pixels	Avg. Predicted NTL	Avg. Grid Distance (km)	Avg. Population	Avg. RWI
Tete	6,891	0.11	66.2	6.1	-0.15
Zambezia	4,860	0.08	20.1	4.5	-0.04
Maputo	4,127	0.15	20.8	0.8	-0.44
Nampula	1,996	0.08	21.6	10.5	-0.19
Cabo Delgado	1,840	0.14	26.0	17.9	-0.21
Inhambane	1,705	0.05	16.3	2.8	-0.07
Gaza	1,646	0.06	15.4	4.8	-0.15
Sofala	1,475	0.08	19.1	5.5	-0.11
Niassa	1,241	0.10	40.8	7.1	-0.11
Manica	1,093	0.07	11.2	4.2	-0.16

Note: LHF pixels are inhabited (log_pop > 0), currently dark (log_ntl = 0), predicted to be lit by the model (y_pred > 0.01), and located outside active conflict zones (conflict_violence = 0).

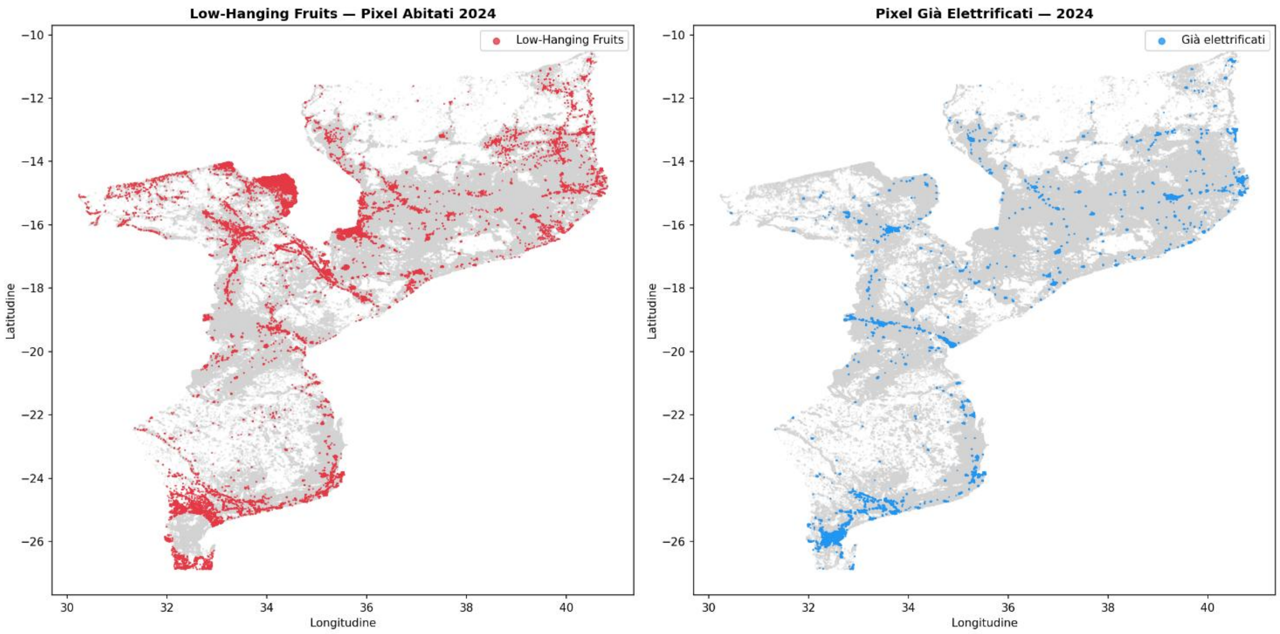


Figure 25-Geographic distribution of low-hanging fruit pixels (left panel, red) and already-electrified pixels (right panel, blue) across Mozambique, 2024. Grey points represent inhabited pixels not classified as LHF.

