An Open-Source Computational Framework for the Scalable Application of Electrification Planning Models

by

Lama Sara Aoudi

B.S. Civil and Environmental Engineering
University of Illinois at Urbana-Champaign (2017)

Submitted to the Institute for Data, Systems, and Society
and the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Master of Science in Technology and Policy

and

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author . . . .
Institute for Data, Systems, and Society
Department of Electrical Engineering and Computer Science
September 17, 2021

Certified by . . . .
Ignacio J. Pérez-Arriaga
Visiting Professor, Sloan School of Management
Professor, Electrical Engineering, Universidad Pontificia Comillas
Thesis Supervisor

Certified by . . . .
James L. Kirtley Jr
Professor, Electrical Engineering and Computer Science
Thesis Reader

Accepted by . . . .
Noelle E. Selin
Director, Technology and Policy Program
Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric and Planetary Sciences

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

An Open-Source Computational Framework for the Scalable Application of Electrification Planning Models

by

Lama Sara Aoudi

Submitted to the Institute for Data, Systems, and Society
and the Department of Electrical Engineering and Computer Science
on September 17, 2021, in partial fulfillment of the
requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science

Abstract

Global efforts to achieve affordable, universal electrification are inextricably linked to the uptake and application of spatially explicit computational electrification planning tools. The tools automate the delineation of potential modes of electrification (grid extension, mini-grids, and stand-alone solar systems) across all customers, at the lowest cost. My thesis aims to expedite universal electrification by designing a scalable and modular framework on how to leverage open-source information, to produce granular estimates of the data needed by these planning tools. The design of the framework is configured to the input requirements of the most data-demanding tool in the field, the Reference Electrification Model (REM). The REM inputs modelled in my framework are the following: 1) the geolocation and demand characteristics of residential customers, 2) the geolocation and demand characteristics of social and productive customers, 3) the electrification status of each customer, and 4) the layout of the medium-voltage distribution network.

The framework prescribes a method, or Python model, to process and analyze existing open data into reasonable estimates of each REM input. The source code of the Python-based framework is available through GitHub, with directional documents on how to run each module to any region. Built-area population datasets are used to estimate the exact geolocation of all residential and smaller load customers (e.g., village community centers). The geolocation of non-residential loads, or community and productive customers (e.g., markets or large industrial plants), are extracted from the Google Maps API. To do so, the model iterates through an area-of-interest and searches the API for the locations and operational status of each customer type. Non-residential loads are ascribed demand patterns from t supporting literature on archetypal behaviors of larger loads. The Falchetta et al. (2019) is used to classify each customer's electrification status and assign them a "tier" or level of electricity consumption. Finally, I propose leveraging the grid-design capabilities of REM to estimate a region's existing medium voltage distribution network when it is unavailable through other means. To apply REM for the task, I initialize its model parameters to force an 'all-grid-extension' output and supply it with available a-priori information on the existing medium voltage network. Overall, my thesis lays the foundation for a complete transition of the energy access sector to granular, open-source modelling.

Thesis Supervisor: Ignacio J. Pérez-Arriaga
Title: Visiting Professor, Sloan School of Management
Professor, Electrical Engineering, Universidad Pontificia Comillas

Thesis Reader: James L. Kirtley Jr
Title: Professor, Electrical Engineering and Computer Science

Acknowledgements

To my advisor, Ignacio Pérez-Arriaga, thank you for being a patient mentor over the years. Your knowledge and guidance have taught me the merits of research and the value of challenging oneself to endeavor into new fields. I could not have achieved this document today without your willingness to read, contribute, and edit. Thank you.

Thank you to Elsa Olivetti for ushering me through this last year. You have been an incredible mentor to learn from and grow with. The attention and care you show your students makes all the difference and I am endlessly grateful for it.

To Pedro Ciller Cutillas and Andrés González García, greatest appreciation for rallying with me these last few months and helping out with REM.

To Barb DeLaBarre, you were my first and longest friend at MIT. Even though Covid rudely brought an end to our daily chats, you continue to bring color to my life anytime we speak.

To the TPP staff and my 2020 cohort, thank you for making the in-person time we had memorable, and for making virtual grad school bearable. Axelle, Erin, Nina, and Sade you were my lifeline these last two years. I am so grateful for our friendship and continue to be wildly impressed by the intelligent, committed, and ambitious women you each are.

To my parents the biggest thank you for being a source of unconditional support and encouragement. To Fatema, you inspire me. You have kept me laughing and sane for as long as I've known you. Bora and Liam, thank you for five years of emotional (and coding) support. To the rest of my family and friends, let's catch-up- I am officially done with my thesis. And Ali, this would been impossible without all your help over the years.

Contents

Figures

Chapter 1

Introduction

1.1 Setting the Scene

Despite the technological leaps of the last several decades, large populations still lack access to necessities like affordable electricity, clean drinking water, or sanitation. Recognizing the scale of this global disparity, all United Nations Member States adopted the 2030 Agenda for Sustainable Development in 2015. The Agenda laid out 17 Sustainable Developmental Goals considered to be pre-requisites for eradicating poverty and ushering developing countries into modern, thriving economies. Sustainable Development Goal Seven, and the subject of this thesis, is the guarantee of "access to affordable, reliable, sustainable, and modern energy for all".

Electricity services encompass everything from the cooling of vaccines to irrigation pumping to the operation of commercial businesses. The global COVID-19 pandemic has highlighted the critical nature of electrification, as healthcare facilities struggle to meet increased healthcare demand with poor or unreliable power (IEA et al., 2020). The International Energy Agency (IEA) estimates that 759 million people lived without electricity in 2019, with 75%, or 569 million, of the unelectrified people residing in Sub-Saharan Africa (SSA) (IEA et al., 2021). Despite considerable investments by public and private organizations, the energy gap has persisted; and at current electrification rates 660 million people will still lack access[1] in 2030 (IEA et al., 2021).

The slow progress demonstrates the techno-economic complexity and scale of the coordination needed between local stakeholders, governmental institutions, private utilities, and businesses.

[1] (IEA et al., 2021) "Access to electricity" or "the electrification rate" refers to the share of the population with access to electricity over a specified time period or geographic area. It is defined as the ability of the end-user to consume electricity for desired services A method for measuring access to electricity (or energy) is defined in the World Bank's Multi-Tier Framework (Bhatia & Angelou, 2015), discussed in more detail in Chapter 4.

Energy access initiatives should build-upon the understanding between electricity demand and cost of supply to ensure that electricity access and development become mutually reinforcing endeavors (Odarno et al., 2017). For instance, if a supply system is poorly designed it can constrain the productive activity of households, community facilities, and enterprises. Or if an area is densely populated but remote, decentralized mini-grid solutions may be more cost-efficient but may inhibit the rapid demand growth made possible by an extension of the local grid. To address these challenges and trade-offs, research organizations have started to develop spatially explicit electrification tools to assist network planners in decision-making and budget-allocation. The following section delves deeper into these electrification planning tools.

1.1.1 Narrowing In: Universal Energy Access Planning

Access to energy is lower in rural areas than in urban areas. In 2019, 21%, or 97 million, of urban residents and 72%, or 471 million, of rural residents in Sub-Saharan Africa lacked access to electricity (IEA et al., 2021). Extending the grid to rural, often remote, communities is usually economically unattractive and therefore these settlements mostly remain un-electrified (Szabó et al., 2011). To meet energy access goals in rural areas, countries have begun to use hybrid solutions that pair networked systems with decentralized technologies. Decentralized electrification technologies include islanded grids (aka. mini- or micro-grids) and stand-alone solar home systems (SHS). The recent proliferation of off-grid solutions is largely due to the rapid cost decline of solar PV, battery technologies, and energy-efficient appliances. Electrification planning tools apply least-cost methods and GIS (Geographic Information Systems) to determine which demand points may be better served by grid extensions or off-grid solutions. By pairing the spatial and temporal dimensions to GIS data, a planner can tailor network analyses and solutions more heterogeneously than traditional frameworks (Khavari et al., 2021). An overview of available electrification planning tools is available in Chapter 2.

This thesis will focus primarily on the challenge of providing the required input data to the electrification tool the presently represents the state-of-the-art in this field: the Reference Electrification Model, or REM.

REM is a computer-based optimization tool that identifies the least cost electrification plans that meet demand at a prescribed reliability level. The tool is designed to provide ministries and regulators quantitative support when formulating policies and informing developers on optimal locations for off-grid technologies. It can be used also by electrification planners, developers of off-grid solutions and others that study electrification for different purposes. REM stands out among other electrification tools because it provides system designs at any geographic scale, working at the individual customer level, employing full representations of each customer's hourly demand patterns, respecting the physical laws and constraints of power systems, explicitly modeling reliability targets, and employing optimizations to find least cost combinations of electrification delivery modes (Ciller et al., 2019). To achieve this level of granularity REM depends on extensive input data such as the distribution and density of population settlements and productive facilities, electricity demand levels at the consumer level, energy resources availability, economic activity, distance from functional infrastructure (e.g., transmission and distribution network, roads, power stations), and more (Ciller et al., 2019). As such, REM's effectiveness depends heavily on credible and up-to-date records of existing infrastructure in the area of interest. *The objective of this thesis is to establish a scalable framework for how to cost- and time-efficiently obtain, process, and validate REM's input requirements for any country.*

1.2 Motivation

Global, regional, and national development policymakers lack critical data. Many governments still do not have access to adequate data related to their populations. Those that do struggle to maintain or afford regular updates to the data through standard methods like field surveys. This lack of data is particularly prevalent in the poorest and most marginalized communities, the very places that leaders will need to focus on if they are to achieve zero extreme poverty by 2030 and zero emissions by 2050 (United Nations, 2017). The magnitude of the data gap is even larger when trying to acquire the detailed inputs required by REM. So far, applications of REM have hinged on established partnerships between researchers and national utilities, governmental agencies, or the sponsoring non-profit organizations (e.g., the World Bank). Through these engagements, existing data is aggregated, and surveys are conducted to impute missing data.

However, this process is resource intensive and limited by procedural red-tape and unpredictable response times. Moreover, annual collection methods and stakeholder coordination do not lend themselves to long-term policy or goal changes since any future re-implementation of a planning tool would require costly re-collection of necessary data. The same holds for wide-scale (i.e., multiple country) analyses which would require personal relationship formation with each country's representatives. Finally, as hinted above, the data that may be available from these on-the-ground sources can be incomplete, unavailable in geospatial format, outdated, or exclude the rural populations most in need of electricity. In recent years, new open sources of data, such as satellite data, and new analytical approaches, such as Machine Learning (ML), have enabled more agile and efficient methods of data collection which can be leveraged instead of costly partnerships to apply REM and other electrification planning tools widely and repeatedly. Open datasets can also better measure progress on SDGs and increase transparency and accountability between citizens and their government

1.2.1 Benefits to Open-Source Datasets

The volume of data in the world has increased tremendously in the last few years. A large share of these data is passively collected by means of remote-sensing satellites and everyday interactions with digital products or services (e.g., mobile phones). The data is growing because the passive methods of gathering and storing it are inexpensive and numerous. Best of all, this data is 'open' or freely available to everyone. The data deluge is a by-product of the larger "open-source" movement, which aims to make software, hardware, knowledge, government, and education openly accessible to anyone. New insights gained from open data collection can complement official statistics and survey data, adding depth, timeliness, and nuance to information on human behaviors and experiences (United Nations, 2017). If applied responsibly, open data can assist governments and other stakeholders measure progress on SDGs and shed light on social or infrastructure disparities previously hidden (United Nations, 2017). Most importantly, open data increases transparency and accountability within governments and between them and citizens.

The integration of open data, ML, and advancements in GIS tools solve the data collection bottlenecks, which have impeded the broader uptake of electrification planning tools like

REM. High-resolution geospatial datasets, derived from Earth Observation (EO) technologies, provide a range of location-specific information that may fulfill REM's input requirements. The location and distribution of consumers can be estimated by the Google Open Building dataset. That dataset uses satellite imagery and ML algorithms to geolocate buildings for the entire continent of Africa (Sirko et al., 2021). Altogether, there are several advantages to using EO technologies, or satellites as inputs to REM.

> "First, compared to ground-based methods, the use of EO technologies, and in particular the use of satellites, allows the production of cost-effective data with a higher frequency over longer periods of time and over larger spatial extents. Second, EO technologies enable the collection of near real-time, objective, and independent data for remote and marginalized areas that have previously been ignored. Third, when combined with traditional data (e.g., field surveys, census data, demographic and socio-economic statistics), EO data (satellite imagery) supplement and/or enhance the quality of the information by improving its spatial resolution and interpretation capabilities (including better visualization)."
> Palacios-Lopez et al., 2021

Open sources also provide a "control" on input data, which is useful when running REM for several countries at once. In other words, when trying to extract the least-cost electrification plan for multiple areas, the inputs are not more or less granular for any country or region, which 'levels the playing field' between data-rich and data-poor countries, minimizing biases in the results.

Thus, developing a framework to run electrification planning tools on open-source data promises to expedite progress on SDG 7 by enabling rapid utilization at lower cost. This in turn mobilizes investment and speeds up the implementation process. Finally, if transparently designed and openly available, the framework can be audited by third parties, which ensures quality control and demonstrates due diligence to investors.

1.3 Why this Framework?

The objective of this thesis is to design and test a framework that enhances the applicability of REM by making it possible to obtain its inputs from open-source data, for any region, without costly engagements with on-the-ground partners. The computational framework applies a variety of open datasets and leverages Python GIS libraries to estimate (or infer) inputs of REM which are not currently publicly available. The list of inputs include: the geolocation and load profiles of households and productive facilities, the electrification status of each consumer, and the layout of the region's existing Medium Voltage distribution network. Pre-acquired ground-truth data from Rwanda and reported energy statistics are used to validate the estimates. The framework is built in a modular format such that improved data sources can be easily substituted. Moreover, though REM has traditionally been used to develop detailed electrification plans for whole countries, future users may prefer to run coarser analyses that only delineate the high-level spatial contours of immediate (or intermediate) investment plans for electrification in a larger region. Building modularity into the framework enables these users to leverage coarser datasets as inputs and avoid the long computation times of granular data. Finally, while the design of the framework is configured to the exact needs of REM, its emphasis on open-data and key attributes of electrification plans should make it applicable and effective at estimating inputs for any electrification planning tool. The framework is synergistic with broader climate goals described below.

1.3.1 Concurrent Efforts with Climate Goals

Unlike today's developed nations, developing countries face paradoxical goals of achieving modernization as quickly as possible, while also investing and prioritizing environmental sustainability and global emissions reduction. In other words, they must meet the UN's Sustainable Development Goals (SDGs) while simultaneously remaining on track with their Nationally Determined Contributions (NDCs). Cleaner technologies are often less available and more expensive in these regions than fossil fuel options (e.g., diesel and petroleum). In the case of SDG 7, electrification needs are currently met via uncoordinated efforts between private and public investors. As a result, researchers are unable to predict how the final delineation of electrification modes, and subsequently the energy mix, will impact a country's greenhouse gas emissions, local air quality, or effectiveness at meeting NDCs. My framework enables researchers to predict the future energy supply in a country by running REM with openly sourced data, equipping them with the means to put preventative policies in place. Several studies have also demonstrated the synergy between electrification, clean-cooking, and regional air-quality (Maji, 2020; Dagnachew et al. 2019). Universal electrification can encourage uptake of cleaner cookstoves and improve local air-quality, reduce the emissions of climate-warming particulate matter (PM2.5), and improve the overall health and safety of women in households. My framework can quantify this synergy by running REM for a variety of policy scenarios and comparing their economic outcomes.

1.4 Structure

Chapter 2 discusses REM in greater detail and provides a review of other least-cost electrification planning tools. Chapter 2 also links the work of this thesis to complementary open data energy access frameworks. Next, Chapter 3 explains the design parameters considered when creating the framework, explicating which REM inputs must be estimated with open data and why. Section 2 of Chapter 3 discusses the Python implementation of the framework and how users are anticipated to interact with it. Chapter 4 is divided into three large sections; each section details a single consumer-related input. These sections include a review of available data sources, the framework's proposed method for processing the raw data into REM-formatted inputs, and preliminary results from applying the method in Rwanda. Chapter 5 describes the method of estimation of a region's medium voltage network, with a qualitative comparison between preliminary results from Rwanda on ground-truth data and competing open data sources. Finally, Chapter 6 discusses possible applications of the framework to complementary developmental goals and suggests future work.

Chapter 2

Literature Review

Section 2.1 introduces electrification planning models and describes their success at assisting governments achieve their universal electrification goals. Section 2.2 then situates the proposed framework (described in Chapter 3) within complementary efforts to improve the state of energy access through publicly available data sources.

2.1 Electrification Planning Tools

With the advent of cheap, modular renewables, providing access to electricity through distributed generation systems is increasingly common. These systems differ from a centralized grid by generating power much closer to the site where it is consumed. As a result, poorer, more remote populations (i.e., 'last-mile' customers) are connected faster and with fewer capital risks (Morrissey, 2019). There are two primary types of distributed generation (also referred to as 'off-grid') systems: mini-grids (MGs) and stand-alone solar home systems (SHS). Different sizes of these off-grid technologies have created new opportunities to supply households with affordable energy. However, their integration has created challenges for electrification planners because planning must now involve the assessment of which technology option (GE, MG, or SHS) can provide energy services to each household at the lowest cost.