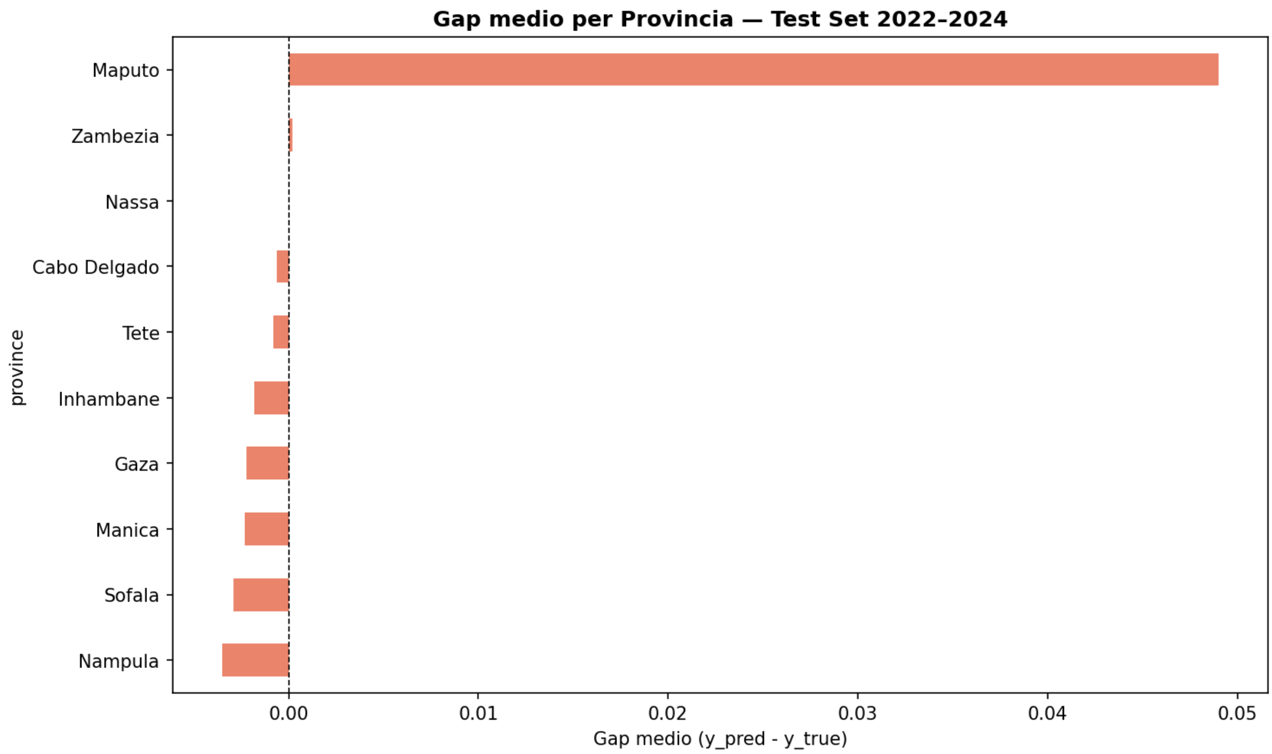


Figure 26-Average gap (y_pred − y_true) by province, test set 2022–2024.

Three distinct clusters emerge from the data, each with different policy implications.

Cluster 1 – Immediate grid extension

Zambezia, Nampula, Gaza, Inhambane, Sofala, and Manica present average grid distances ranging from 11 km (Manica) to 21 km (Nampula). These pixels are relatively close to existing infrastructure and have sufficient population density to justify grid extension investment. They represent the most immediately actionable candidates for short-term intervention. The SHAP profiles of these provinces, where RWI is the dominant driver and infrastructure distances carry secondary weight, suggest that the binding constraint is not physical remoteness from the grid but the economic capacity of households to afford connection costs. This implies that physical grid extension alone may be insufficient: it should be accompanied by instruments that reduce connection costs for low-income households, such as connection subsidies, instalment payment plans, or income-adjusted social tariffs..

Cluster 2 – Urban last-mile

Maputo presents 4,127 LHF pixels with an average grid distance of 20.8 km, comparable to Cluster 1 provinces, but an average RWI of −0.44, the lowest of all provinces. These pixels are most likely informal and peri-urban settlements where the electricity grid is physically accessible but connection has not occurred, plausibly for economic reasons. The last-mile urban phenomenon is well-documented in the energy access literature: in the peripheries of large African cities, the grid often reaches areas physically but the poorest households cannot afford connection costs, which in many SSA countries can equivalent to several months of household income(18). In this context, extending the grid further does not solve the problem: the constraint is economic, not infrastructural. The appropriate policy instruments for this cluster are therefore fundamentally different from those of Cluster 1: direct connection cost subsidies, concessional financing programmes, reduced connection tariffs for informal settlements, and prepayment solutions that lower the entry barrier are the most relevant tools.

Cluster 3 – Off-grid and specific cases

Tete presents the largest number of LHF pixels (6,891) but with an average grid distance of 66.2 km, substantially higher than all other provinces. At these distances, conventional grid extension is economically difficult to justify for communities with the income levels observed in these areas(29). The appropriate solutions for this cluster are individual off-grid systems, primarily solar home systems, or community mini-grids powered by renewable sources. Niassa (40.8 km average grid distance) occupies an intermediate position and may be served by a combination of mini-grid solutions and targeted grid extension where population clusters are sufficiently dense.

Cabo Delgado presents 1,840 LHF pixels with the highest average population per pixel of all provinces (17.9 inhabitants), indicating significant demographic potential. However, as discussed in Sections 5.3.2 and 5.4.2, the conflict context renders these pixels not immediately actionable with standard infrastructure instruments. The recommendation for Cabo Delgado is sequential: political stabilisation and security restoration must precede infrastructure investment. In the short term, community-driven approaches involving local actors and small-scale off-grid systems, less vulnerable to destruction than grid infrastructure, may represent a transitional solution for the most exposed communities.

The gap analysis is subject to three methodological limitations that must be made explicit. First, the model is trained on 2015–2021 data and applied to 2024: the pixels identified as LHF reflect the structural conditions of the training period, which may not be fully updated to 2024, particularly in areas that have undergone significant change such as post-conflict reconstruction zones or areas affected by recent infrastructure expansion. Second, the threshold y_pred > 0.01 is operationally justified but ultimately arbitrary: variations in this threshold produce different priority lists, and a systematic sensitivity analysis would be necessary to assess the robustness of the ranking. Third, the gap analysis does not account for intervention costs: a LHF pixel close to the grid is not necessarily cheaper to serve than a more remote one if terrain conditions, settlement characteristics, or land tenure constraints are unfavourable. Integration with geospatial cost modelling tools such as OnSSET(30) would transform the priority list from a potential-based classification into a cost-benefit ranking, substantially increasing its operational utility for decision-makers..

6. Discussion and Conclusions

6.1. Answer to the Research Question

Before presenting the analytical findings, it is important to acknowledge the structural complexity of the context in which this study is situated. Mozambique's electricity access deficit is not a purely technical or infrastructural problem, it is embedded in a multi-layered reality that encompasses deep socio-political fragility, persistent ethnic and religious tensions, endemic corruption at multiple levels of government, extreme institutional weakness, and the largely insufficient response of multilateral institutions to the country's structural challenges. This study adopts a socio-technical systems engineering approach and does not engage with the full political, diplomatic, and institutional complexity of the Mozambican context. Its results and recommendations should therefore be read as contributions to evidence-based energy planning within this complex reality, not as a comprehensive assessment of all the conditions that ultimately determine the feasibility of any intervention.

This study set out to answer the following research question: what factors explain disparities in electricity access in Mozambique over the period 2015–2024?

Table 8. Summary of Principal Findings

Finding	Key result
Drivers of access	RWI and log_pop explain ~74% of total feature importance. Electricity access is primarily a socioeconomic phenomenon.
Geographic heterogeneity	HDI is the dominant driver in Maputo; RWI ≈ log_pop in Cabo Delgado consistent with conflict-driven disruption; dist_hospital_km unusually high in Niassa.
Structural stability	R^2 stable across 2022–2024 (0.796, 0.775, 0.785) with no degradation trend, confirming stable structural relationships between drivers and access.