The least-cost choice between off-grid technologies depends on multiple, dynamic, spatial and techno-economic variables. On one hand, the centralized grid generates electricity at a lower unit cost[2] ($/kWh) than decentralized, off-grid, technologies. On the other hand, the cost of building out the grid to deliver electricity to customers is expensive (with or without expansion of the transmission network). As a result, for households which consume very small amounts of

[2] The unit cost of electricity ($/ kWh) is also referred to as the Levelized Cost of Electricity (LCOE). LCOE measures the expected revenue required to build, operate, and fuel a generator over a specified cost recovery period. (Energy Information Administration, 2021)

electricity or are remote (or both), the higher unit cost of electricity from distributed generation is offset by the money saved by not extending the grid. Conversely, the high cost of grid extension is justified in areas that are close to central infrastructure or have high demand (or both) (Morrissey, 2019; Szabó et al., 2011). In response to the dynamic complexity and magnitude of information needed to address these planning challenges, several computational models are used to automate the delineation of electrification technologies which achieve universal access to electricity at the lowest cost (Morrissey, 2019; Peña Balderrama et al., 2020). These computational models use region-specific, techno-economic and GIS data to group customers into geometric cells, calculate the Levelized Cost of Electricity (LCOE) of grid and off-grid alternatives, and output large-scale electrification plans (regional, country, or continent level) for a specified time-horizon (Ciller et al., 2019; Peña Balderrama et al., 2020).

Notable examples of large-area, GIS-based, electrification modelling tools include the Re^{2}nAF, Network Planner (NP), the Reference Electrification Model (REM) and the Open-Source Spatial Electrification Tool (OnSSET) (Amatya et al., 2018; Ciller et al., 2019; Mentis et al., 2015, 2017; Ohiare, 2015; Szabó et al., 2011). Re^{2}nAF is an open access web mapping application that enables geographic exploration of off-grid technologies in the African continent. It overlays population settlements, infrastructure features (grid network, power plants and roads) and solar resources indicators (kWh/m^2) to provide a comparison between diesel and solar photovoltaic (PV) based technologies. Unlike other large-area plans, the underlying model of Re^{2}nAF is not a project-customizable tool. Proceeding sections will discuss the remaining large-area modelling tools in greater detail. Other noteworthy literature include independent electrification studies for Uganda (Kaijuka, 2007), Kenya (Zeyringer et al., 2015), and Nigeria (Bertheau et al., 2016).

2.1.1 The Network Planner

Network Planner (NP) is an open-source model developed at Columbia University (Ohiare, 2015). The model identifies the optimal electrification technology mix for currently unserved demand centers, or nodes, and maps the estimated grid extension. Households and productive facilities are aggregated into a single demand node (i.e., model aims to make decisions at the village level). NP has been used to model Kenya (Parshall et al., 2009), Senegal (Sanoh et al., 2012), Liberia (Modi et al., 2013), Ghana (Kemausuor et al., 2014), and Nigeria (Ohiare, 2015). The necessary inputs for NP include the following: data on electricity costs and demand (often from local utilities), the location of existing grid infrastructure, population, and other socio-economic data. With this information, the NP model compares the cost of meeting demand in every settlement by calculating the respective LCOE of each of the model's prescribed distributed energy technologies. These technologies are (1) off-grid, defined as a hybrid of PV and diesel generators for household and productive use, respectively; (2) mini-grid, defined as a diesel generator plant with low voltage (LV) supply for all types of demand; and (3) grid extension, sub-divided into two cost groupings: internal and external (Ohiare, 2015). Internal, non-transmission, costs comprise the cost of equipment, distribution costs, and generation costs. External costs entail the cost of connecting a transformer at the settlement to the closest medium voltage (MV) network.

Selecting the least-cost technology involves estimating the cost of connecting a demand node to off-grid then mini-grid technology options. The option with the lowest cost is chosen and compared to only the internal costs of grid extension to the node. If the distributed generation is cheaper than the internal cost, it is selected. If not, the difference in cost determines the budget available for the external part of the grid connection costs, for that demand node. If the external cost is less than the difference, the settlement is grid connected; otherwise, the least-cost distributed energy alternative is recommended for electrification of the node. Parshall et al. (2009) describe the heuristic that NP uses to calculate the grid-extension layout.

There are multiple limitations to the NP model. First, it fails to recompute the cost of grid extension for other settlements, after a nearby settlement is recommended for grid extension and therefore the proximity of grid infrastructure has decreased and thus the external cost (Morrissey, 2019). Second, while the model aims to make village-level decisions, the data in Parshall et al. (2009) is aggregated at the sublocation level (within an average area of 15 km^2) (Ciller et al., 2019). Finally, NP does not account for the effects of grid reliability or demand increases, which play an important role in grid-extension decisions in rural areas (Ciller et al., 2019; Morrissey, 2019).

2.1.2 Open-Source Spatial Electrification Tool

The Open-Source Spatial Electrification Tool (OnSSET) is a GIS-based model developed to identify the least-cost electrification option between seven alternative configurations: grid connection; mini-grid systems with either PV, wind turbine, diesel, or small-scale hydropower generation; or stand-alone systems powered by either PV or diesel generators. OnSSET has been applied in the IEA's energy access outlook publications (IEA, 2014, 2017), to all of Sub-Saharan Africa (SSA) (Mentis et al., 2017), as well as the following countries: Ethiopia (Mentis et al., 2016), Nigeria (Mentis et al., 2015), Kenya (Moksnes et al., 2017), Afghanistan (Korkovelos et al., 2017), Madagascar (Kappen, 2019), Tanzania and Zambia (The World Bank). The model was developed by the Royal Institute of Technology at Stockholm (KTH).

OnSSET is built on a gridded population density map that segments a region into evenly sized cells (2.5-km). The model calculates the LCOE for each technology for each cell by applying a least-cost model developed by Fuso-Nerini et al. (2016). Four parameters are considered for the LCOE calculations: (a) target level and quality of energy access (measured in kWh/household/year), (b) population density (measured in households/km), (c) distance from the closest grid connection (km) and the national cost of grid electricity ($/kWh), and (d) local availability of renewable sources (solar, wind, and whether hydro potential exists within 10 km of settlements) (Fuso-Nerini et al., 2016; Mentis et al., 2015). Electrified cells are identified by means of using night-light datasets in combination with local population density, the transmission network, and the road network (Mentis et al., 2017). Initial iterations of OnSSET assumed that a population within a certain distance from the high-voltage (HV) network are electrified.

OnSSET's tree-search algorithm traverses iteratively through each unelectrified cell and tests if the conditions for their connection to nearby electrified cells are fulfilled. These conditions are: (1) a minimum number of people live in the cell, and therefore a minimum demand that justifies connection to the grid, and (2) confirmation that cell connection does not cause the total additional MV grid length to exceed 50 km. If these conditions are verified, the cell status is switched to electrified. Unlike Network Planner, OnSSET stores the length of the additional MV grid built by the model. Storing this length ensures that 1) newly electrified cells comply with the 50km limit as the grid expands, and 2) the respective cost increases of each additional MV line extension is measured. The cost increases represent the expense of building-up the generation and strengthening the centralized (existing) grid to meet the additional demand. The tree-search algorithm continues throughout the region until all cells to which the grid can be economically extended to are reached (Mentis et al., 2017). After which, the LCOE of each off-grid technology option is calculated for each of the remaining unelectrified cells and the cheapest technology is chosen. A limitation of OnSSET is that it does not include the grid's current reliability levels in its cost comparison of off-grid and on-grid electrification. Second, OnSSET considers grid extension for all cells before shifting to distributed technologies, so the lowest cost technology is not always chosen. Rather, grid extension is chosen only if it can be sustained at a breakeven price, otherwise a potentially more expensive distribution generation is selected.

Ultimately, the OnSSET model provides the optimal technology mix, capacity, and investment requirements for achieving electricity access goals under a user's pre-defined time series. A notable study by Mentis et. al. (2017) produces least-cost electrification strategies for 44 countries across SSA (25.8 million 1 by 1 km cells) for 2030 under ten alternative scenarios. The ten scenarios encompass five different future electricity demand scenarios under both high and low diesel price scenarios. Each tier represents different levels of electricity services provided, starting from basic lighting to services that provide comfort, such as air conditioning. A major limitation to this study is that the results of each scenario are calculated as if each tier is homogenously applied over the continent, which is drastically different to reality where all five tiers are likely to co-exist in each country, and therefore within the whole continent. Future continent-level studies should aim to model demand heterogeneously at high granularity (e.g., between cells or between neighboring households and productive loads).

2.1.3 The Reference Electrification Model

Each of the models introduced above work at the village-, settlement- (i.e., aggregated households), or cell-level instead of at the individual consumer level. Their network designs and cost-optimization processes are based on geometric considerations involving distances and population-based demand estimations, as opposed to power flow and electrical constraints. Both NP and OnSSET fail to optimize the generation design of off-grid systems or account for reliability constraints when delineating between electrification modes (Ciller et al., 2019). The Reference Electrification Model (REM) addresses these limitations in several ways. First, it defines and meets consumer demand at very high levels of spatial and temporal granularity, specifying hourly demand patterns for each individual customer. Second, REM meets this demand at a minimum total cost while satisfying power system technical constraints, as well as other user-defined constraints regarding the reliability of supply, generation mix, administrative requirements, or limits on modes of electrification (e.g., a preestablished target for grid extension) (Ciller et al., 2019). Third, REM supplies users with a detailed and georeferenced network layout for the prescribed grid extensions or mini/micro-grids, including the expected bill of materials. REM serves as both a least-cost electrification planning tool and a stopgap for sophisticated power flow and distribution network analyses.

REM was developed by the MIT-Comillas Universal Energy Access Laboratory (UEAL). It has been used in the design of electrification master plans in several regions including Mozambique, Rwanda, Uganda, India, Kenya, Colombia, and Indonesia. However, unlike other planning models, REM is not open source. REM delineates between on- and off-grid electrification by grouping individual customers into hierarchal structures of grid extension and off-grid clusters; the first level of the structure contains grid-extension clusters, the second level contains MG clusters, and the third level contains individual, SHS customers. REM uses a Delaunay triangulation to obtain the potential connections among customers, deciding whether to group (cluster) customers based on two conflicting driving factors: 1) the savings in generation, operation, and management costs driven by economies of scale and customers' simultaneity factors in larger MGs and 2) the increase in network costs associated with grouping customers together (Ciller et al., 2019). Figure 2.1.1(a) provides an anecdotal example of the clustering process. In Figure 2.1.1(a), customers labeled 1 and 2 should intuitively be electrified together, while a connection between 1 and 32 is only justified if economies of scale suggest aggregating all customers into one large cluster.

Once grouped, REM obtains the least cost electrification solution for a cluster by calling the greenfield network-design software RNM (Reference Network Model) to create precise network designs for a grid extension and MG, then compare the solutions by their corresponding costs. RNM employs equipment from a prescribed catalog, designs networks that meet electrical constraints, and accounts for topological features in the considered region (Domingo et al., 2011). Two possible network designs are evaluated for a MG, a LV-only network (no transformers needed) and a MV-LV network (contains MV/LV transformers with LV subnetworks; the least expensive design is selected when comparing with the grid extension option. Figure 2.1.1(c) shows examples of both types of networks. To determine these cost optimal combinations of SHS, MG, and grid extension clusters REM will simultaneously consider: 1) the layout, technical characteristics, supply costs, reliability, and catalog of equipment of the existing MV distribution network; and 2) the availability and costs of local resource generation (e.g., solar or diesel), customer preferences between SHSs and MGs, desired reliability levels, and potential dispatch strategies for off-grid supply systems. Figure 2.1.1(b) shows a possible final electrification solution for the anecdotal example in (a), where GE1 is electrified with a grid-extension design, while the remaining grid-extension clusters have lower costs when electrified with OG systems that are consistent with the hierarchical structure.

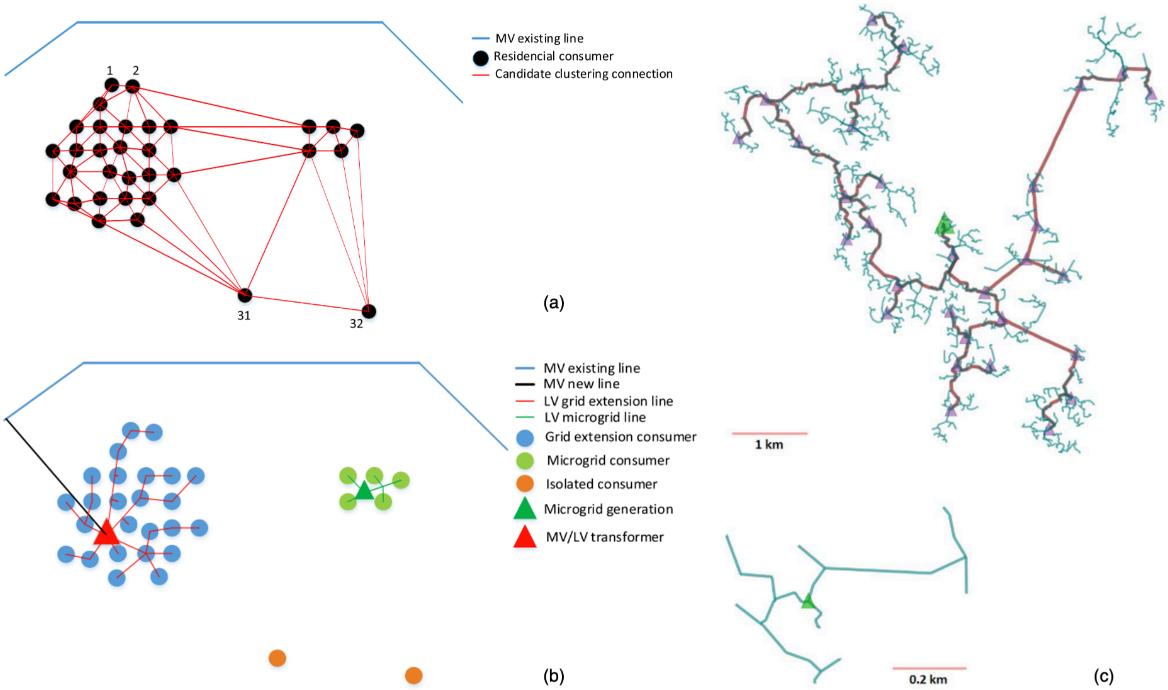


Figure 2.1.1- (from Ciller et al., 2019) (a) Clustering Candidate Connections; (b) Final Electrification Solution; (c) Two examples of MG network layouts. For (c) red line signifies a MV line, blue is a LV line, generation is a green triangle, and MV/LV transformer are the red triangles.

The final outputs of REM include the optimal groupings of individual customers into electrification mode clusters, the optimal generation mix and network layout for each of the MGs, the optimal layout for each cluster that will be connected to the grid, cost breakdowns, GIS files, generation and storage specifications, bill of materials, summary charts, georeferenced system designs. Figure 2.1.2 illustrates REM's final electrification plan for Rwanda. An important distinction between REM and other planning tools is that REM considers two types of costs when optimizing clusters: incurred and social. Incurred costs are related to investment, operation, maintenance, and management activities, which can be directly quantified. Social costs measure the loss of utility or welfare to the end consumers resulting from poor service quality and potential limitations in electricity utilization associated with their mode of connection (Ciller et al., 2019).

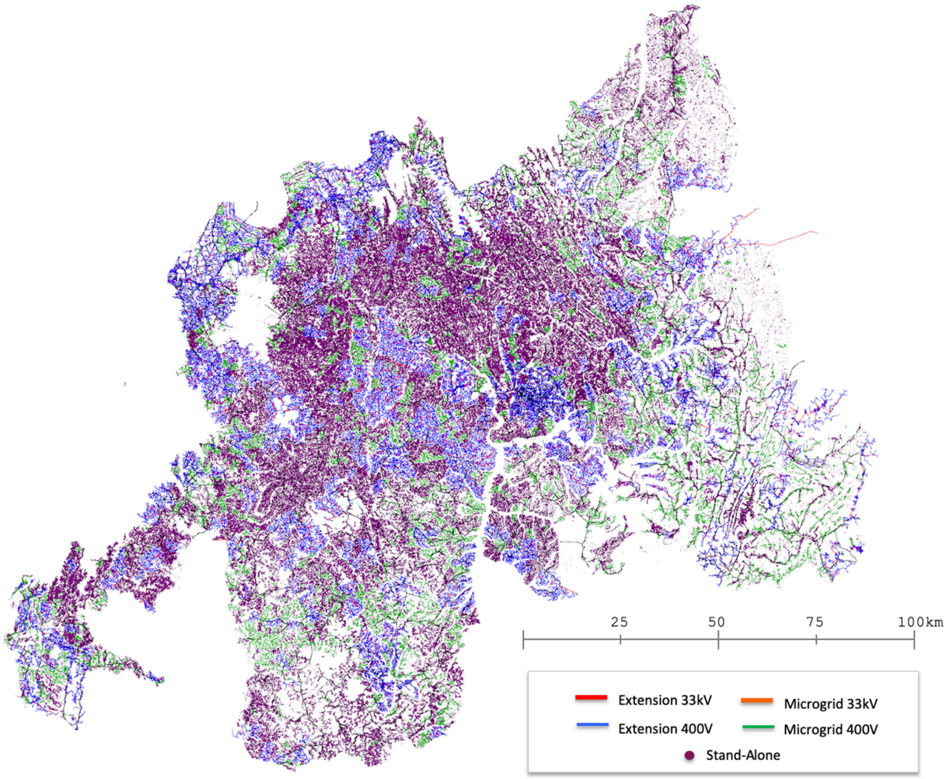


Figure 2.1.2- REM-informed plans for Rwanda's National Electrification Strategy Plan. Accessed through the UEAL website.

Describing REM Inputs

Since REM operates at the individual level, accurate, ground-truth inputs are key to its usefulness as a network designer and optimal design planner. The key inputs that must be specified prior to its use are:

1. Determining Building Locations and Building Electrification Status

Coordinates of individual buildings and characterization of the latent load profile associated with each building and whether it is electrified or unelectrified.

2. Demand Profiles for Each Type of Building/ Load

Expected hourly demand for every building- (or load) type modeled. Each load archetype is defined by a unique hourly consumption trend, modeled as either a single power utilization pattern or linear combination of several patterns. Each demand pattern is characterized by several demand samples. The samples provide alternative consumption chronologies to account for the uncertainty in customer usage, for each profile. Each demand sample specifies two consumption series: critical and non-critical electricity demand. The series consists of 8,670 values, one for every hour of the year, representing the minimum amount of demand needed to satisfy the customer's load at 100%, 24 hour a day. However, since REM is not constrained to meet all the demand, it may choose to reduce the reliability of the electricity to reduce costs and avoid significantly scaling the system. The allowed scaling of the system is defined by the user via the "cost of non-served energy (CNSE)" for critical and non-critical demands. CNSE translates supply failures to social costs.