The results converge on a clear answer: electricity access in Mozambique is primarily a socioeconomic phenomenon. The Relative Wealth Index (RWI) is the dominant predictor, with a mean SHAP value of 0.338, approximately double that of the second predictor, log_pop (0.162). Together, these two factors account for approximately 74% of total model feature importance. Infrastructure variables, in particular dist_grid_km (0.026), carry significantly less weight than the infrastructure-first narrative would suggest.

This result must however be interpreted with methodological caution. The fact that dist_grid_km has low SHAP importance does not imply that grid distance is causally irrelevant. In a predictive framework such as the Random Forest, highly correlated variables tend to share importance, and dist_grid_km is structurally correlated with both RWI and log_pop: areas far from the grid also tend to be poor and sparsely populated. The model allocates importance to the most marginally informative variables, not necessarily to the causally determining ones. A more cautious interpretation is therefore the following: in the Mozambican context, income and population density are more informationally powerful predictors of electricity access than grid distance alone, which suggests that socioeconomic constraints are at least as relevant as purely infrastructural ones. The two types of constraint are not in opposition; they are co-present and mutually reinforcing. However, the relative weight of evidence points to socioeconomic barriers as the primary binding constraint.

This interpretation is further supported by the robustness check excluding RWI from the feature set, which reduces R^2 by only 3 percentage points (from 0.785 to 0.754) while leaving the main finding intact: log_pop becomes the dominant predictor in the absence of RWI, and socioeconomic variables collectively continue to account for the majority of feature importance (Section 5.3.1). This confirms that the primacy of socioeconomic drivers is not an artefact of RWI's construction methodology.

This interpretation carries a direct policy implication: strategies that focus exclusively on grid extension, without addressing the income constraints that prevent connection even where the grid is physically present, risk being less effective than anticipated. The 26,874 LHF pixels identified in Section 5.5 illustrate this point concretely: these inhabited pixels possess structural characteristics favourable to electrification but remain unserved, and in many cases, as in the peri-urban pixels of Maputo with an average RWI of −0.44, the grid is physically nearby but connection has not occurred. The barrier is economic, not infrastructural.

The second principal finding is the geographic heterogeneity of access drivers. The provincial SHAP analysis demonstrates that the factors explaining electricity access vary systematically across provinces in ways that a single national model cannot fully capture.

Maputo is the only province where HDI carries a weight comparable to RWI (0.241 vs. 0.342). In no other province does HDI exceed 0.013, a difference of nearly an order of magnitude. This indicates that in the capital, the country's only mature urban context, aggregate human development is a discriminant of electricity access beyond individual income. In a context where basic infrastructure is already widespread, the variable that differentiates who has access from who does not is no longer raw income alone but the overall level of development of the surrounding environment, including institutional quality, public service coverage, and formal employment, all dimensions captured more by HDI than by RWI.

Cabo Delgado presents the most anomalous profile. It is the only province where RWI and log_pop carry nearly identical weights (0.157 vs. 0.156), while in all other provinces RWI exceeds log_pop more clearly. This profile is consistent with the structural disruptions caused by the armed conflict that has afflicted the province since 2017(15,16): in a context of prolonged conflict, the normal relationships between income and electricity access are distorted, and the demographic dimension, specifically the concentration of population in particular locations often determined by displacement dynamics rather than economic activity, acquires greater relative weight.

An apparent paradox is worth addressing explicitly: conflict_violence has a SHAP value of only 0.002 in Cabo Delgado, the lowest of all provinces. This might seem counterintuitive given that conflict is widely recognised as the dominant obstacle to infrastructural development in that province. The explanation is methodological: SHAP measures the marginal contribution of a variable to the variation in predictions within the sample. If conflict_violence has low intra-provincial variance, meaning the conflict is pervasive and approximately uniformly distributed across the province, its SHAP value will be low regardless of its actual causal impact on electricity access. Conflict in Cabo Delgado does not emerge as a SHAP driver because it functions as a constant rather than a variable. The indirect evidence of its impact is precisely the low provincial R^2 (0.609): the model struggles to explain access in Cabo Delgado because conflict, which cannot be identified as a discriminating intra-provincial variable, is the dominant driver.

Niassa presents dist_hospital_km with a SHAP value of 0.045, nearly double the value observed in any other province. This pattern suggests that in contexts of extreme geographic remoteness, distance to health facilities functions as a proxy for general geographic marginalisation: a pixel far from the nearest hospital in Niassa is likely far from everything, including the electricity grid. This is an example of how proxy variables can carry different meanings in different geographic contexts, a result that emerges only from the disaggregated provincial analysis.

The northern provinces (Nampula, Zambezia, and Niassa) show a slightly higher weight for dist_grid_km compared to the central and southern provinces. This is consistent with the lower infrastructure density of northern Mozambique: where the grid is sparser, distance from it retains greater explanatory power compared to contexts where the grid is more capillary and distance varies less across pixels.

The third principal finding concerns the predictability and structural stability of electricity access over time. The model trained on 2015–2021 data explains 78.5% of the variance in the 2022–2024 test set, with stable performance across the three test years (R^2 of 0.796, 0.775, and 0.785 for 2022, 2023, and 2024 respectively). The absence of performance degradation over time is a substantive finding, not merely a methodological one.

It indicates that the structural relationships between socioeconomic variables and electricity access in Mozambique remained substantially stable between the training period and the test period. The drivers of access have not changed significantly: RWI and population density continue to be the dominant predictors in the same order of magnitude. Had a major structural shift occurred, for instance a large-scale electrification policy that altered the relationship between income and access, or an infrastructure expansion that reduced the explanatory weight of grid distance, we would expect progressive deterioration in out-of-time performance. The absence of such deterioration suggests that Mozambique has not undergone structural transformations in electricity access large enough to alter the fundamental drivers during the period analysed.

The comparison between the temporal split ($R^2 = 0.785$) and the random split ($R^2 = 0.881$) quantifies the cost of temporal generalisation: approximately 10 percentage points of R^2 are lost when the model is asked to predict future years rather than years mixed into the training set. This gap carries a broader methodological implication: studies that evaluate their models using random splits on temporal panel data may systematically overestimate real predictive capacity. The temporal split, while producing apparently lower metrics, is the methodologically correct evaluation for any policy application that requires forecasts over future periods.

A final result deserving explicit attention is the R^2 on lit pixels only: 0.487 in the temporal split. This value, substantially lower than the overall R^2 of 0.785, indicates that the model captures variance less well in already-electrified pixels than in the full dataset. The dynamics of lit pixels are inherently more complex: an electrified pixel can vary in NTL intensity for contingent reasons, including fluctuations in usage intensity, supply interruptions, changes in consumption habits, and seasonal patterns, none of which are systematically linked to the structural variables included in the model. This result points to an intrinsic limitation of the approach: the framework developed here is effective at identifying where access exists or does not exist, but is less suited to modelling the quantitative nuances of access intensity in already-served pixels.

6.1.1. Positioning relative to existing electrification planning tools

The framework developed in this study differs from existing geospatial electrification planning tools, most notably OnSSET(30) and its derivatives(29), in both its analytical objectives and its methodological approach. Understanding these differences is important for positioning the contribution of this study within the broader energy access literature.

OnSSET and similar least-cost electrification models are optimisation tools: given a set of techno-economic parameters (grid extension costs per km, mini-grid capital costs, solar irradiance, population demand projections) they identify the least-cost technology mix to achieve a defined electrification target at a given horizon. Their output is a technology assignment (grid, mini-grid, or standalone system) for each settlement or pixel, together with an associated cost estimate. These tools are primarily normative: they answer the question "what is the optimal way to electrify a given area?"

The framework developed in this study is instead a predictive and explanatory tool. It does not assume a techno-economic cost structure or optimise an objective function. Instead, it learns from nine years of observed NTL data, what actually happened on the ground, to identify which structural characteristics are most associated with electricity access across the Mozambican territory, and where structural conditions are already favourable for electrification but access has not yet been achieved. Its output is not a technology assignment but a feature importance hierarchy and a geographic priority list grounded in observed patterns rather than modelled assumptions.

This distinction generates three specific advantages relative to OnSSET-type tools.

First, the approach is data-driven rather than assumption-driven. OnSSET requires a large number of input parameters (grid extension cost per km, population density thresholds, demand tiers, technology cost curves) that are often uncertain or unavailable in data-scarce contexts like Mozambique. The Random Forest framework requires only the variables included in the dataset, all of which are publicly available, and does not impose any parametric structure on the relationships between them.

Second, the SHAP analysis produces an interpretable decomposition of access drivers that OnSSET cannot provide. OnSSET tells you what technology to use where; it does not tell you why access exists or does not exist in structural terms. The provincial SHAP disaggregation developed in this study reveals that drivers vary systematically across provinces , a finding with direct implications for policy design that no cost-optimisation model can generate.

Third, the framework is updatable annually at minimal cost. New VIIRS NTL data becomes available every year, and re-running the gap analysis requires only updating the dependent variable, no re-calibration of cost curves or demand projections is needed. This makes the framework suitable as a dynamic monitoring tool for tracking progress toward Mozambique's 2030 electrification targets.

The two approaches are therefore complementary rather than competing. OnSSET provides the techno-economic costing that this study lacks; this study provides the data-driven driver analysis and priority ranking that OnSSET does not produce.

6.2. Policy Implications for Energy Planning

The results of this study do not exhaust themselves in an academic contribution, but produce an output directly usable for energy planning in Mozambique. The gap analysis has identified 26,874 inhabited pixels where structural conditions are favourable to electrification but access has not yet been achieved. These pixels represent the areas where an investment in grid extension or off-grid solutions has the highest probability of short-term success, precisely because the underlying socioeconomic conditions are already present.

The overarching methodological implication emerging from the provincial SHAP analysis is that no single solution exists for Mozambique. Access drivers vary geographically in a systematic manner, and this implies that intervention strategies must be calibrated to the local characteristics of each province. A uniform national policy focused exclusively on grid extension or income subsidies would be appropriate for some provinces but ineffective or even counterproductive for others.

The provincial SHAP analysis and the gap analysis together demonstrate that the drivers of electricity access and the appropriate intervention strategies differ systematically across Mozambique's provinces. However, translating these analytical results into effective policy action requires recognising that responsibilities for electrification are distributed across multiple levels of governance, from international multilateral cooperation to local community implementation, and that the actors, instruments, and timelines appropriate at each level are fundamentally different. Table 9 summarises these multi-level policy implications and the key stakeholders responsible for implementation at each level.

Table 9. Multi-level policy implications, stakeholder responsibilities, and implementation instruments.

Level	Policy implication	Key actors	Instruments
International / Multilateral	Mobilise concessional finance for grid extension and off-grid solutions; coordinate donor support; address institutional fragility and governance failures; support conflict resolution in Cabo Delgado through multilateral mechanisms.	World Bank, African Development Bank, IEA, USAID Power Africa, EU, UN agencies, African Union, bilateral donors (UK Aid, SIDA)	Concessional loans, results-based financing, technical assistance, multilateral coordination frameworks, sanctions and incentive mechanisms for governance reform
National	Design differentiated electrification strategies by province; reform EDM's tariff structure toward cost-reflectivity; strengthen ARENE as independent regulator; allocate public investment to priority LHF areas identified by the gap analysis.	Government of Mozambique, Ministry of Mineral Resources and Energy, EDM, ARENE, FUNAE, Ministry of Finance	National electrification plans (ETE 2023), tariff reform, public investment budgets, regulatory reform, EDM unbundling
Regional / Provincial	Implement cluster-specific interventions: grid extension in Zambezia, Nampula, Gaza, Inhambane, Sofala, Manica; connection subsidies and social tariffs in Maputo peri-urban areas; off-grid and mini-grid solutions in Tete and Niassa; conflict-sensitive transitional approaches in Cabo Delgado.	Provincial governments, EDM regional directorates, local NGOs, private sector IPPs, international development organisations with provincial presence	Grid extension investment, mini-grid licensing, connection subsidy programmes, conflict-sensitive planning, off-grid procurement
Local	Facilitate household-level connection uptake; manage community mini-grids; address land tenure and informal settlement challenges; build local technical capacity for operation and maintenance.	Municipal authorities, community organisations, local EDM offices, solar home system distributors, local civil society	Community-driven electrification schemes, prepayment solutions, local capacity building, land regularisation, productive use promotion

This multi-level framework reflects a broader lesson from the energy access literature in sub-Saharan Africa: national electrification strategies that are designed and implemented exclusively at the central government level, without effective coordination with international financing institutions, provincial authorities, and local communities, tend to underperform relative to their stated targets(25).