3. Existing Distribution Network

Grid extensions require REM to both model and interpret the existing location and capacity of connected MV distribution lines. Users should supply the existing MV network as a set of line segments, each defined by two coordinates. REM uses the candidate MV connection points to evaluate the network cost (and layout) of a grid-extension solution for all consumer clusters considered. To do so, REM chooses an appropriate subset of representative MV points for each cluster as potential connection points; it then submits the problem to RNM, which is free to use any of the points to provide an optimized layout. Each MV segment is allocated an hourly reliability level and an energy price in $/kWh. The energy price accounts for differences in the locational price of delivering electricity (which allows users to translate different generation costs into the local wholesale energy costs at MV distribution level). Cost of delivering electricity is used to compute the cost of energy losses in RNM.

4. Catalog of Components: Networks and Generation Sites.

To accurately account for the cost and technical constraints associated with physical assets in the distribution network, users should supply precise technoeconomic parameters of the equipment to REM (and thus RNM). Examples of these parameters are cost and electrical features of lines and transformers. Similarly, users must provide a catalog of the cost and capacity of available generators (e.g., diesel gensets, PV panels, battery storage, charge controllers and inverters).

5. Cost Drivers and Financial Models

Includes direct monetary costs and indirect societal costs associated with the potential electrification options. Direct costs include initial investment and on-going expenditures. Social costs are estimated with the CNSE (differentiated for critical and noncritical demands). REM imposes the CNSE as a per-kilowatt hour penalty for every unit of energy demand that is not supplied. This penalizing factor ensures that system reliability is properly accounted for, while guaranteeing that supply does not become prohibitively expensive as the direct monetary costs rapidly grow with high levels of reliability.

6. Topography and Administrative Decisions

REM and RNM can adjust network costs in accordance with data on local altitudes, ground slopes, or forbidden and penalized zones. These adjustments are penalties applied during network design. The penalties capture the additional costs related to building lines over changing altitudes or crossing forbidden or prohibitively expensive terrains (e.g., lakes, forests).

7. General Electrification Criteria

Additional parameters can be used to define overall targets of the electrification plan. The most important parameters refer to required levels of quality of service and the mix of delivery modes (i.e., grid extension, MGs, and SA systems) used to meet the targets. Reliability of supply is dealt with by setting minimum acceptable levels of quality of supply in off-grid systems and including the social CNSE in the optimization.

Comparing Least-Cost Model Input Requirements

Table 2.1.3- Comparison of the required inputs for each of the large-area planning models.

Input Description	Network Planner	OnSSET	REM
Administrative Areas	●	●	●
Gridded Population Density		●	
Building Locations	●		●
Electrification Status (or Nighttime Lights)		●	●
Household Demand Profiles	●	●	●
Transmission Network	●	●	●
Distance to Grid	●	●	●
Medium Voltage Network Layout	●	●	●
Grid Equipment Catalog			●
Topography Data (Elevation and Slope of Terrain)		●	●
Travel Time to Big Cities or Road Network		●	
Power Plants and Economic Acitvities		●	●
Wind Energy Potential		●	
Solar PV Potential		●	●
Small Hydro Potential		●	
Price of Fuel (e.g., Diesel)	●	●	●

Table 2.1.3 compares the inputs required for the three most widely applied large-area least-cost electrification planning models. At first glance, OnSSET appears to require the most inputs to run and NP requires the least. However, because REM operates at the individual building level while OnSSET operates at the 1-km cell level, the same inputs required by both must be several magnitudes more precise for REM. This in turn makes assuring the quality and specify of the input datasets to REM more critical to the legibility and applicability of REM's outputs. Moreover, given REM's emphasis on technical feasibility and spatial distribution of demand, its outputs are more useful for planners especially when exploring off-grid network designs. As a result, ascertaining the availability and accuracy of its inputs should be of primary importance – especially when trying to maximize assistance to governments and private utilities in their pursuit of energy access by 2030.

2.2 Open-Access Planning Frameworks

Several frameworks aim to streamline and standardize the use of least-cost electrification planning tools. The following section spotlights two frameworks in the literature that complement my thesis objectives. Taneja (2019) develops a qualitative framework to assess whether electrification efforts are successful at driving development and investment in developing countries. Mentis et al. (2019) describe the online open-access platform *Energy Access Explorer (EAE)*, that aggregates and analyzes spatial data related to a country's unique energy supply and demand attributes. The construction and lessons learned from these frameworks were guiding design principles for my own framework, described in Chapter 3.

2.2.1 Qualitative Framework by Taneja, 2019

To develop his framework, Taneja (2019), conducts a data-driven assessment on the substantial challenges faced by Kenya when trying to secure universal, low-cost electricity to its population. From his study, Taneja identifies two primary factors which can improve the viability of electrifying all customers in Kenya without eventual insolvency (or vast profit losses) of major investors and utilities. The factors are: 1) incentivizing the consumption of electricity by all customer types and 2) the efficient and optimal allocation of scarce resources, namely money. To tackle the first factor Taneja (2019) presents three pathways for improving electricity consumption in Kenya:

1. Enabling rural customers to develop positive electricity habits. For example, improving their access to micro-financing schemes and supply chains for daily appliances. Appliances are likely to yield increased electricity consumption. Educational programs on the potential usefulness of electricity would also help weave consumption into the "fabric of their lives".
2. Developing programs that encourage and bolster local enterprises, especially in rural areas. Commercial and industrial customers in rural areas would benefit from better access to appliances, as well as gain greatly from business model support. Moreover, cross-subsidization of smaller (low-demand) consumers with large industrial customers, can make rural electrification more favorable.

3. Improving system reliability to build confidence in the electricity delivery of local distribution utilities. In response to poor grid or off-grid reliability, many customers "stack" energy sources (i.e., rely on multiple sources of power beyond their current connection). This results in redundant fees for two different electricity systems, draining the customers' scarce resources (ability to pay). Improving grid reliability would foster better relationships with industrial customers who require consistent power for the operation of their services. Grid reliability also empowers local customers to pursue opportunities of economic growth and quality-of-life improvements (e.g., enables cost-effective production of often-underdeveloped small businesses).

Altogether, the first half of Taneja's report makes the strong argument that there is a direct social cost to no, or poor electricity reliability. There is also a direct cost benefit to grid and off-grid distribution companies when they guarantee more consistent services and access to various appliances. Of the large-area planning models, REM singularly incorporates these costs and benefits into its optimization function via the CNSE parameter which further corroborates the importance of building my framework around ensuring the frequent and scalable application of REM to any country.

The second factor asks, "how can electricity planners ensure that the needs of customers are met while minimizing the cost of electricity, both now and in the future?" In order to answer this question, modelers must specify the appropriate features during the optimization of electrification modes. These features are: 1) considering the full slate of technology and productive enterprises local to a region; 2) acquiring more specific data for a region, especially local demand and proximity to existing supply sources; 3) promoting planning that enables more dynamic deployment of the different electrifying technologies, scaling the systems as customers demand more electricity (Taneja, 2019). Once again, each of these features are modelled by REM, which emphasizes the usefulness of customizing an open-data framework to its inputs.

2.2.2 Energy Access Explorer

The Energy Access Explorer (EAE) is an online, open source, interactive platform developed by researchers at the World Resources Institute. The EAE is a data aggregation and processing framework for analysts and decision-makers within the energy planning sector (e.g., rural electrification agency) to strategically plan for electrification, explore new markets for clean energy technologies, or invest for impact (Mentis et al., 2019). The platform identifies where customers are likely to be located and where there is a concentration of demand to support energy entrepreneurs. EAE complements large-area planning tools by providing a bottom-up representation of local affordability and demand. It also reduces the high transaction costs associated with data collection and sharing by serving as a database of up-to-date information.

The Energy Access Explorer incorporates data from global, national, subnational, census, remote sensing, and crowdsourcing databases either publicly available, or provided by international and local stakeholders. The data collected includes current energy demand and supply. Demand is estimated with information on household demographics and data on the social and productive uses of electricity (e.g., mines, health care facilities...etc.). Energy supply is estimated via current location-specific resource availability and infrastructure data. The platform synthesizes this data with a multi-criteria analysis tool which produces four indices for geographically targeted energy planning: 1) Demand Index, 2) Supply Index, 3) Energy Access Potential Index, and 4) Need for Assistance Index. The Energy Access Potential Index approximates where a population has the ability to pay for electricity and is close to social and productive uses of energy. The Need for Assistance Index is used to approximate where financial assistance is needed most; for example, areas with high potential to develop economically (i.e., large number of productive loads) but with low ability to pay. Finally, an important feature of this data aggregation platform is its commitment to assuring high-quality data. This is done by upholding a specific criterion for each of the included datasets, outlined below from Mentis et al. (2019).

1. *Credibility of the Source*: Must be peer-reviewed or include comprehensive citations.
2. *Metadata*: A detailed methodology is provided and clearly written.

3. *Accuracy*: Data is complete and free of errors.
4. *Accessibility*: Data is available in geospatial format and at high spatial resolution.
5. *Timeliness:* Recent-enough to be used for decision-making.
6. *Public License:* Data is public, shareable, and available for download via open license.

Chapter 3

Framework Overview and Design

3.1 Framework Design

3.1.1 Contextualizing the Framework

The effectiveness of GIS based electrification plans depends largely on the availability of credible, and up-to-date records of existing infrastructure in areas of interest (Bazilian et al., 2012; Korkovelos et al., 2019). As it stands, most of these records are gatekept by the private or governmental institutions that produced them. As a result, planners are forced to enter tedious, time consuming data-sharing agreements which can hinder or slow electrification progress (Korkovelos et al., 2019; Mentis et al., 2015, 2017). The planners' next best alternative is publicly accessible GIS data, or data (typically) produced by multiple organizations coordinating to find, organize, and openly redistribute gathered data. A noteworthy example of this is the Global Power Plant Database by the World Resource Institute (Global Energy Observatory et al., 2018). The frequently updated database contains the geolocation, generation capacity, ownership, and fuel type of 28,700 power plants across 164 countries. Cumulatively the data covers 80.2% of global installed capacity (Byers et al., 2021).

Unfortunately, few public datasets uphold similar recordkeeping standards or scale. Instead, open access information can be incomplete, of unknown origin, poorly maintained, and lack sufficient meta data (Korkovelos et al., 2019). Moreover, open access data is mostly available at low spatial or temporal resolutions and thus insufficient for most electrification planning models (e.g., REM). Higher granularities are sometimes available, but only at a premium or under special agreement (i.e., time or resource consuming, again stymieing electrification progress). Due to the limitations of manually compiled datasets, practitioners have begun taking advantage of recent advancements in satellite imagery, remote sensing capabilities, and machine learning to create open, high-resolution, geospatial datasets on global infrastructure (Adkins et al., 2017; Arderne et al., 2020; Falchetta et al., 2019; Korkovelos et al., 2019; Lee et al., 2019; Mentis et al., 2017; Sirko et al., 2021; Xie et al., 2015).

The following chapter defines a framework that makes use of both types of publicly available data to produce granular, better-quality information. Equipped with this framework, users can run REM (and its counterparts) in the absence of ground-level, proprietary, datasets. The result is much shorter lead times on characterizing regional energy infrastructure and optimal paths to meet unserved energy demand. Note that, while the framework should be transferable to alternative electrification planning tools, its design is configured specifically for REM, which is reflected in the remainder of the chapter. This choice is a reasonable one, since REM is the most sophisticated model presently available in terms of spatial, temporal and demand characterization, and thus the most data demanding model.

3.1.2 Data Availability

The extent and type of data available for a region typically varies by country. Nigeria, for example, is a data-rich country relative to its neighbors. Nigeria's data density is attributed to recent efforts by the local Rural Electrification Agency (REA) to centralize and make available energy statistics and community data, which have been collected by government agencies, donors, and private entities via the Nigerian Energy Database. The database provides information on both off-grid (e.g., potential mini-grid locations) and on-grid infrastructure (e.g., transmission and distribution lines), roads, the location of mines, healthcare services, and education facilities. Inversely, information around many other sub-Saharan African nations is often sparse and limited only to global-scale remote sensing datasets (e.g., Facebook's building locations map). Chad, for example, has a national electrification rate of 11% and lacks any ground-data on existing infrastructure or potential loads (World Bank, 2019). To account for the gulf between each country's data availability, the framework is designed in modular form such that each module represents a single REM input, and each input can have multiple underlying data modules. Section 3.1.3 provides further exploration on the modular design.

3.1.3 Specifying Framework Modules

Examining REM's Input Requirement

Table 3.1.1- Description of the inputs necessary to run REM, their relative availability, and relevant open-access datasets. Cells in black do not have a single dataset that meet the requirement completely.

Category	Input Type	Description	Status	Relevant Datasets[1]
Demand	Building Locations	(X, Y) Coordiantes of buildings	Incompletely Available	**Facebook High Resolution Population Density Maps** (30-m resolution)
	Electrification Status	Binary variable classifying building as electrified or unelectrified.	Incompletely Available	**High Resolution Gridded Dataset of Electricity Access** (1-km resolution)
	Type of Customers	Each customer type is related to a demand pattern (or a linear combination of "basic" demand patterns).	Available By-Country	
	Demand Patterns		User-Specified[2]	
Supply	(conditional) Exisiting MV Transfromer Locations	Used to derive MV grid (if unavailable).	Available By-Country	**EnergyData.Info** (platform of open data on the energy sector)
	Exisiting MV Network Layout	Used to derive grid extensions and assign region for off-grid clusters.	Available By-Country	**GridFinder** (Open source tool for predicting transmission and distribution maps.)
	Equipment Catalog	***Technical and economic*** parameters of available electrifiecation components for distribution network extension and off-grid electricity supply (e.g. solar panels, diesel generators, batteries, and power electronic equipment)	User-Specified[3]	
Regional Parameters	Topography & Geography	[1] Administrative Boundaries [2] Elevation [3] Slope of Terrain [4] Forbidden Areas	Publicly Available	**Global Administrative Area Database** (GADM, version 3.6) [1] **Aster Global Digital Elevation Model** (30-m resolution) [2-3] **MODIS Land Cover** (GLCF, v. 5.1) [4]
	Solar Input Characterization	Local solar irradiation.	Publicly Available	**Global Solar Atlas**

[1] This is not a comprehensive list of publicly available datasets per input-type, only an exmaple of the datasets applied previously.
[2] There is an ongoing effort to use advanced statistical methods and remote sensing to characterize and forecast future demand. (Lee et al., 2019)
[3] A standard catalog is available with amendable values to correspond to regional voltages and characteristics. However, for more correct implementations, users are required to use region-based grid specifications and off-grid equipment (e.g., solar panels, batteries, etc.).

Table 3.1.1 summarizes the minimum inputs required for a successful run of REM. For each input, the table provides an explanation of the data needed, classifies its data availability as one of four statuses, and links to a dataset that has been, or could be, used to fulfil the input requirement. The availability of information around each input is demarcated as one of the following classes: 'Publicly Available', 'Available by Country', 'User Specified', or 'None'. The classes are distinguished in the following way:

- *Publicly Available* means that at least one, publicly available, peer-reviewed dataset exists at the necessary resolution for all countries in SSA. In our context, peer-reviewed refers to the demonstration of sufficient due diligence around the data's quality. For instance, a guarantee that the data's accuracy has been tested and reviewed by scientific (or adjacent) organizations. Or similarly that the geospatial data and its meta information is regularly monitored and improved over time. If there are open access datasets on the input data needed, but they do not meet the requirements stipulated above, then they are not considered 'publicly available'. Similar quality standards are upheld by the Energy Access Explorer.

- *Available by Country* refers to inputs which may have datasets that meet the 'Publicly Available' requirements described above, but only for some countries or regions. Typically, the publicly available datasets which correspond to this class are compiled by local stakeholders and shared as part of a larger initiative or commitment (e.g., Nigeria's REA database).

    - A notable endeavor that must be highlighted here is the Global Electrification Platform (GEP). The GEP is a multi-phase project led by the World Bank to standardize the use of geospatial tools for least-cost electrification planning. The platform provides a high-level overview of the technology mix (grid and off-grid) required to achieve universal access by 2030[3]. For our context, GEP is most useful for its preliminary step of aggregating all public energy-related information, producing a summary per country via the open database EnergyData.Info. Their summary reduces the tedious task of ascertaining which data exists for a given study region.

- *User-Specified* inputs can be compiled to acceptable accuracy by relying on assumptions from literature, nearby regions, previous experiences. In other words, the input can be manually created and customized by users. However, this does not

[3] The current form of the GEP uses KTH's Open-Source Spatial Electrification Tool (OnSSET) to produce its findings. All results are based on publicly available information on demand and existing infrastructure.