The results suggest three intervention clusters with distinct operational logics, each requiring different priorities and instruments.

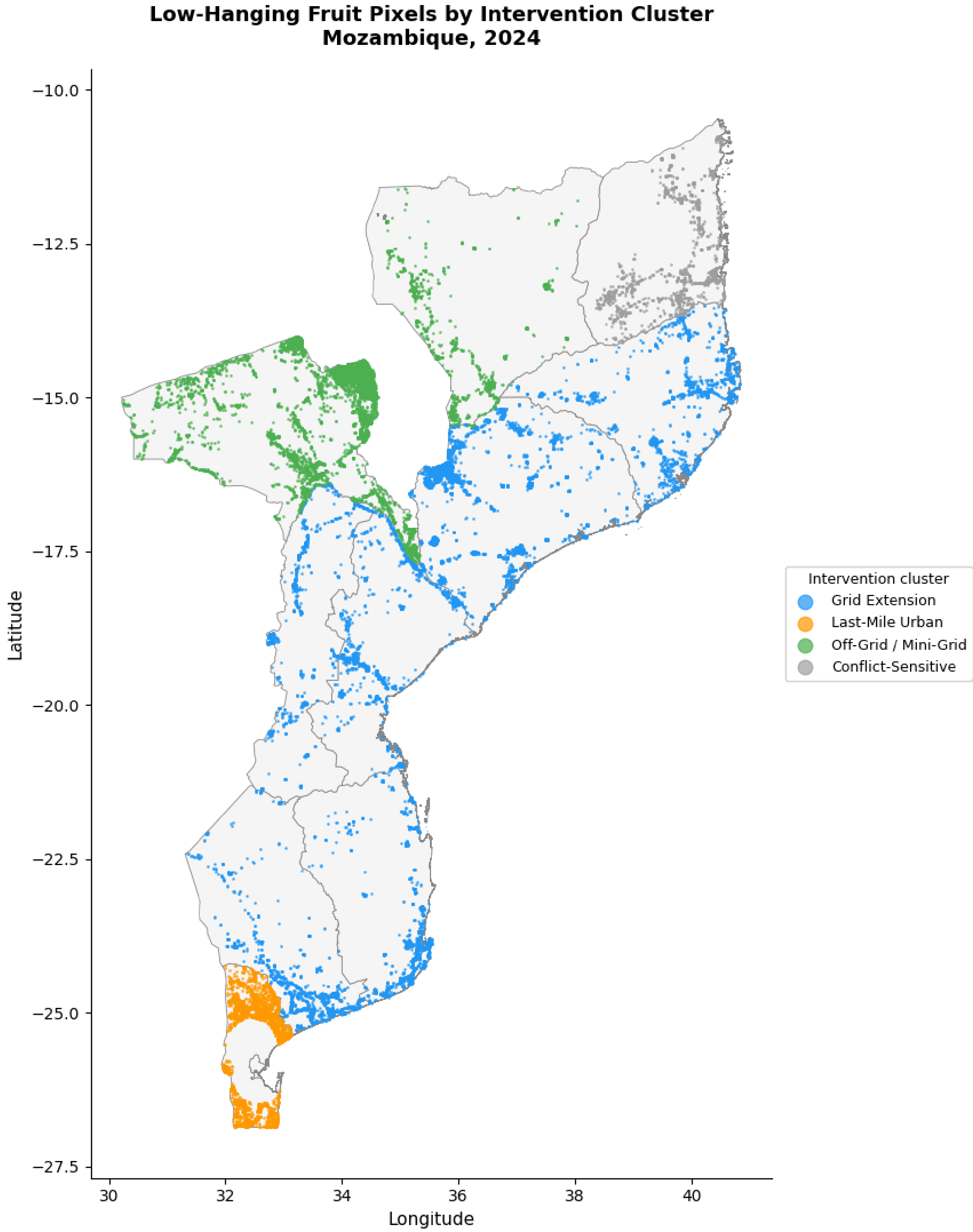


Figura 27-Geographic distribution of low-hanging fruit pixels by intervention cluster, Mozambique, 2024.

Cluster 1 - Immediate grid extension

The provinces of Zambezia, Nampula, Gaza, Inhambane, Sofala, and Manica present the most favourable configuration for short-term grid extension interventions. LHF pixels in these provinces have average grid distances ranging from 11 km (Manica) to 21 km (Nampula), distances that are generally considered economically viable for distribution network extension(29). Average population per pixel is sufficient to justify the investment, and the absence of active conflict renders these territories accessible for infrastructure works.

The SHAP profile of these provinces, where RWI is the dominant driver and infrastructure distances carry secondary weight, suggests that the binding constraint is not physical remoteness from the grid but the economic capacity of households to sustain connection costs. This implies that physical grid extension alone may be insufficient: it must be accompanied by instruments that reduce connection costs for low-income households. Connection subsidies, instalment payment plans, and income-adjusted social tariffs are complementary tools that would significantly increase the effectiveness of the infrastructure investment, as documented in the energy access literature for comparable SSA contexts(18).

In terms of operational priority, Zambezia (4,860 LHF pixels) and Nampula (1,996 LHF pixels) emerge as the provinces with the greatest unserved demographic potential in this cluster, combining a high number of LHF pixels with contained grid distances and significant population concentrations.

Cluster 2 - Urban last-mile

Maputo presents a radically different profile from the other clusters. Its 4,127 LHF pixels have an average grid distance of 20.8 km, comparable to Cluster 1 provinces, but an average RWI of −0.44, the lowest of all provinces. This profile indicates that Maputo's LHF pixels are not remote rural areas but most likely informal and peri-urban settlements where the electricity grid is physically accessible but connection has not occurred for economic reasons.

The urban last-mile phenomenon is well documented in the energy access literature(18): in the peripheries of large African cities, the grid often physically reaches areas but the poorest households cannot sustain connection costs, which in many SSA countries can amount to several months of household income. In this context, further extending the grid does not resolve the problem: the constraint is economic, not infrastructural.

The appropriate policies for this cluster are therefore fundamentally different from those of Cluster 1. Direct connection cost subsidies, concessional financing programmes, reduced connection tariffs for informal settlements, and prepayment solutions that lower the entry barrier are the most relevant instruments. The SHAP analysis of Maputo, where HDI carries an anomalous weight of 0.241, also suggests that integrated interventions combining electricity access with other human development services could generate synergistic effects larger than those of purely energy-focused interventions.

Cluster 3 - Off-grid and mini-grid solutions

Tete (average grid distance 66.2 km) and Niassa (40.8 km) represent contexts where conventional grid extension is economically difficult to justify. At these distances, the cost per kilometre of transmission line typically exceeds the threshold of economic viability for communities with the income levels observed in these provinces(29).

The appropriate solutions for this cluster are individual off-grid systems, primarily solar home systems, or community mini-grids powered by renewable sources. Mozambique has abundant solar resources across its entire territory, with average irradiance exceeding 5 kWh/m²/day in the central and northern provinces(14). This makes photovoltaic systems technically and economically competitive relative to grid extension for remote communities. The government's Energy Transition Strategy already recognises this, with a target of 30 to 35 percent of universal access to be achieved through off-grid and mini-grid solutions by 2030(14).

Cabo Delgado: a separate case

Cabo Delgado does not fit into any of the three preceding clusters because its fundamental problem is neither infrastructural nor economic, but political and security-related. The 1,840 LHF pixels identified have the highest average population per pixel of all provinces (17.9 inhabitants), indicating significant demographic potential. However, as documented in the literature on the conflict-energy nexus(16), infrastructure investments in active conflict zones are exposed to risks of destruction and interruption that negate their economic value.

The recommendation for Cabo Delgado is therefore sequential: political stabilisation and security restoration must precede infrastructure investment. In the short term, community-driven approaches involving local actors and small-scale off-grid systems, which are less vulnerable to destruction than grid infrastructure, could represent a transitional solution for the most exposed communities.

A cross-cutting element of all recommendations is the need for continuous monitoring. The framework developed in this study is replicable annually without substantial methodological modifications: new VIIRS NTL data are available each year, and updating the model and the gap analysis would allow tracking of progress toward Mozambique's 2030 universal electrification targets while updating priority lists as conditions change. In particular, monitoring LHF pixels over time, verifying how many are effectively electrified year after year, would allow empirical validation of the model's predictions and finer calibration of the operational thresholds of the gap analysis. This would transform the present study from a static analysis into a dynamic decision support system for Mozambican energy planning, directly usable by EDM, the government's energy ministry, and international organisations operating in the country's energy sector.

.

6.3. Limitations

This study presents certain methodological and substantive limitations that must be discussed for a more accurate interpretation of the results. The limitations can be organised into five main areas.

6.3.1. The dependent variable as a proxy

NTL is a proxy for electricity access, not a direct measure of it. The nighttime light intensity detected by the satellite reflects the presence of artificial light sources, which include not only the electrification of residential areas and dwellings but also industrial activity, commercial operations, and street lighting. The VNP46A4 dataset includes filtering procedures to reduce these interferences, but does not eliminate them entirely. This introduces a measurement error in the target variable that is difficult to quantify without ground truth data.

A more substantive limitation concerns small-scale off-grid and mini-grid systems. A dwelling served by a solar home system or a community mini-grid may not produce sufficient luminosity to be detected by the satellite at 1 km^2 resolution. This implies a systematic underestimation of real access in remote areas, precisely those where off-grid systems are most prevalent. The results of the gap analysis, and in particular the identification of LHF pixels, must be interpreted in light of this limitation: some areas classified as dark may in reality have partial access through systems not connected to the grid.

6.3.2. Static variables

The variables expressing distance from the grid, distance from services, and built-up area are treated as constant over time. In reality, these variables change: new transmission lines are built, new hospitals and schools are opened, new settlements are created. The assumption of stationarity introduces a measurement error that grows with the length of the analysis period: a pixel classified as far from the grid in 2015 may have become close to the grid by 2022 following new infrastructure investments. This error is particularly relevant for dist_grid_km, which is the static variable with the greatest potential temporal variability.

A direct consequence is that the model may overestimate the relative importance of dynamic variables (RWI and log_pop) compared to infrastructure variables, since the latter do not capture the real changes that occurred during the 2015 to 2024 period. Updating static variables with annual grid expansion data, where available, would represent a significant methodological improvement for future research.

6.3.3. Dataset imbalance

With 99.4% of pixels dark, the model is trained predominantly on observations with log_ntl = 0. The R^2 on lit pixels only (0.487 in the temporal split) reflects this structural difficulty and indicates that the dynamics of already-electrified pixels are less well captured. This imbalance also has implications for the gap analysis: the operational threshold y_pred > 0.01 was chosen to identify pixels where the model predicts light in a non-trivial manner, but remains an arbitrary choice. Variations in the threshold produce different priority lists; a systematic sensitivity analysis on the threshold would be necessary to assess the robustness of the gap analysis results.

6.3.4. Cabo Delgado and conflict contexts

The results for Cabo Delgado must be interpreted with additional caution relative to the rest of the country. Armed conflict creates structural discontinuities, including population displacement, infrastructure destruction, and changes in economic activity, that the model cannot capture through the available variables. The low provincial R^2 (0.609) is an indirect measure of this limitation, but does not quantify it precisely. The LHF pixels identified in Cabo Delgado may reflect pre-conflict conditions that are no longer current in 2024.