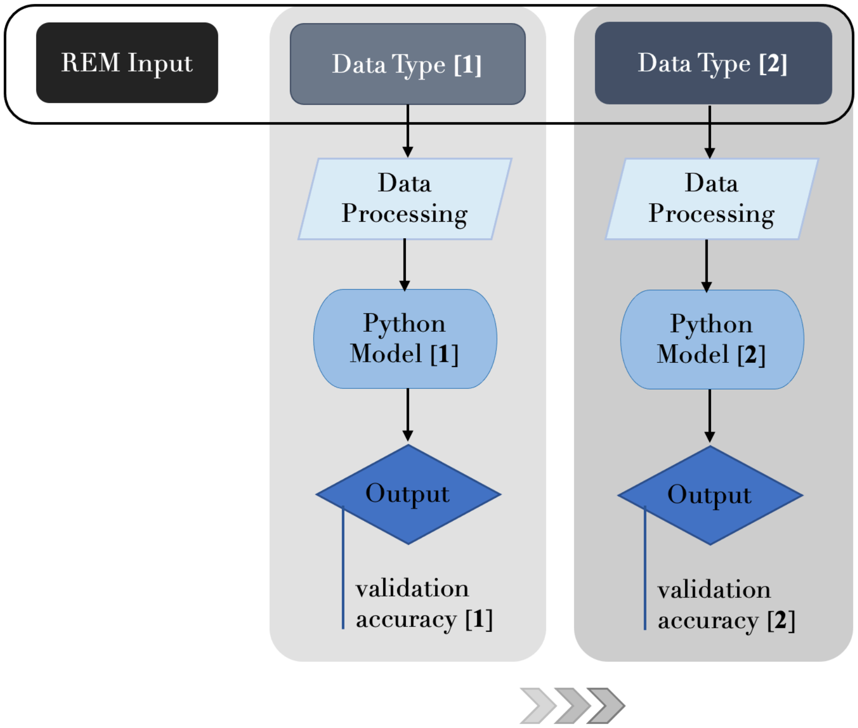
mean that users and future implementations should avoid identifying alternative, more accurate methods of fulfilling these input requirements. For example, each country typically uses specific types of cables, transformers, and poles and this should be reflected in the equipment catalog used. The same principles apply to the available batteries and solar panels by the off-grid developers. However, given the heterogeneity between what some countries share, and others don't, default grid asset types can be used in lieu of better information. Specifically, the framework provides a "standard" equipment catalog with cells that are left blank if they must vary by country (e.g., voltage levels). Depending on the purpose of the REM implementation this may or may not be OK. Coarser studies with little interest in predicting precise grid layouts can forgo the accuracy improvement of a customized equipment catalog.

- Special consideration should be given to the 'demand patterns' input. Each building will correspond to a typical electricity consumption profile, or demand patterns. Since similar building types are likely to have similar patterns and can be grouped under a broad consumer archetype. Examples of these archetypes are similar-sized households, healthcare facilities, or schools. Users should be able to provide an hourly, daily, and annual pattern for each of these archetypes. Several studies have shown that characterizing these patterns at a high level of heterogeneity and granularity is necessary for electrification planning (Kivaisi, 2000; Lee et al., 2019; Louie & Dauenhauer, 2016; Mandelli et al., 2016; Riva et al., 2019). Inaccurate models of the spatial variation of demand increases the risk of under- or over-building of infrastructure and generation, as well as non-optimal delineation of off-grid regions (Alova et al., 2021; Lee et al., 2019; Louie & Dauenhauer, 2016; Mandelli et al., 2016; Taneja, 2019). However, modelling demand in rural areas of SSA is made notoriously difficult by the innumerable confounding factors and external circumstances that influence consumer demand in these regions. Therefore, incorporating or estimating a regionally specific dataset of demand patterns for all load types from publicly available information requires further study and is beyond the scope of this thesis. In the meantime, users should refer to previous studies on consumer behavior and load consumption to guide the creation of synthetic demand patterns (Williams et al., 2017). A good example of these open studies is the World Bank's *Beyond Connections* report. For the report, the World Bank conducted multiple surveys to collect and share data on the energy demand and consumption behaviors of rural and urban residents in several SSA countries (Bhatia & Angelou, 2015). Users may also utilize data from nearby countries as a proxy for how consumers in their region of study demand and use power.

- Inputs labelled *Incompletely Available* lack any existing public dataset, which perfectly meets the requirements of *Publicly Available* inputs described above. Building locations, for example, is often imputed with the High-Resolution Population Density Maps, however this dataset lacks precise coordinates of buildings. Nor does the dataset have any reference for the size of the building or its associated customer type; thus, incompletely fulfilling the input requirements. Exceptions may occur for inputs labelled as *Incompletely Available* while high-quality public data for a single country is available, ergo users are always encouraged to begin by identifying the plethora and status of available information for their study-area.

Moving forward, the elaboration of the framework will consider only the REM inputs classified as *Available by Country* or *Incompletely Available.* This expands to the following inputs: 'Building Locations', 'Electrification Status', 'Types of Customers', 'Existing MV Network Layout'. Note that the location of 'Existing MV Transformers' is adapted into the framework as a sub-module and elaborated on in Chapter 5. The framework will utilize all information that currently exists and guide users on how to manipulate, restructure, and append the data to fulfil each input-of-interest's requirement.

The Modular Design



Each python model is written according to the type of data used.

In the direction of decreasing data type quality & validation accuracy.

Figure 3.1.2- Visual Example of how Data Types fit into the Framework. The following is an example of a two-data-type scenario for a single Input Module.

The framework is designed to consider each input separately producing a total of five modules. Each input can be satisfied using different datasets, and each dataset is handled according to a prescribed Python model, or estimation method. Consider the 'Existing MV Network Layout' input in Table 3.1.1. Although high-voltage data on transmission networks and even on high voltage distribution networks are often available, medium- and low-voltage distribution networks data are often non-existent or unavailable (Arderne et al., 2020). The framework will accordingly provide alternative estimation methods depending on the grid assets publicly *knowable* for the given study region. For example, is the geolocation and voltage of the region's MV to LV distribution transformers available? If not, then is the location of their HV to MV substations available? Some information is better than others, in this case MV/LV distribution transformer locations are preferable to HV/MV substations. The framework reflects these preferences by assigning an ordered rank to each 'baseline data' module. Here "better" refers to the input-data combinations with the relatively higher validation accuracies. Validation methods are described later in the thesis. To demonstrate how the framework functions per input and collectively see Figure 3.1.2 and 3.2.1, respectively.

The choice to make the framework modular across both the REM-input and dataset used was critical to guaranteeing it could be iteratively updated and expanded. Under the framework's current structure, higher-resolution versions of existing datasets can substitute their older counterparts at the related module. Alternatively, if a new type of dataset is created, or if a new energy-related data imputation method is developed, either can be added on as a new 'data' module within an 'input' module. Or in the case of demand, if consumer load consumption data becomes available for SSA, then it can be included as additional 'input' module in the framework. Since the objective is to expedite long-term global efforts around electrification, the framework must survive the constantly changing paradigm of data collection and sharing.

3.2 Python

3.2.1 Framework Implementation Overview

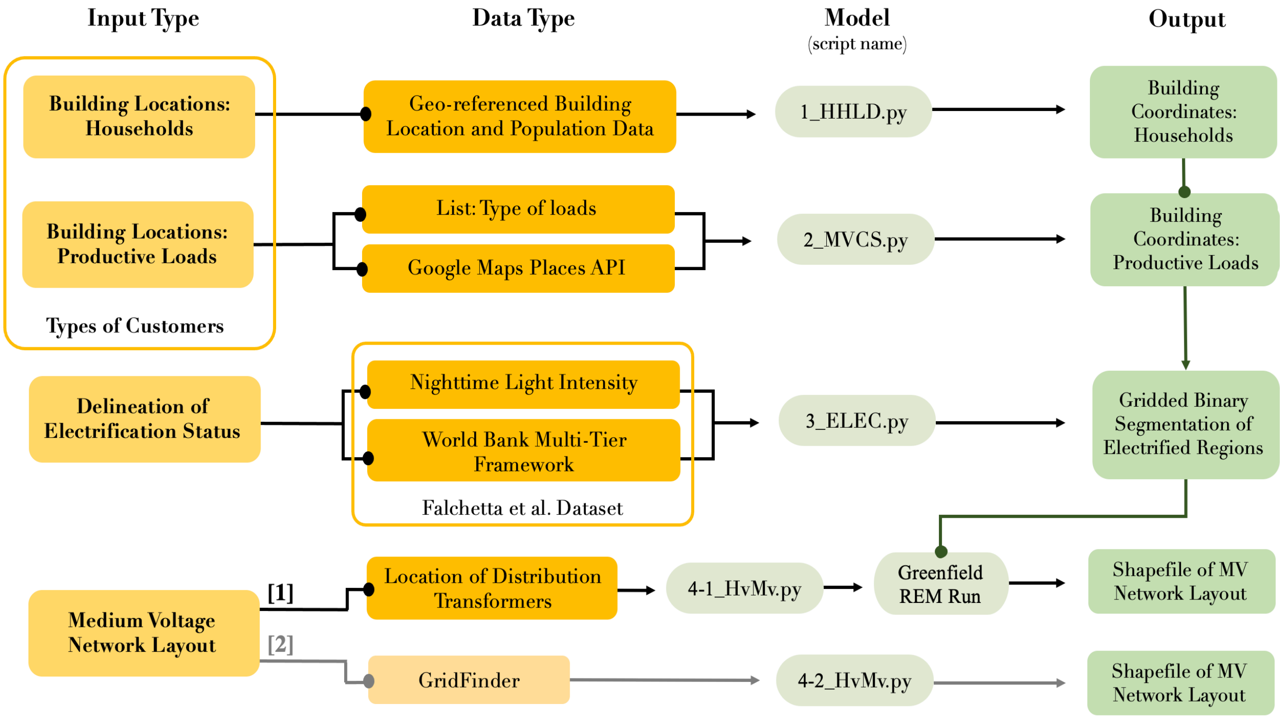

Input Type
Data Type
Model
(script name)
Output
Building Locations: Households
Building Locations: Productive Loads
Types of Customers
Geo-referenced Building Location and Population Data
List: Type of loads
Google Maps Places API
1_HHLD.py
2_MVCS.py
Building Coordinates: Households
Building Coordinates: Productive Loads
Delineation of Electrification Status
Nighttime Light Intensity
World Bank Multi-Tier Framework
Falchetta et al. Dataset
3_ELEC.py
Gridded Binary Segmentation of Electrified Regions
Medium Voltage Network Layout
[1]
[2]
Location of Distribution Transformers
GridFinder
4-1_HvMv.py
4-2_HvMv.py
Greenfield REM Run
Shapefile of MV Network Layout
Shapefile of MV Network Layout

Figure 3.2.1- Schematic of how all Input and Data Modules connect to meet REM input requirements.

Figure 3.2.1 describes the final implementation of the framework. Each row is an adaptation of the per-module process introduced in Figure 3.2. The objective of this section is to explain how users should apply the framework and the considerations made to make their interaction seamless, scalable, and open source.

Explaining the Framework Schematic

To interpret Figure 3.3, think of each yellow square (furthest left) as a 'row' and each step title (e.g., Data Type) as a 'column' in the overall framework. Then each row represents an input type, and each column represents the steps necessary to produce its estimate (refer to Figure 3.2 for additional context). The black horizontal lines signify the process to follow for the 'best' estimate of the desired output, or more precisely the method with the highest validation accuracy in my implementation. For example, two lines are drawn from the Medium Voltage Network Layout square, each tracking a separate estimation method (theoretically there can be any number of methods, each with its relative accuracy and applicability). The first line is black and the second is grey; similarly, one is adorned a [1] and the latter a [2], respectively. Both the stepwise color-lightening and ascending numeric values, symbolize the reduction in accuracy between the processes – we can think of these as rank ordered 'tiers' with higher values translating to relatively poorer outputs. In practice, a user should only use a higher numbered (less accurate) Data Type when the preceding (more accurate) tier is not available for their specific region. Figure 3.2, which narrows in on the components of a single row at a time, uses a similar color and number scheme. The vertical lines are used to convey interconnectedness between modules, for example the need to produce a specific set of inputs prior to producing another.

If a horizontal line has a dot at the end of it, then it is a *connective* node in the framework, thus the user must consider both modules at once. If a line has an arrow at the end, it signifies a process node – a flow of information between the former module and the latter. This distinction is especially important for estimation method [1] in the Medium Voltage Network Layout module, whereby the line-ends describe how preceding modules (Input Types) relate to the prescribed [1] process. Users should therefore use the framework (or acquire inputs) in the set-out order of rows shown in the figure. A user can forgo this order only in the case that some of the inputs, in the earlier or later modules, are already available for the specific study region being considered. To guarantee that later modules (which depend on earlier modules) can still be applied, instructions on the necessary data format of each output will be provided.

Design Choices Made During Implementation

During implementation of the framework the choice was made to incorporate 'Types of Customers' into the attainment of 'Building Locations'. Building Locations was subsequently bisected into two-broad building types (1) households and (2) productive loads. This is due to the lack of semantically meaningful distinctions between buildings provided in current publicly available datasets. For example, information on building footprint size or shape is usually only available at a premium. Households are low-voltage consumers with a narrow range of differing load consumption patterns. Productive loads include everything from small commercial businesses (consuming at the low-voltage level) to industrial customers such as mining companies (consuming at the medium-voltage level). In more advanced implementations, these broad building types can be further delineated into those located in rural vs. urban areas, with the presumption that rural loads consume power differently and less reliably. For example, rural households often consume outside of archetypal peak hours (between 8:00 am- 5:00 pm), using power mostly at night. Rural customers also tend to consume much less power than those in urban regions due to prohibitively expensive electricity prices. Rural households primarily use basic appliances like fans, radio, and lighting (Bhatia & Angelou, 2015; Szabó et al., 2011). I propose a work around for the lack of segmentation between rural and urban buildings in Section 4.3.

3.2.2 Incorporating Python

All processing and formatting of data happens in the Model column using Python. The source code is available through GitHub, with a readme.md file to support users in applying the framework. The choice to use Python stems from a commitment to creating an open, scalable, reproduceable framework. Today, Python is the industry and academic 'gold-standard' for open-access implementations. As a well-documented and widely used programming language, Python is an ideal common interface on which future users can build new processing chains and audit previous implementations with. Ensuring that my framework is transparent is critical to assuring quality control and demonstrating adequate due diligence for stakeholders to trust its findings. In this way, the framework fosters open-science and open-source coding in the field of energy access research – yielding rapid techno-economic screening analyses and expediting the achievement of SDG7 (Korkovelos et al., 2019).

Each python model is set up to run independently of the other modules. Each model consists of three files: a functions module, a script, and a data input requirements file. All the data processing occurs between the function module and the main script file. The data requirements file explains where to download the module's prescribed datasets and how to store them locally (including the appropriate naming nomenclature). For models that have vertically integrated inputs (i.e., its inputs are outputs from previous modules) the data requirement file also describes how alternative inputs can be integrated instead. The description specifies the exact scale, structure, and format the data should be in. The supplemental instructions are necessary for contexts where a country may already have publicly shared the REM-input of a preceding module. The functions module is imported into the corresponding model's script file. Users interact with each model by first inputting the country code they are interested in, e.g., 'RWA' for Rwanda. The model will extract the key country statistics that it needs to run the script and validate its output. Note, the framework currently runs at country-scale; therefore, if a user is interested in just a small region within a country, they will need to clip all output files using a Mask (geo-referenced vector outline) of their region-of-interest. Recall, the outputs of each model will be configured to meet REM input-file format requirements. Finally, a 'requirements.txt' file is provided with the framework to guarantee any user can run all Python scripts. The text file specifies all the packages needed to run the code and is used during the initialization of the Python virtual environment.

The GitHub Repository can be accessed at:
https://github.com/lamaoudi/OpenSourceFrameworkforLCEMs

3.2.3 Validation Processes and Metrics

In this thesis we use previously gathered data from a project with Rwanda to check our estimates against ground-truth data. However, the data is outdated in some instances and limited in scope, thus an imperfect benchmark. Furthermore, using a single country for validation does not measure the robustness of our methods across different countries. As an alternative to ground-truth information, the framework also employs publicly shared country-statistics as easy metrics to check the quality of the estimates. Further discussion on validation metrics will be explored in coming chapters.

Chapter 4

Data Sources and Processing: Consumer Modules

This Chapter looks at each Input Type module, provides an overview of useful datasets, and proposes a method for estimating the Input. Note that the proposed methods are not intended as the 'best' or prescribed method of imputation; only as my suggestion for producing REM inputs quickly and to any country in SSA (i.e., the applied data types are available across the region). Extensive research is being conducted elsewhere on ascertaining electrification status, identifying settlements and settlement types, as well as predicting grid infrastructure. Hence, more advanced techniques or datasets may be available to implement the framework with, producing more accurate outputs. Typically, these methods exist only for limited regions, require additional validation, or are costly to apply at scale. Therefore, the users must discern for themselves which parameters are higher priority: accuracy, simplicity, or speed – then choose the input and data modules accordingly.

Section 4.1 covers the estimation of residential-building coordinates. The section includes a review of publicly accessible population and building footprint datasets and describes how to apply them to locate households in any study-region. Section 4.2 proposes using the Google Maps API to extract the coordinates of productive and community facilities. Section 4.3 looks at the existing literature on estimating the electrification status of buildings, and how to couple this information with the World Bank's definition of five energy access tiers. The tiers are then used to estimate the maximum power demanded by each load-type. In each Section I compare the outputs of my implementation to ground-truth data from Rwanda.

Chapter 4 should motivate how powerful open-access frameworks can be for scalable technoeconomic analyses in energy access and related fields. It should also re-iterate the importance of a modular system, given the value of substituting my implementations with better datasets and models as they are made available.

4.1 Identifying Building Locations: Households

One of the most crucial inputs to run any electrification planning tool is the accurate knowledge of where people reside and at what density. This question arises for a variety of disciplines including public security, health policy, urban growth, vulnerability and risk assessment among others (Palacios-Lopez et al., 2021). Geospatial population data has also been shown to support, implement, and monitor more than half of the SDGs and their related indicators (Kavvada et al., 2020; Kuffer et al., 2020; Palacios-Lopez et al., 2021; Qiu et al., 2019). As a result, research on compiling and improving population data is plentiful and ongoing. The following section explores notable datasets and highlights how one can leverage the information to fulfill REM's building coordinates requirement. The alternative Data Types enables users to apply the framework at their preferred granularity and scope. For example, broader continental-level studies may favor speed over accuracy, and therefore choose to employ coarser, smaller file-sized population datasets which minimize processing time.

4.1.1 Identifying Building Locations: Households

The following section is a comprehensive review on alternative Data Types that users can investigate as potential sources of building location information.

Review of Population and Built-Area Datasets

Population modelling historically consisted of gridded population datasets. The grid is comprised of spatially identical cells, where the size of the cells determines the spatial resolution of the data, or the land-area covered per cell. There are several openly available, large-scale (continental and global) population gridded datasets, each has been produced using a different "top-down" dasymetric modelling approach. In other words, each employs its own method of disaggregating administrative unit-based official population counts into the grid cells (Leyk et al., 2019; Palacios-Lopez et al., 2021). Each of these disaggregation methods choose a different set of ancillary geospatial datasets to model, and in some cases restrict, the distribution of population across space (Palacios-Lopez et al., 2021). In recent years, population grids have improved in quality, accuracy, and spatial resolution by the emergence of built-area datasets. Built-area datasets use remote sensing techniques (optical and radar imagery) to accurately describe the extent, location, and characteristics of urban and built-up areas. The detected infrastructure is then used to accurately delineate the distribution of human settlement (Palacios-Lopez et al., 2021; Reed et al., 2018). Hence, population attributes (e.g., counts) are only assigned to areas with detected infrastructure, providing a high-resolution map of human settlement.

As REM requires the explicit coordinates of buildings in its study-region, built-area datasets serve as suitable substitutes for ground-truth settlement information. Representative, publicly available, examples include: the World Settlement Footprint 2015 (WSF-2015) map; the new WorldPop Sub-Saharan Africa gridded building dataset (coupled with 100-m resolution population counts); the Global Urban Footprint (GUF); the High Resolution Settlement Layer (HRSL), and the Global Human Settlement Layer (GHSL). I provide a brief review on each of these datasets for additional context to how they differ and might be applied:

1. (Linard et al., 2012) WorldPop Sub-Saharan Gridded Building & Population Datasets; The gridded population layers are available for many (not all) SSA countries at 100-m, and available at a 1-km resolution for continent level density information. Annual population counts are available from 2000 to 2020. These layers use interpolation techniques which sometimes give rise to inaccurate population estimates, and at times infeasible values (e.g., <1) (Korkovelos et al., 2019). Their gridded building dataset is available at 100-m resolution for 51 countries in SSA. The underlying footprint information was not validated prior to compilation or processing.

2. (Ouzounis et al., 2013) Global Human Settlement Layer (GHSL); A global 100-m spatial resolution built-up dataset that focuses on three primary products: built-up areas, population grids, and urban/rural classification. The residential population estimates are provided for target years 1975, 1990, 2000, and 2014, expressed as number of people per grid cell. Built-up areas are semi-automatically extracted from Landsat-8 optical imagery, collected in 2014 at a resolution of 38-m (Marconcini et

al., 2020; Tiecke et al., 2017a). GHSL applies CIESIN's same census disaggregation method as HRSL, described in detail below. It was created by and is available through the European Commission's Joint Research Center.

3. (Esch et al., 2017, 2018) Global Urban Footprint (GUF); Available globally at 12-m resolution and referring to the year 2012. Generated from single-date scenes, which can be strongly affected by the specific acquisition conditions, hence resulting in misclassification errors. GUF exclusively uses optical and radar imagery, which are sensitive to different structures on the ground. Bare soil and sand tend to be misclassified as settlements in optical imagery. Complex topography areas or forested regions can be wrongly categorized as settlements with radar imagery.

4. (Tiecke et al., 2017a) High-Resolution Settlement Layer (HRSL); Created in partnership with Facebook's Connectivity Lab and CIESIN[4]. HRSL provides estimates of human population distribution in 140 countries at 30-m resolution (1 arc-second) for the year 2015. The Connectivity Lab used computer vision techniques to classify pixel data as containing buildings or not, then CIESIN allocated population data from subnational census data to the settlement extents. The method of extracting building footprints using Machine Learning (ML) algorithms is described in detail below. Census unit boundaries and associated population estimates from the Gridded Population of the World Collection (GPWv4) are used. The census unit boundaries collected for GPWv4 vary by administrative level, number of units, average unit size, and census year for each country (Tiecke et al., 2017a). Since the population density values are equally distributed to settlements in a census unit, its usefulness increases at higher administrative levels (i.e., tighter districts). The HRSL has demonstrated superior accuracy in estimating settlements in rural areas compared to the GHSL and GUF datasets, shown in Figure 4.1.1 (Tiecke et al., 2017a).

[4] Center for International Earth Science Information Network at Columbia University

5. (Marconcini et al., 2020) World Settlement Footprint (WSF) 2015: A 10-m resolution (0.32 arc sec) global map of human settlements on Earth for the year 2015. Validated on 900,000 ground truth samples collected by crowdsourcing photointerpretation, carried out in collaboration with Google. (Marconcini et al., 2020) show that WSF-2015 outperforms the GUF and GHSL on the validation samples and is best at detecting very small settlements in rural regions and scattered suburban areas. Settlements are derived by means of 2014-2015 multitemporal Sentinel-1 (S1) radar and Landsat-8 optical imagery. Unlike the GUF, WSF-2015 addresses misclassification by utilizing multitemporal optical and radar data. The rationale is that the temporal dynamics of human settlements over time are different than non-settlement classes, and therefore can be used to distinguish between them. Unlike its competing datasets, WSF-2015 has not disaggregated population data to its settlements, therefore relying on the dataset as a proxy for household distribution requires that the data is segmented into individual building footprints, or something similar. Figure 4.1.2 provides a visual comparison of WSF, GHSL, and GUF.

The following list is not comprehensive, nor intended as a comparative analysis of publicly available population datasets. Leyk et al. provides a complete review on large-scale gridded population products (Leyk et al., 2019). Additional resources on built-area datasets include (Khavari et al., 2021; Lloyd et al., 2019; Palacios-Lopez et al., 2021; Reed et al., 2018).

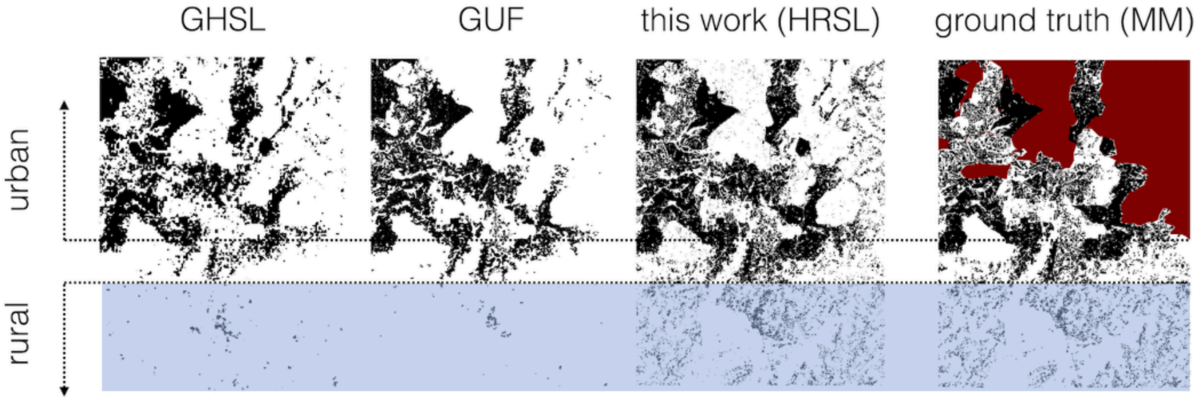


Figure 4.1.1- (from Tiecke et al., 2017) Qualitative comparison of GUF, GHSL, HRSL, and labeled ground truth data (via Missing Maps, not mapped areas are indicated in red) for a single region near Blantyre, Malawi. The recall values for the southern rural areas (below the lower dotted line, shaded in blue) are 0.04, 0.06, and 0.84 for the GUF, GHSL, and HRSL respectively.

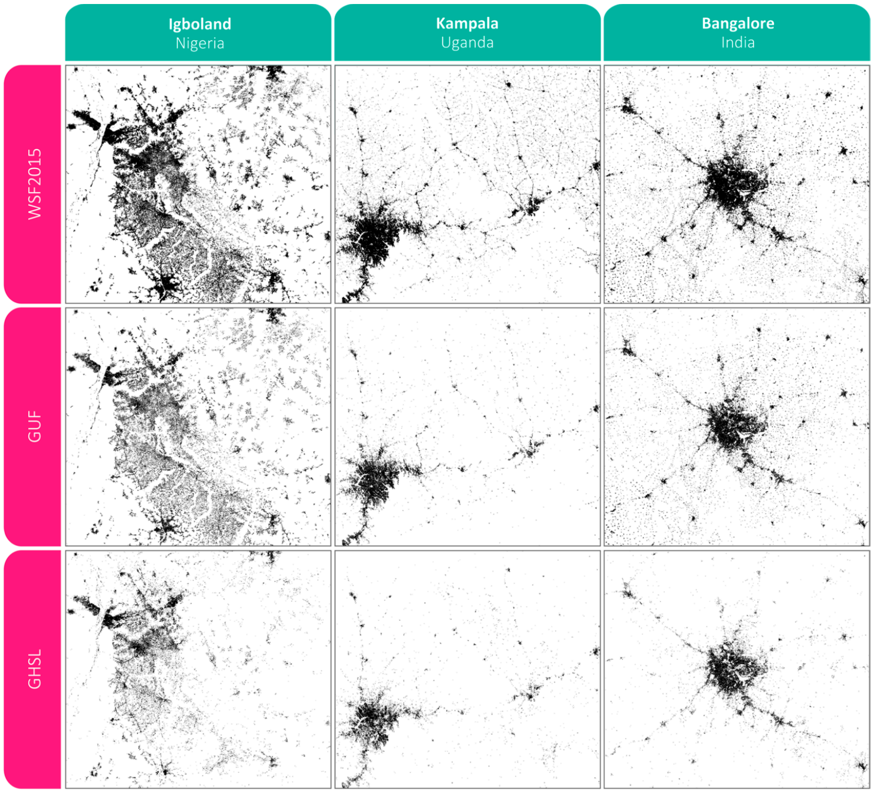


Figure 4.1.2- (from Marconcini et al., 2020) Qualitative Comparison of the WSF-2015 against GUF and GHSL. Samples are reported for Igboland, Nigeria; Kampala, Uganda; Bangalore, India. Each sample is characterized by the presence of medium and large size cities, surrounded by very small settlements. The WSF-2015 outperforms other datasets by detecting a higher number of small villages and better outlining the fringes of major urban areas. The GUF performs equally good only in the Igboland region and not in Kampala or Bangalore.

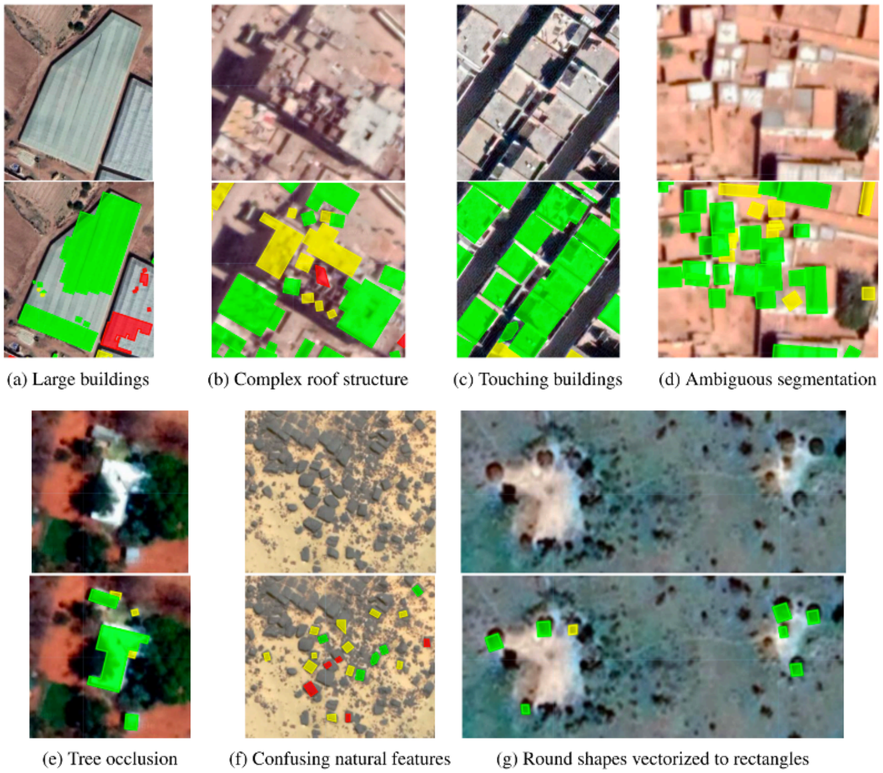
Machine Learning Techniques for Building Location Estimation

The primary alternative to using built-area population datasets for building coordinates is building footprint extraction. Building footprint extraction involves classifying each pixel of a high-resolution satellite image as either containing a building or not. Manual analysis of pixels remains the most accurate and precise way to carry out this method. Its drawback is the high cost and prohibitive time needed to compile the data. More importantly, since electrification planning tools are updated frequently, manual extraction will need to be repeated at each update, which would not allow for the wide application of REM, or adjacent planning tools. A notable manually labelled building database is OpenStreetMap (OSM). OSM provides free and open detailed building and street annotations through crowdsourcing: millions of participant's conduct ground-based surveys and perform manual labelling on the top of aerial imagery (Ciller et al., 2019). However, OSM data is inconsistent and can be sparse or non-existent in rural regions of SSA – hence not useful in our context.

The automatic detection of building footprints is dominated by machine learning (ML) computer vision techniques, namely Convolutional Neural Networks (CNNs). Previous work has demonstrated the success of CNNs at detecting and delineating building footprints from satellite imagery. Typically, semantic segmentation is first carried out to classify each pixel in an image as building or non-building. Then shapes and boundaries of individual buildings are extracted, either by layering different CNN architectures, contouring algorithms, or leveraging a host of other image post-processing techniques. Two wide-coverage datasets apply CNNs to produce high-resolution settlement datasets: Facebook's HRSL and Google's Open Buildings Dataset. HRSL produces binary labeled pixels at 30-m resolution (building or no building), while the Open Buildings dataset identifies the extents of building footprints at 50-cm resolution.

A limitation to automatic detection methods is the reliance of their performance on available pixel-wise labeled training samples (Sirko et al., 2021). Without adequate inclusion of Africa's diverse terrain and building types, the classification models may struggle to consistently detect settlements across the continent. For example, aerial images of geological or vegetation features may be confused with built infrastructure, while buildings constructed with natural materials (or non-generic roof structures) may visually blend into the surrounding areas in rural or desert regions – going undetected (Sirko et al., 2021). The Open Buildings dataset addresses the potential misclassifications by curating a training set with a mix of rural, medium-density and urban in areas in different regions of the continent, including informal settlements in urban areas as well as refugee facilities. Since contiguous, densely packed, buildings were difficult to delineate a 'dense building' class was introduced. Examples of the diverse structures labelled are: round, thatched-roof structures (distinguished via cues like pathways and clearings); compounds containing dwelling places as well as smaller buildings (such as grain stores) (Sirko et al., 2021). The Open Buildings training set was human-annotated and a laborious accomplishment. Examples of its final outputs are shown in Figure 4.1.3.

Preceding its competitor by several years, the HRSL opted against a supervised-learning approach, afraid of its intractability when scaling globally and the large possibility of accumulating errors during supervision. Instead, the HRSL employs a weakly-supervised learning approach by using only easy-to-get image level categorical supervision into training and preforming pixel-level prediction (Bonafilia et al., 2019; Tiecke et al., 2017b). This in turn makes it less robust to heterogeneous landscapes and reduces overall performance of the dataset.



(a) Large buildings (b) Complex roof structure (c) Touching buildings (d) Ambiguous segmentation

(e) Tree occlusion (f) Confusing natural features (g) Round shapes vectorized to rectangles

Figure 4.1.3- (from Sirko et al., 2021) Examples of the Open Buildings Dataset footprint extraction. A powerful, meta-component of the dataset is its "confidence score" assessment- which quantifies the relative accuracy of detection for a particular building. The color of the polygons in the figure above reflects the confidence score range: red [0.5; 0.6), yellow [0.6; 0.7), and green [0.7; 1.0]. Panel (f) is a good example of an instance where natural terrain is mistaken for built-up infrastructure, fortunately most of the mis-identified footprints in (f) are marked with low confidence levels and can be ignored if a user applies threshold-confidence values to their implementation of the dataset.

4.1.2 Applying the Framework: Rank Order Data Types

An essential component of my framework is its incorporation of multiple Data Types, ranked by their respective accuracy. Given that no formal quantitative assessment was conducted between all the datasets listed in Section 4.1.1, I rely on the validation and sensitivity analyses run by each source to ascertain its data's performance quality. To select the top three Data Types for the Building Coordinate input, I prioritize the following data characteristics: (1) high granularity, (2) broad coverage, and the availability of either (3) segmented building extents, or (4) population counts. Accordingly, Google's Open Building Dataset, HRSL, and WSF-2015 are singled out. Recall, GUF and GHSL have been proven to perform worse than the WSF-2015 and HRSL (Marconcini et al., 2020; Tiecke et al., 2017b).

The Open Building Dataset is ranked 'best' because it detects buildings at the highest granularity (50-cm resolution satellite imagery), provides the exact extents of building footprints, and couples each footprint with a meta statistic on its accuracy. And with the geographic heterogeneity of the Open Building's training set, we can expect the model to perform consistently across the continent. This is unlike HRSL and WSF-2015 which struggle to discern atypical roofs or building-looking natural features (e.g., rectangular rock formations) and hence perform best in areas with more traditional settlement-types and little vegetation. The HRSL and WSF-2015 also fail to provide building extents or a formal count of the buildings detected within the land-area of a pixel. Instead, each dataset only classifies pixels as with- or without- settlement. However, unlike its competitor, the HRSL layers its binary classifier with the average population density of the local area – information which can be used to estimate building counts. As a result, even though the WSF-2015 is available at a higher resolution (10-m or 100-m^2 vs. 30-m or 900-m^2, respectively), the HRSL is preferred. While a user can manually overlay gridded population datasets with the WSF-2015 data, it adds additional post-processing steps and becomes less tractable for fast and easy implementations. Furthermore, the HRSL is easily downloadable for all countries in SSA via EnergyData.Info, with each country's density raster being continually updated as CIESIN releases improved gridded population datasets. However, if a user prefers the higher-resolution of the WSF-2015 they can easily interchange it for HRSL in the coordinate extraction method outlined in Section 4.1.3.