More generally, the framework developed in this study implicitly assumes that the structural relationships between socioeconomic variables and electricity access are stable over time. This assumption is reasonable for most of the country, as demonstrated by the stability of R^2 across test years, but does not hold for contexts where exogenous events such as armed conflicts or natural disasters structurally alter socioeconomic conditions.

6.3.5. Absence of cost data

The gap analysis identifies priority intervention areas based on the structural characteristics of pixels, but does not account for intervention costs. A LHF pixel close to the grid is not necessarily cheaper to serve than a more remote one if terrain conditions, settlement characteristics, or land tenure constraints are unfavourable. The integration of connection cost data, for instance through geospatial cost modelling tools such as OnSSET(30), would transform the priority list from a potential-based classification into a cost-benefit ranking, substantially increasing the operational utility of the results for decision-makers.

6.4. Future Research

The results of this study open several directions for future research that could extend and deepen the developed framework.

6.4.1. External validation with survey data and stakeholder engagement

The integration of survey data on electricity access would allow external validation of the results. Comparing model predictions with real access data, such as those from the Demographic and Health Surveys (DHS) or the World Bank Living Standards Measurement Studies (LSMS), would allow quantification of the accuracy of the NTL proxy and calibration of the operational thresholds of the gap analysis on more solid empirical foundations. In particular, the threshold y_pred > 0.01 could be optimised against ground truth data rather than fixed arbitrarily.

However, quantitative survey validation alone would be insufficient to capture the full complexity of electricity access dynamics in Mozambique. The results of this study suggest that the drivers of access vary not only spatially but also institutionally: what explains why a pixel remains unelectrified in Maputo is fundamentally different from what explains the same condition in Cabo Delgado or Niassa. Understanding these differences in depth requires complementing the quantitative analysis with qualitative fieldwork and structured stakeholder engagement at multiple levels of governance.

A comprehensive external validation framework should therefore include interviews and consultations with actors across the four levels identified in Section 6.2. Table 10 provides an indicative mapping of the key stakeholders to be engaged at each level for both validation and policy implementation purposes.

Table 10. Stakeholder mapping for external validation and policy implementation.

Level	Policy focus	Key stakeholders
International / Multilateral	Concessional finance, donor coordination, governance reform, conflict resolution	World Bank, African Development Bank, USAID Power Africa, EU, UNDP, African Union, bilateral donors (UK FCDO, SIDA)
National	Electrification strategy, tariff reform, regulatory independence, public investment allocation	Ministry of Mineral Resources and Energy, EDM, ARENE, FUNAE, Ministry of Finance
Regional / Provincial	Cluster-specific interventions (grid extension, last-mile urban, off-grid, conflict-sensitive approaches)	Provincial governments, EDM regional directorates, local NGOs, private sector IPPs, FUNAE regional offices, mining companies (Tete)
Local	Household connection uptake, community mini-grid management, land tenure, local capacity building	Municipal authorities, community organisations, local EDM offices, SHS distributors, district administrators

This stakeholder mapping reflects a central conclusion of this study: complex, multi-layered problems such as electricity access in Mozambique cannot be adequately understood or addressed through quantitative analysis alone. The pixel-level findings identify where to intervene and what structural conditions are associated with access, but only structured engagement with actors at all four levels of governance can produce the depth of understanding required to design interventions that are technically sound, institutionally feasible, and socially legitimate.

6.4.2. Annual updating and monitoring

The framework developed is replicable annually without substantial methodological modifications: new VIIRS NTL data are available each year, and updating the model and the gap analysis would allow tracking of progress toward Mozambique's 2030 electrification targets. In particular, monitoring LHF pixels over time, verifying how many are effectively electrified year after year, would allow empirical validation of the model's predictions and the construction of a dynamic decision support system for energy planning.

6.4.3. Integration of dynamic infrastructure variables

The principal methodological limitation of this study is the treatment of infrastructure variables as static. Future research could integrate annual data on electricity network expansion, if available from EDM (Electricidade de Moçambique) or from satellite sources, to transform dist_grid_km into a dynamic variable. This would allow the model to capture the effect of infrastructure expansion on electricity access over time, separating it from the effect of socioeconomic variables.

6.4.4. Extension to other SSA countries

Extending the framework to other sub-Saharan African countries with similar characteristics, including low electricity access, strong spatial heterogeneity, and the presence of conflicts, would allow assessment of the generalisability of the results. In particular, it would be valuable to verify whether the dominance of RWI over dist_grid_km as a driver of access is a pattern specific to Mozambique or a more general characteristic of low-income SSA countries.

6.4.5. Complementary causal framework

The integration of a causal framework, such as Two-Way Fixed Effects (TWFE) with pixel and year fixed effects, would complement the predictive results of the Random Forest with estimates of the causal effect of dynamic variables over time. TWFE estimates how much electricity access changes in a given pixel when a variable changes over time, controlling for all fixed pixel characteristics. This is particularly relevant for conflict_violence and HDI, whose causal effect on electricity access is of direct interest to policy makers but is not identifiable with a purely predictive model.

6.4.6. Separate models for conflict contexts

The low R^2 of Cabo Delgado (0.609) suggests that the structural relationships between socioeconomic variables and electricity access in conflict zones are sufficiently different to justify a dedicated analytical framework. Future research could develop models specific to contexts of political instability, integrating variables that better capture conflict dynamics, such as conflict intensity, distance from active fronts, or population displacement, and applying them not only to Cabo Delgado but to other African contexts where conflict and energy deficit coexist.

List of figures

Bibliography

1. Africa Energy Outlook 2022 [Internet]. Paris: IEA; 2022. Available from: https://www.iea.org/reports/africa-energy-outlook-2022

2. Falchetta G, Pachauri S, Parkinson S, Byers E. A high-resolution gridded dataset to assess electrification in sub-Saharan Africa [Internet]. 2019. Available from: https://www.nature.com/articles/s41597-019-0122-6

3. Henderson V, Storeygard A. Measuring Economic Growth from Outer Space [Internet]. 2012. Available from: https://www.aeaweb.org/articles?id=10.1257/aer.102.2.994

4. Doll CNH, Pachauri S. Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. Energy Policy. 2010 Oct;38(10):5661–70. doi:10.1016/j.enpol.2010.05.014

5. Gibson J, Boe-Gibson G. Nighttime Lights and County-Level Economic Activity in the United States: 2001 to 2019. Remote Sens. 2021 Jul 12;13(14):2741. doi:10.3390/rs13142741

6. Román MO, Wang Z, Sun Q, Kalb V, Miller SD, Molthan A, et al. NASA's Black Marble nighttime lights product suite. Remote Sens Environ. 2018 Jun;210:113–43. doi:10.1016/j.rse.2018.03.017

7. Elvidge. VIIRS night-time lights [Internet]. 2017. Available from: https://www.tandfonline.com/doi/full/10.1080/01431161.2017.1342050 doi:10.1080/01431161.2017.1342050

8. Yeh C, Perez A. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa [Internet]. 2020. Available from: https://www.nature.com/articles/s41467-020-16185-w

9. Ratledge N, Cadamuro G, de la Cuesta B. Using machine learning to assess the livelihood impact of electricity access [Internet]. 2022. Available from: https://pubmed.ncbi.nlm.nih.gov/36385544/ doi:10.1038/s41586-022-05322-8

10. Chi G, Fang H, Chatterjee S, Blumenstock JE. Microestimates of wealth for all low- and middle-income countries. Proc Natl Acad Sci. 2022 Jan 18;119(3):e2113658119. doi:10.1073/pnas.2113658119

11. Belgiu M, Drăguț L. Random forest in remote sensing: A review of applications and future directions [Internet]. 2016. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0924271616000265 doi:10.1016/j.isprsjprs.2016.01.011

12. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 2018 Aug 29;6:e5518. doi:10.7717/peerj.5518

13. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. 2017. Available from: https://arxiv.org/abs/1705.07874

14. Mozambique 2024 [Internet]. Paris: IEA. Available from: https://www.iea.org/reports/mozambique-2024

15. Zamponi M. Dove va Cabo Delgado? Crisi politica, islamismo e violenza nel Nord del Mozambico. Nuovi Autoritarismi E Democr Dirit Ist Soc NAD-DIS. 2022 Jun 30;4(1). doi:10.54103/2612-6672/18123

16. Shettima A, Elheddad M, Bassim M, Alfar AJK. The impact of conflict on energy poverty: Evidence from sub-Saharan Africa. Resour Policy. 2023 Oct;86:104090. doi:10.1016/j.resourpol.2023.104090

17. Falchetta G, Pachauri S, Byers E. Satellite Observations Reveal Inequalities in the Progress and Effectiveness of Recent Electrification in Sub-Saharan Africa [Internet]. 2020. Available from: https://www.sciencedirect.com/science/article/pii/S2590332220301433

18. Salite D, Kirshner J, Cotton M, Howe L, Cuamba B, Feijó J, et al. Electricity access in Mozambique: A critical policy analysis of investment, service reliability and social sustainability. Energy Res Soc Sci. 2021 Aug;78:102123. doi:10.1016/j.erss.2021.102123

19. Gibson J, Olivia S, Boe-Gibson G, Li C. Which night lights data should we use in economics, and where? J Dev Econ. 2021 Mar;149:102602. doi:10.1016/j.jdeveco.2020.102602

20. Tatem AJ. WorldPop, open data for spatial demography. Sci Data. 2017 Jan 31;4(1):170004. doi:10.1038/sdata.2017.4

21. Smits J, Permanyer I. The Subnational Human Development Database. Sci Data. 2019 Mar 12;6(1):190038. doi:10.1038/sdata.2019.38

22. Didan K. MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006 [Internet]. NASA Land Processes Distributed Active Archive Center; 2015 [cited 2026 Mar 17]. Available from: https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod13a3-006 doi:10.5067/MODIS/MOD13A3.006

23. Raleigh C, Linke A, Hegre H, Karlsen J. Introducing ACLED: An Armed Conflict Location and Event Dataset. J Peace Res. 2010 Sep 1;47(5):651–60. doi:10.1177/0022343310378914

24. Schon J, Koren O. Introducing AfroGrid, a unified framework for environmental conflict research in Africa. Sci Data. 2022 Mar 29;9(1):116. doi:10.1038/s41597-022-01198-5

25. Salite D, Kirshner J, Cotton M, Howe L, Cuamba B, Feijó J, et al. Electricity access in Mozambique: A critical policy analysis of investment, service reliability and social sustainability. Energy Res Soc Sci. 2021 Aug;78:102123. doi:10.1016/j.erss.2021.102123

26. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens. 2016 Apr;114:24–31. doi:10.1016/j.isprsjprs.2016.01.011

27. Ogbe MA, Abdullahi MS, Ding Y. Measuring how armed conflict impacts economic growth in sub-Saharan Africa through spatial analysis. Front Polit Sci. 2024 Sep 13;6:1433584. doi:10.3389/fpos.2024.1433584

28. UNITED NATIONS DEVELOPMENT PROGRAMME. HUMAN DEVELOPMENT REPORT 2023/2024;BREAKING THE GRIDLOCK - REIMAGINING COOPERATION IN A POLARIZED WORLD. S.l.: UNITED NATIONS DEVELOPMEN; 2024. 1 p.

29. Egli F, Agutu C, Steffen B, Schmidt TS. The cost of electrifying all households in 40 Sub-Saharan African countries by 2030. Nat Commun. 2023 Aug 21;14(1):5066. doi:10.1038/s41467-023-40612-3

30. Mentis D, Howells M, Rogner H, Korkovelos A, Arderne C, Zepeda E, et al. Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa. Environ Res Lett. 2017 Aug 1;12(8):085003. doi:10.1088/1748-9326/aa7b29