Figure 4.1.4 visualizes the ranked ordering between each Data Type. Because the preference between HRSL and WSF-2015 is not necessarily associated with higher accuracy levels and only the ease-of-use of dataset, the shading between [2] and [3] are identical. Keeping the coloring consistent eliminates any visual implication that a choice between either data type would sacrifice input estimation performance. Finally, due to the modular design of my framework users can easily interchange between all three options if it better suits their study parameters.

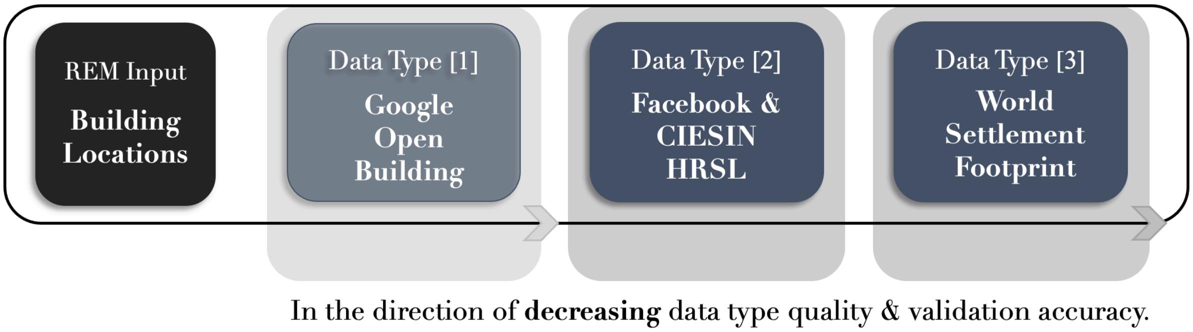


Figure 4.1.4- Shortlist of Building Information Data Types. The depicted sequence is the preferred order by which users should choose a Data Type. The shading between [2] and [3] is identical since there is no implication that the latter would reduce quality of the data, only the ease-of-use of the data. The assumed accuracy level is inferred from each dataset's attributes and not from formal quantitative assessments.

4.1.3 Applying the Framework: Estimating Household Locations with the High-Resolution Settlement Layer

The following section outlines how to use the HRSL dataset[5] to produce the building coordinates input. Recall that my framework segments building types into 'households' and 'productive loads'. Productive loads encompass larger loads with unique consumption profiles, often associated with economic returns. Examples include health centers (ranging from small clinics to state hospitals), educational facilities (including universities and technical institutes), mines, and large industrial plants. Distinguishing between the type of loads and their proper spatial allocation is critical during REM's estimation of on- and off-grid network layouts. Further detail on types of productive loads and their role in electrification planning is available in Section 4.3. Given that HRSL does not provide any ancillary information on the type or size of detected settlements, little can be said from its data on which pixels contain those larger loads. As a result, I assume all settlements by the HRSL are households, or low-voltage loads.

The HRSL is available in two primary formats: a Raster file (GeoTIFF) or a CSV. In Raster format a pixel is non-zero only if a settlement is detected and zero everywhere else,

[5] The Open Buildings dataset was not utilized in this study because it was only made available recently, in July of 2021.

and the band value is set to the local population density. The CSV file provides the centroid of each non-zero pixel and the associated population density value. Since DataFrames are more tractable than images in Python, my method uses the CSV format. Figure 4.1.5 demonstrates the differences between both file types, by layering both on daytime satellite images of the Kimisagara (blue outlined) and Kigali (orange outlined) neighborhoods of Rwanda. Each sector's data is marked in a different color in accordance with its band (population density) value. A closer look at Figure 4.1.5(c) shows how large a 30-by-30 m area is on-the-ground and the multiple buildings (or households) that can reasonably occupy that area. To produce the coordinates of all households in a country, I first estimate the number of houses per pixel and assign each of those houses to a set of coordinates within the pixel.

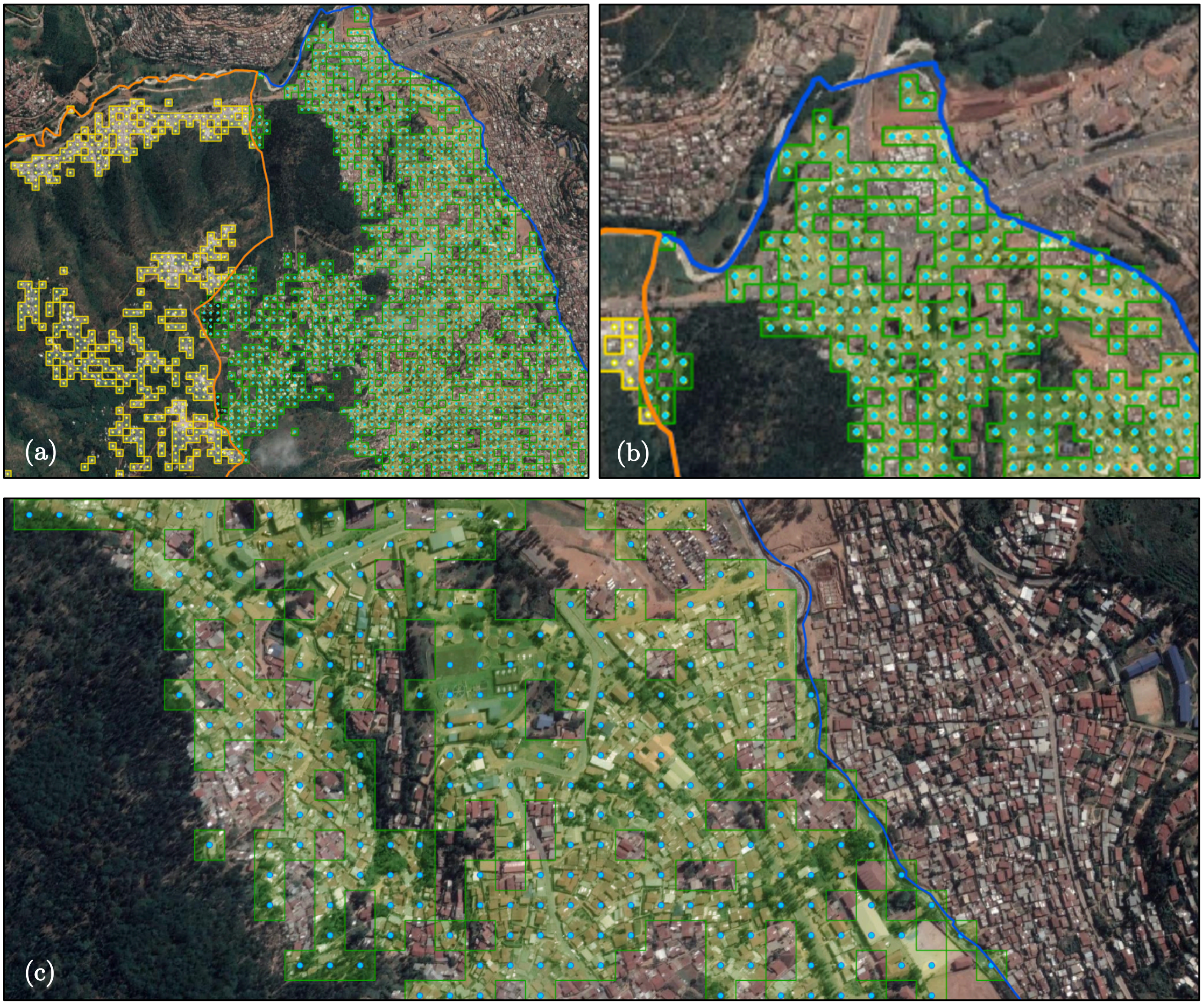


Figure 4.1.5- Raster and CSV HRSL data visualized on satellite imagery of Rwanda. Image (b) and (c) are closer-up perspectives of the north and west areas in image (a), respectively. The region outlined in blue is the Kimisigara neighborhood and the Kigali neighborhood is outlined in orange. The tiles are the raster image and the circular marker in the center of each tile is the CSV's centroid coordinates. Color of each tile or centroid signifies the local population density.

The density rasters are created by dividing the UN WPP-adjusted population count raster for a given target year by the land area raster. HRSL assumes equal population distribution per building within a census area, as a way to make no assumptions on the number of people per building and constrain systematic errors to one census area (Tiecke et al., 2017a). I instead assume the density values are a reasonable approximation for the number of people living within a square. I also assume that the average size of a household is the same everywhere in a country, and I obtain these statistics from the Open Data for Africa[6] portal. Under these assumptions, I divide a pixel's density value by its country's average household size, round down to the nearest integer and approximate the number of houses within the pixel's land-area. The number of houses per pixel will be hereon referred to as hhld_num. To assign latitude and longitude coordinates to each of these houses, I apply the following protocol to each pixel:

- Extract the bounding x- and y- coordinates of the pixel, in other words the minimum and maximum coordinates of the orthogonal sides. Recall that x-coordinates refer to longitudinal degrees and y-coordinates refer to latitudinal degrees. To extract the bounds, I use the cell size and measure the coordinates half a distance away in both directions from the centroid. Since the size of the cell may differ incrementally depending on where you are with respect to the equator, I suggest manually re-calculating the size of the cell per country. Also, if either the latitude or longitude of a pixel is negative, it will affect (flip) how you set the maximum and minimum values.

- For each direction (x- and y-): Produce a 1-dimensional array of 10 equally spaced coordinates (each 3-m apart), that are bounded by the pixel's extents. Together the arrays create 100 unique combinations of x- and y-coordinates, forming a grid of 100 smaller land-areas within the pixel. I assign each house belonging to the pixel to one of these 100 land areas, at random. Figure 4.1.6 provides a diagram of how each of these arrays combine to form the 10-by-10 grid. In the diagram the digits in grey are the indices of each 1-D array. The grid consists of the pairs of indices which produce the respective [x-coordinate, y-coordinate] centroid of the small land-area.

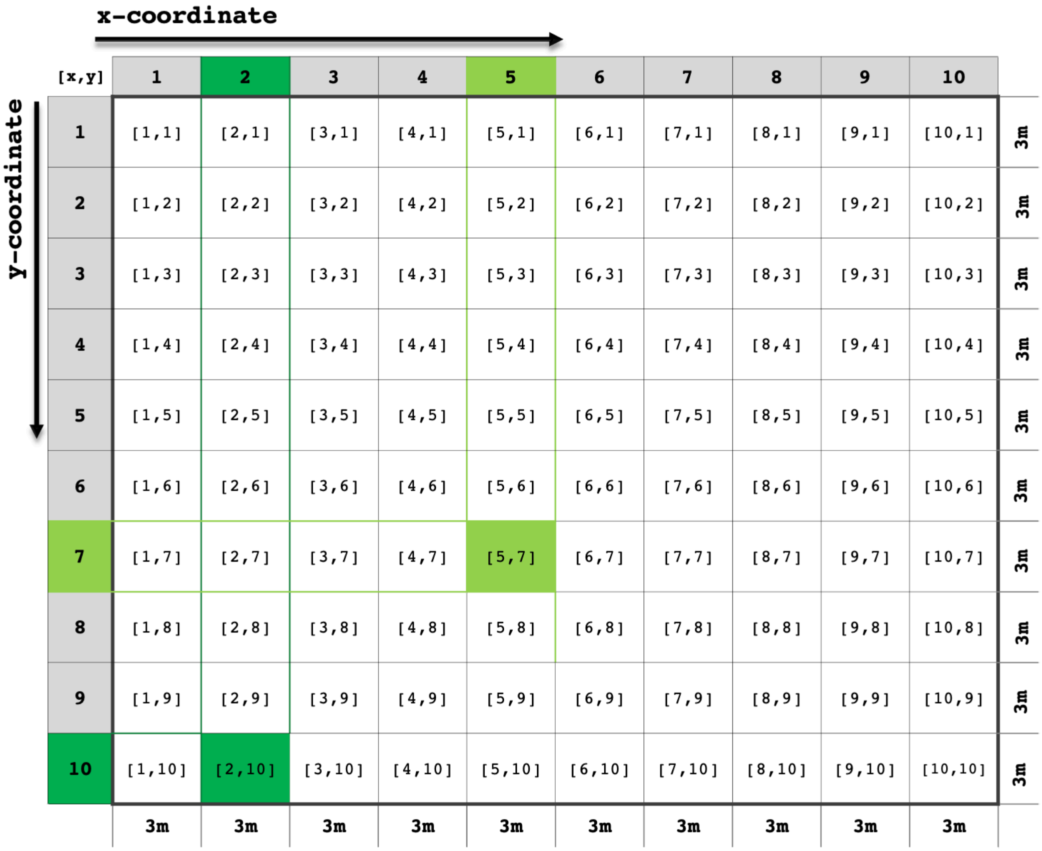
- Random sampling land areas: Sample an integer between [0, 10) from a uniform distribution. I then use the integer to index the x-coordinate array and extract a

[6] Open Data for Africa is a statistical data portal put together by the African Development Bank. The portal aggregates all statistical information and socio-economic indicators relating to African countries.

coordinate. I repeat the sampling process for the y-coordinate array and assign the house to the small land-area with that centroid.

- Sampling x- and y- coordinates is repeated as many times as hhld_num evaluates to per pixel – in other words, for as many houses estimated to be within the pixel. If hhld_num evaluates to a value below 1 then we assume no household exists.

- Store the coordinates of each household in the same list while iterating between houses in a pixel and pixels in a study-region.



[x,y]	1	2	3	4	5	6	7	8	9	10	
1	[1,1]	[2,1]	[3,1]	[4,1]	[5,1]	[6,1]	[7,1]	[8,1]	[9,1]	[10,1]	3m
2	[1,2]	[2,2]	[3,2]	[4,2]	[5,2]	[6,2]	[7,2]	[8,2]	[9,2]	[10,2]	3m
3	[1,3]	[2,3]	[3,3]	[4,3]	[5,3]	[6,3]	[7,3]	[8,3]	[9,3]	[10,3]	3m
4	[1,4]	[2,4]	[3,4]	[4,4]	[5,4]	[6,4]	[7,4]	[8,4]	[9,4]	[10,4]	3m
5	[1,5]	[2,5]	[3,5]	[4,5]	[5,5]	[6,5]	[7,5]	[8,5]	[9,5]	[10,5]	3m
6	[1,6]	[2,6]	[3,6]	[4,6]	[5,6]	[6,6]	[7,6]	[8,6]	[9,6]	[10,6]	3m
7	[1,7]	[2,7]	[3,7]	[4,7]	[5,7]	[6,7]	[7,7]	[8,7]	[9,7]	[10,7]	3m
8	[1,8]	[2,8]	[3,8]	[4,8]	[5,8]	[6,8]	[7,8]	[8,8]	[9,8]	[10,8]	3m
9	[1,9]	[2,9]	[3,9]	[4,9]	[5,9]	[6,9]	[7,9]	[8,9]	[9,9]	[10,9]	3m
10	[1,10]	[2,10]	[3,10]	[4,10]	[5,10]	[6,10]	[7,10]	[8,10]	[9,10]	[10,10]	3m
	3m	3m	3m	3m	3m	3m	3m	3m	3m	3m	

Figure 4.1.6- Diagram of how a single pixel is segmented into 100 small land-areas.

The green-shaded squares are examples of how the uniformly sampled values between [0, 10) leads to a unique pair of x- and y- array indices, which refer to a specific square (or land area) in the formed 10-by-10 grid.

Once the protocol has been applied to all pixels in the study-region, save the data as GeoDataFrame and export the file as a shape file. To see how well this method estimates the local population, I multiply the total number of households by the average household size and compare the estimated population count by the country's most recent reported Total Population. If the estimate is far below the total population count, I recommend rounding up instead of rounding down hhld_num to the nearest integer. So far, my results show that this method overestimates the population count by approximately 15%. Future efforts should explore when the instances of overfitting or underfitting occur and propose more robust solutions. Potential questions to explore include: How does the following method hold up in rural, peri-urban, and urban regions? Are there drastic differences? Additionally, I opt for a broad statistic validation exercise because it is the type of data that is uniformly available for all countries. However, the exercise is imperfect at capturing proximal differences between ground-truth and estimated data. More sophisticated validation exercises are recommended if actual household data is available for some regions in a user's country. Figure 4.1.7 is a summary of how to use HRSL (or WSF-2015) to extract building coordinates. Three additional studies propose methods of translating high-resolution gridded datasets to more granular settlement information. Korkovelos et al. (2019) outline how to process the HRSL dataset to provide more accurate vector type settlement layers to apply on electrification planning tools. Khavari et al. (2021) use open-source data to suggest a methodology that translates all high-resolution raster population data into vector-based population clusters. Fobi et al. (2021) demonstrate a data processing strategy to convert settlement locations from satellite imagery to estimated household locations using census-data at the country level. They apply and validate their strategy on Kenya.

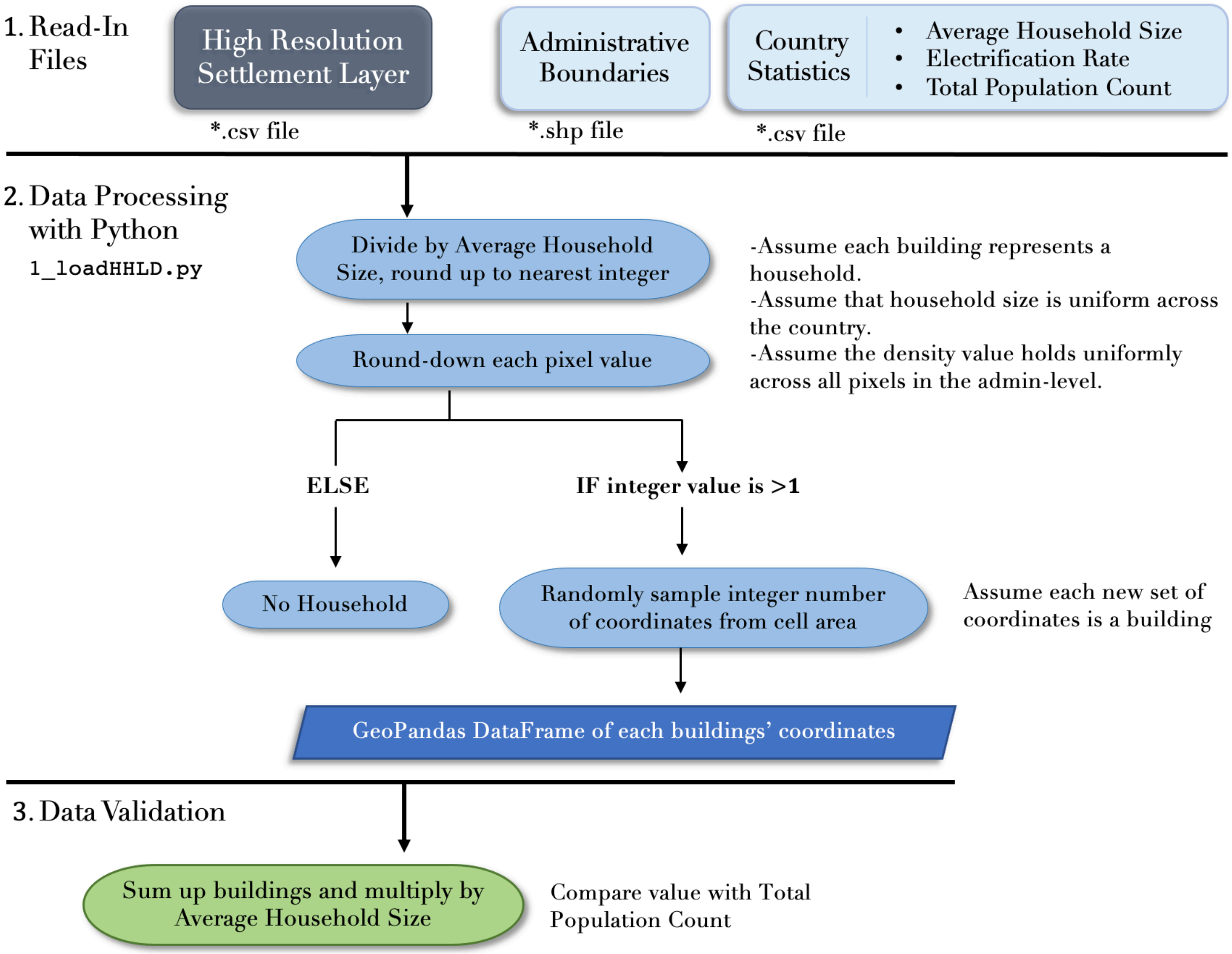

1. Read-In Files
High Resolution Settlement Layer
*.csv file
Administrative Boundaries
*.shp file
Country Statistics
• Average Household Size
• Electrification Rate
• Total Population Count
*.csv file
2. Data Processing with Python
1_loadHHLD.py
Divide by Average Household Size, round up to nearest integer
Round-down each pixel value
-Assume each building represents a household.
-Assume that household size is uniform across the country.
-Assume the density value holds uniformly across all pixels in the admin-level.
ELSE
IF integer value is >1
No Household
Randomly sample integer number of coordinates from cell area
Assume each new set of coordinates is a building
GeoPandas DataFrame of each buildings' coordinates
3. Data Validation
Sum up buildings and multiply by Average Household Size
Compare value with Total Population Count

Figure 4.1.7- Diagram of Building Estimation Method. The first row lists the necessary datasets, with the darkest square signifying the chosen Data Type for the following implementation of the framework. The round-edged modules represent subsequent steps in the method. The final blue parallelogram is the output of the process, and the final step (in green) signifies the validation test for the proposed method.

4.1.4 Applying the Framework: Rwanda Case-Study

Throughout this thesis Rwanda will be used as the case-study for my proposed methods. This is because the Universal Energy Access Lab (UEAL), which houses the REM model, recently completed drafting a long-term universal electrification plan for Rwanda. The project involved the full cooperation of public and private stakeholders in the country. As a result, UEAL, was afforded access to ground-validated REM inputs, which can be applied to test the reliability of the publicly available data sources.

For the task of using HRSL to identify residential buildings I run both a qualitative and quantitative comparison. The qualitative comparison will look how the distribution of buildings in pixels compares to the footprints visible on satellite imagery. It will also compare where the approximate buildings are relative to government-provided settlement locations. The quantitative comparison computes the total number of houses estimated by HRSL and my building distribution algorithm and multiplies it by Rwanda's average household size. The computed residential population estimate is evaluated against the country's total population count.

A Qualitative Comparison

The top-left and top-right images in Figure 4.1.8 are extracted from sparsely populated villages in the North-East region of Rwanda. The bottom-left and bottom-right images, in Figure 4.1.8, are identical and represent a small subset of the Kigali neighborhood in Rwanda. I chose to juxtapose these widely different landscapes to evaluate the model's distribution of buildings in rural and urban locations. The green and pink squares in the top and bottom rows, respectively, are the raster cells of the HRSL dataset. The yellow circles in the squares of the top row images are the HRSL's CSV centroid data. The pink circles in the top row and the blue circles in the bottom row are the outputs of the building estimation method described in Section 4.1.3. Studying the outputs, the model is much more precise in areas with low density and therefore fewer houses to distribute within the pixel. In some instances of the village images, the model outputs exactly overlay the visibly discernable building footprints. However, since a 30-by-30 m (or 900-m^2) land-area is so big, randomly assigning buildings coordinates to anywhere within it means that these houses will sometimes end-up nonsensically on a road or nearby plot of vegetation. However, the implications of this on REM's recommended designs of the distribution network is likely small. For denser regions, where many buildings are sprinkled within the 900-m^2 cell area, this effect is magnified, and buildings rarely seem to match what the underlying satellite image shows. However, given that these areas are also characterized by congruent building footprints, it would be difficult to devise a method that produces more 'sensical' results. The most accurate alternative would be the building footprint extraction models discussed in Section 4.1.1.

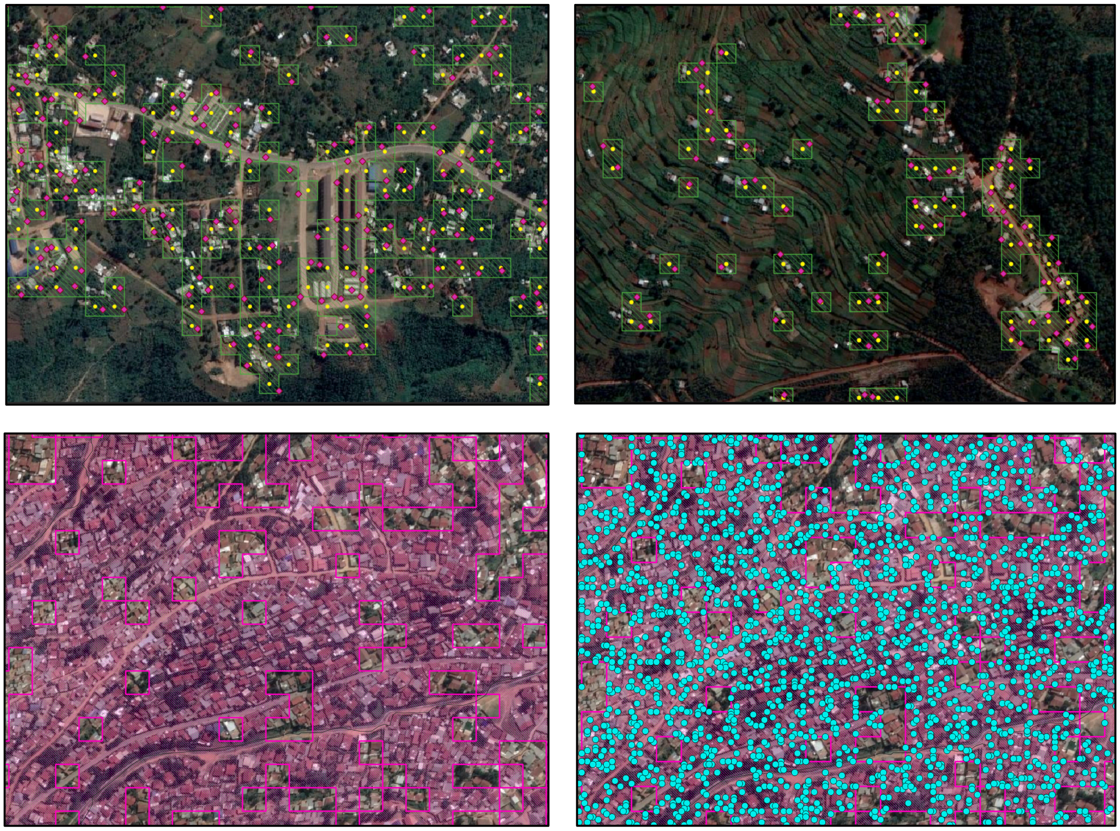


Figure 4.1.8- Building Estimation Method Performance in Multiple Regions of Rwanda. The top-left and top-right images are extracted from sparsely populated villages in the North-East region of Rwanda. The bottom-left and bottom-right images are of the same area of the Kigali neighborhood in Rwanda.

Figure 4.1.9 compares the output of the model (marked by a blue dot) with both the original HRSL raster (pink grid), and the ground-truth data supplied by the Rwanda government (marked by a green-star). Zooming into closely, we see that the ground-truth data is a poor measure of the total number of buildings in an area. In the bottom-left image of Figure 4.1.9, there are multiple instances where we can visibly see more buildings than the green-stars indicate. It is less obvious if the model's output accurately accounts for those missed buildings or slightly overestimates the number of buildings in the area. Overestimation is likely since population densities are adopted from coarse administrative data, and therefore may not be consistently representative of local population distributions.

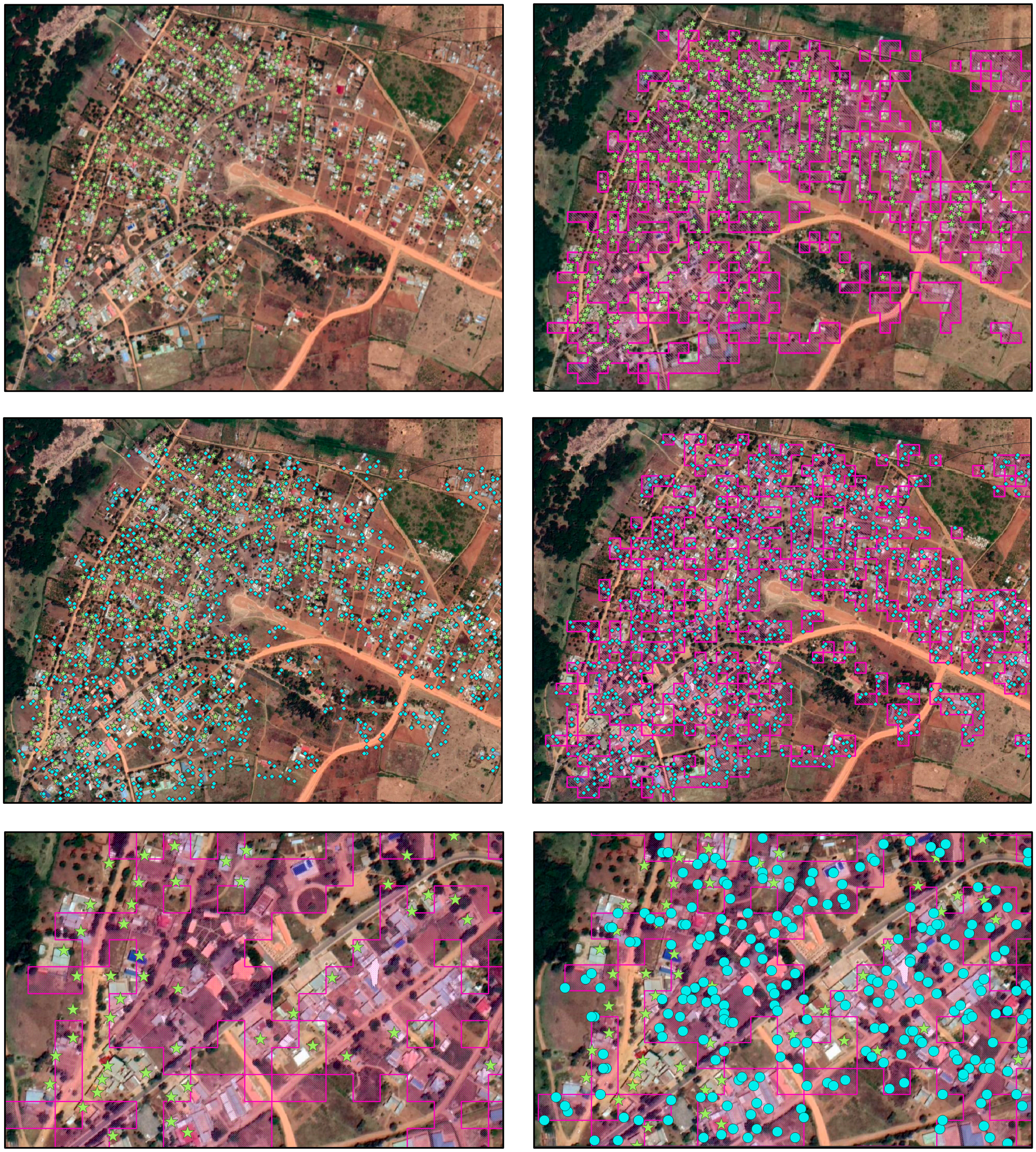


Figure 4.1.9- Building Estimation Method Performance Compared to Ground-Truth Data for Medium-Sized Village in North-East Rwanda. All images correspond to the same area except at different zoom levels. Green stars are the ground-truth data, blue dots are the outputs of the model, and the pink grid is the raster of the HRSL dataset.

A Quantitative Comparison

The model's total estimated population underestimates Rwanda's recorded population by approximately 2 million people. It is unclear if this underperformance is unique to Rwanda or ubiquitous when applying the HRSL dataset. As a result, I suggest future work should run this method on all countries in SSA and plot the estimated population counts by the official counts. This way, researchers can observe how sensitive this model is to how HRSL data differs across countries, for example how the respective granularity at which countries report population counts might affect the applicability of population density values.

Table 4.1.10- Quantitative Comparison of Rwanda Population Counts.

Estimated Number of Households	Estimated Population Count	Total Population Count	Error (%)
3,395,081	14,598,852	12,630,000	15.6

4.2 Identifying Building Locations: Productive Loads

In electrification planning we care about the location of buildings for two reasons: to know where individuals still lack electricity and to provide feasible, cost-effective plans on how to connect the buildings to power. Section 4.1 has given a thorough overview on how to locate people (or settlements). This section will enrich these findings by providing methodology to answer crucial questions about the nature of these settlements: what type of consumers the settlements represent, how much electricity they are going to need, and how planners should anticipate how their load behaviors will change over time. These parameters are especially critical to running REM. Unlike other electrification planning models, REM incorporates engineering design capabilities into its evaluation of a cost optimal electrification plan and its prescription of network layouts for Mini-Grids and Grid Extensions. Therefore, modelling the load characteristics of the identified buildings becomes crucial to the legibility and useability of REM's outputs.

In my implementation of the framework, I use building data to delineate the extent of household settlements and small loads. Examples of small loads are local commercial businesses, village community centers, or household-operated services in rural or peri-urban areas. These loads usually share similar consumption profiles and are always fed by low-voltage lines. Other loads like health centers, education facilities, mines, and large industrial consumers will have unique consumption patterns, require larger and more reliable amounts of electricity, and might need to be fed by higher voltage lines. As such, they become important to REM's generation capacity studies and spatial delineation of Off-Grid (OG) and Grid Extension (GE) clusters. As shown in Figure 3.2.1 in Section 3, the Building Coordinates module is used as an intermediary input to the estimation of the MV Grid Input when the information is unavailable for a user's study region. In which case, feeding REM the most correct spatial distribution of large customers is critical to ensuring it closely approximates the country's existing grid.

While my framework emphasizes distinguishing between productive and residential (or smaller) loads, a user may opt to skip this step if they are less concerned with modelling future capacity requirements and detailed grid layouts. Instead, they can randomly allocate a pre-determined list of customer archetypes to the settlements they detected from Section 4.1. Similarly, if the user has access to the layout of the MV grid when running REM, sidestepping productive load estimation can also suffice. This holds only under the assumption that most large, critical loads are considered high-priority and profitable to the local utilities and thus likely to be already electrified and not relevant to the planning of network expansion to all customers. Still, medium-sized loads in unelectrified regions are deciding parameters in REM's assignment of clusters as either sophisticated Mini-Grids (MG) or small solar-home kits (SHS). The following sections provide an operational background on REM, and a review of open datasets that can help estimate productive loads and introduces an experiment on using the Google Places API to supplement our building coordinates input.

4.2.1 Background on REM: How does it consider individual building locations?

The objective of REM is to determine the optimal grouping of individual customers into electrification clusters such that total system costs are minimized (Ciller et al., 2019). Each cluster is assigned one of the three electrification modes (MG, SHS, or GE). When deciding among the electrification modes, REM evaluates the cost of the internal network of a MG or a GE network that meets all prescribed technical requirements. Technical requirements include physical constraints of power flows, capacity sizing, a catalog of existing grid components, compliance with grid codes, reliability requirements, and dispatch strategies (Ciller et al., 2019). REM employs the Reference Network Model (RNM) to run these technical network-design tasks.

RNM is a software tool for large-scale network design, which can also be applied in the design of the network components of micro- or mini-grids (Ciller et al., 2019; Domingo et al., 2011). RNM designs a quasi-optimal distribution network that meets demand requirements under prescribed reliability and electrical constraints, while minimizing capital investment costs (CAPEX) and corrective, or operational maintenance costs (OPEX). Electrical constraints include grid feasibility metrics like maximum allowed voltage drop and the maximum capacity of the system. To produce this optimal network layout, RNM requires the location of customers, their estimated peak power usage, the location of transmission substation, topological features (e.g. slope, altitudes in the terrain, forbidden or penalized zones and village or other administrative boundaries), and a catalog of technoeconomic information (Ciller et al., 2019). Once RNM produces its optimal layout and the corresponding costs for a particular cluster under a specific electrification mode, REM proceeds with the least cost cluster design. Importantly, the quality of RNM's design and cost estimates are only as good as its inputs. Hence, prioritizing the accurate specification of building locations, and appropriately characterizing each building as a particular load type is necessary for building a scalable, open-access framework that does not majorly compromise the outputs of REM (or similar electrification planning tools). This is especially true for GE clusters where customer types are more heterogenous than small villages composed of a (usually) standard list of load-types. Cities, on the other hand, tend to be more diverse and consist of unique industries and consumer preferences. As a result, I recommend considering productive loads separately, since it would only be to the advantage of the electrification planner.

Customer Archetypes, Demand Profiles, and REM

REM classifies customers into broad customer archetypes like 'Secondary School', 'Health Center', or 'Local Government Office'. The demand of each archetype is modelled as either a single power utilization pattern or as a linear combination of several consumption patterns. The demand patterns are typically estimated by considering the consumption of similar loads in rural regions of other countries, or by using the expected appliances and machinery used to estimate what demand needs will be through the day or year. Compiling these demand profiles is beyond the scope of the thesis, and is the primary subject of multiple studies in literature (Alova et al., 2021; Fabini et al., 2014; Fobi et al., 2021; Lee & S, 2019; Williams et al., 2017). Nonetheless, the consensus across this literature is that optimal electrification planning without modelling the spatial heterogeneity of demand will lead to misleading, often more expensive, results at the local level. For example, Lee & Eduardo (2019) demonstrate that modeling demand heterogeneity produces REM plans that are 9% less costly than modeling assuming one single customer type and increases the prevalence of MGs and SHSs. Consequently, developing a method for locating different consumer types and thus inputting spatial variations in demand, enables planners to accurately model the impact of settlement patterns on the components of the distribution system, and avoid under- or over-building the local system.

While I am not concerned with estimating energy demand, my choice to formulate a method to distinguish households from productive loads, and segment load types to buildings does support the effort of better demand characterization. I refer to REM's broad customer archetypes to define the scope of non-residential loads this section should explore. The list includes:

- Community Consumers: *Primary and Secondary Schools, Children Care Facilities; Technical Institutes and Universities; Health Centers; Government Offices; Large Community Spaces (e.g., Markets)*
- Productive Consumers (aimed at economic rewards): *Irrigation Sites (Agricultural Industry); Mining Facilities; Manufacturing or Distribution Factories; Large Retail Sites; Telecommunication Towers; Airports; Tourist Attractions (Museums, Amusement Parks...etc.)*

4.2.2 Open Data Sources and Limitations

Open Databases and Manually Compiled Datasets

Several existing datasets detail the location and characteristics of the customer archetypes listed above. These datasets are typically compiled manually, by the provision of governmental institutions, developmental agencies, or non-governmental organizations (e.g., the WHO and the World Bank), each of whom are stakeholders in the developmental progress of nations in SSA, encouraged to create and share these data types to monitor and expedite the improvement of infrastructure, sanitation, and health across the continent. Unfortunately, these datasets are limited to few load types, and vary in availability between countries. Moreover, given that these datasets are collected mostly by survey and engagement with on-the-ground resources, it takes exceptional effort to remain aware of changes on the ground, then recompile and update the data; therefore, they may quickly become outdated. Table 4.2.1 provides an overview of these publicly available datasets. The table also includes a list of databases which regularly aggregate and share new data on African (or global) nations, for academic and developmental purposes (e.g., Open Africa). Given that our objective is to find a scalable, publicly available method for estimating inputs of electrification planning tools, a more robust and consistent approach to identifying the location of Community and Productive customers is preferred.

Table 4.2.1- List of Publicly Available Datasets on Productive Customer Types.

Category	Dataset Name, Institution	Description	Relevant Publications
Healthcare Facilities	Public Health Facilities in Sub-Saharan Africa, World Health Organization (WHO)	Geospatial inventory of 98,745 public health facilities, compiled from a variety of government and non-government sources from 50 countries and islands in SSA.	(Maina et al., 2019)
	Healthsites.io, International Committee of the Red-Cross	Global geospatial inventory representing more than 150,000 health facilities. OpenStreetMap was the primary dataset used, as well as several databases from trusted partners.	(René Saameli et al., 2016)
Large-Scale Databases for SSA and other regions	OpenStreetMap	Large crowdsourcing platform that contains data about various infrastructure types, including public and private institutions. The information is plenty but inconsistent across the continent.	None.
	OpenAfrica, Code for Africa	Large independent repository of open data on the African continent.	
	EnergyData.Info, the World Bank Group (WBG)	Open data platform providing access to datasets and analytics that are relevant to the energy sector.	
	Open Data Impact Map, Center for Open Data Enterprise	Public database of organizations that use open data from around the world.	
Cropland Extent	Global Cropland Extent Project, NASA	Utilized 250m MODIS satellite data to map global production cropland extent, compiled over the period of 2000-2008.	(Pittman et al., 2010)
	Global Food Security Analysis-Support Data at 30m for the African Continent, United States Geological Survey's (USGS)	Provides cropland extent maps at 30-m resolution by utilizing multiple satellite imagery composites and Machine Learning (for pixel-level classification of cropland).	(Xiong et al., 2017)
Mining Sites	Geospatial map of company-owned and community-managed mines across SSA$^{\delta}$, University of Sheffield	Identified and mapped the extent of 469 mines across SSA. Data is compiled form a variety of sources which include USGS, Google Earth, and other databases (e.g., Mining-atlas.com)	(Ahmed et al., 2021)

δ No link available. Data available upon request only.

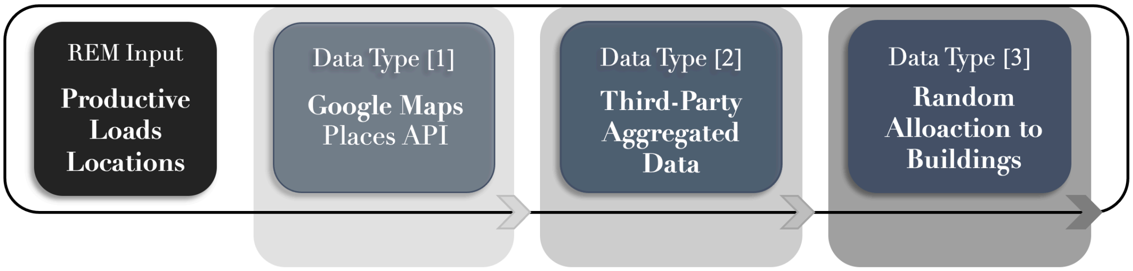
Google Places API for Locating Productive Load Types

A yet unexplored source of data on the geolocation of different businesses or services (i.e., electricity consumers) in an area, is Google Maps. Google Maps covers over 200 countries and territories, is updated 50 million times a day, and has 1 billion monthly active users. In other words, it is an expansive, global resource for geospatial data. To enable program developers to expand and apply the Google Maps library, the company created the Google Maps API with three products: Maps, Routes, and Places. The Google Places API lets one search for places in the same way one would type in 'coffee-shop' or 'market' into Google Maps and get a list of nearby businesses meeting those keywords. With the Places API Nearby Search, a user specifies a set of search parameters as an HTTP request (i.e., a https:// URL), and the API returns the results of the search in JSON or XML notation (user specified in the 'output' field). The required search parameters are: (1) one or more descriptive keywords of the place, (2) the coordinates of their region of interest, (3) the desired search radius (in meters), and (4) an API key. The API response (results) will consist of a list of places that match the terms of the search and supporting information on each place. Of the information returned, the most useful fields for our application are the following:

- **Geometry:** Latitude and longitude coordinates of the location.
- **Name:** Human readable name of the business or establishment.
- **Place ID:** A unique textual identifier assigned to all places on the Maps API.
- **Types:** Array of keywords that describe the Place. The keywords are restricted to a pre-defined list of 100 place-types established by the Google Maps API.
- **Vicinity:** A simplified address for the place that usually contains the administrative locality it belongs to.

The specificity and ubiquity of Google Maps makes it incredibly powerful at extracting information on a variety of customer types and their locations. Information from this service is likely the most complete and reliable than any other compiled dataset or open database today. However, users of this data source should begin by reviewing the Google Maps Platform and Terms of Service, and specifically reference Section 3.2.3(a) and (b) on the storage and sharing of API results. So far, no published literature has employed the Google Places API for infrastructure detection. Unlike publicly available datasets discussed earlier in the chapter, using the API comes at a cost. To use their products, Google requires users to register for an API key[7]. Each time the registered API key is used to request data from the platform (e.g., we execute a URL search for nearby places) the user is billed. The cost is calculated by multiplying the price per use of the product with the number of uses in a month. Each Google Maps API product has a registered SKU. For this study, the product used is Places Nearby Search and it is billed at 0.0112 USD per use. This translates to 112 USD for 1,000 uses of the product, or 1,000 individual URL searches; and users are given 200 USD free Google Maps credit per month. While not inexpensive, the costs are expected to be much less than the time and resources needed to acquire the information by alternative means, like time-intensive relationships with on-the-ground stakeholders or surveyors.

4.2.3 Applying the Framework: Rank Order Data Types


REM Input
Productive Loads Locations
Data Type [1]
Google Maps Places API
Data Type [2]
Third-Party Aggregated Data
Data Type [3]
Random Alloaction to Buildings

In the direction of **decreasing** data type quality & validation accuracy.

Figure 4.2.2- Rank Ordered List of Community and Productive Load Data Types. Ranking of validation accuracy is presumed and not quantitatively measured. Google Maps API is more complete and thus ranked ahead of the aggregated data sources.

When ranking the Data Types, I prioritized data that could be used in any region, at any level of specificity, and for any type of load. The only Data Type to meet all the requirements is the Google Places API. Data Type [2], which encompasses all the datasets

[7] Several online resources are available on how to set-up a Google Cloud account and obtain a unique API Key.

described in Table 4.2.1, is ranked second because it does not include information for all types of loads or for all regions. Data Type [3] involves the random assignment of buildings in REM's clusters to pre-defined load types. Therefore, it is theoretically available for any country and for all loads. However, it lacks any relation to what's happening in the country in real-time, therefore it necessarily is less accurate than using [2] wherever it is applicable and [3] overall. As such, I still recommend [1], or [2] if possible, especially if the user of the framework intends to also apply the information to replicating a country's existing MV grid.

4.2.4 Applying the Framework: Google Maps API for Load Type Identification

My method can be separated into three steps: (1) Specifying the latitude and longitude coordinates for the full search area, (2) Selecting an optimal search radius, (3) Compiling the list of keywords to search for each load type with.

Specifying Latitude and Longitude Coordinates for Study-Region

The Places Nearby Search tool works by looking for places within a radial area that match the supplied keywords. The centroid of the area and its extent is determined by the user-specified spatial coordinates and search radius, respectively. To simplify the task of selecting coordinates that encompass the entirety of the study-region, I take advantage of the HRSL CSV dataset. In this way, I have the complete list of centroids of all 900 m^2 squares that demarcate my country of interest. For smaller, intra-country, regions a user can simply apply a vector mask of the region's extent on the HRSL data using any GIS software, for example QGIS or ArcGIS, and extract only the centroids (and squares) of interest.

Selecting an Optimal Search Radius

Choosing a search radius is a non-obvious task. The computational time for running multiple requests on smaller search radii is much larger than the computational time needed to run fewer requests on larger radii. At the same time, the API returns a maximum of 60 places in a single request, therefore if the radius is too large then places that meet the search criteria but exceed the first 60 responses will be left out of the API results. Since the centroids in the HRSL CSV files are extracted from an original raster file, we know that each coordinate is the center of equally sized, sequential squares (or pixels). We also know that the dimensions of the square are equal to the spatial resolution of the raster file (30m in our case). Figure 4.2.3(a) demonstrates the results of the setting the diameter of the search area equal to the length of the square (or pixel). The regions in red are sections of the land-area that are excluded from the search, and therefore places in those sections will be unaccounted for. Figure 4.2.3(b) circumscribes the square in a circle, which we know from geometry will have a diameter equal $\sqrt{2} \times \text{length of square}$ (or 42.4m in our case). This approach will ensure no areas are missed but might lead to duplicate responses from adjacent pixels. Fortunately, removing duplicates from an array post-compilation of all API responses is very quick and easy, since each place has a unique textual identifier: Place ID.

To improve the computational efficiency of the task I leverage the sparsity of several load types. For example, there will never be more than a few airports or mining sites in a country. Health centers or hospitals are also typically distributed uniformly across a country and therefore checking each pixel for these load types would be redundant and costly. As a solution I use a larger search radius and only call coordinates of every other, every 10, or every 100 pixels. I skip more or fewer pixels depending on how infrequently or frequently I expect to see the load type, respectively. The size of search radius should be scaled accordingly. For example, one should increase the size of the search radius when skipping more pixels. to guarantee all relevant loads are encapsulated. Again, it costs less to search a larger area than it does to make multiple, redundant API requests – both in money and computational time. Figure 4.2.3(c) demonstrates an example where every other pixel is skipped, and the search diameter is twice the size of the cell.

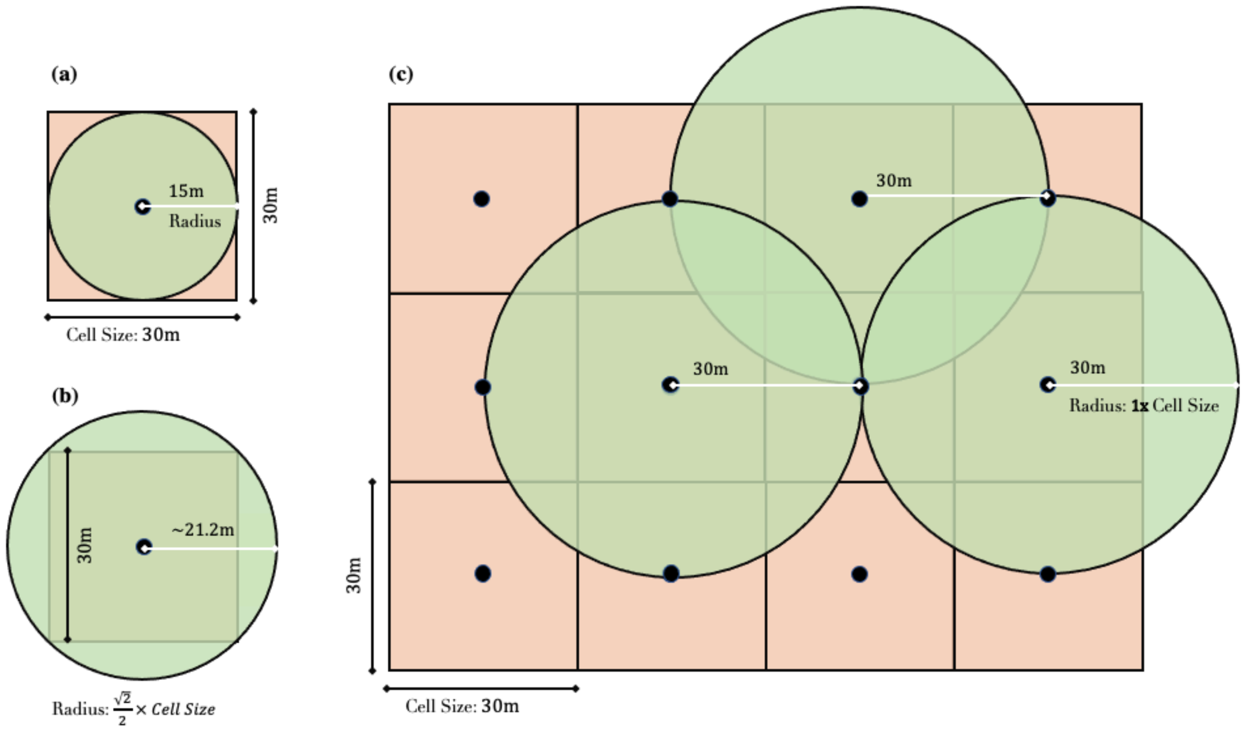


Figure 4.2.3- Visualizing How Different Search Radii Impact the Areas Encapsulated. Red shading indicates land-areas left out of the API search, while green indicates areas included.

Compiling the List of Place Types to Search

Google Maps classifies all its Places as 1 of 100 pre-defined 'types'. Each 'type' is a one-to-two-word identifier of a Place. In Section 4.2.2 I introduce the customer archetypes I want to locate. I expand and explicate this list into 18 total archetypes in Figure 4.2.4. Instead of searching for Places which belong to those archetypes using my own descriptive keywords, I decide to leverage Google's own classification nomenclature, the pre-defined 'types', as the keyword in the URL API request. The advantage of using Google's own terminology is that it eliminates the guesswork on how loads-of-interest were categorized internally.

I started by exporting the list of 'types', provided by Google Maps, into a CSV file. I then manually discarded 'types' which were: too specific and irrelevant (e.g., 'bookstore' or 'locksmith'), not electrified (e.g., parking lots), or not culturally consistent with the SSA region (e.g., RV Park). I classified each 'type' into one of the 18 general consumer archetypes I am interested in. And will use each of these keywords when compiling the geolocation of all loads that fit the archetype in my region of interest. Note that when implementing this method, users are free to reduce the consumer archetypes or use fewer (e.g., only the most specific) 'types' in their search. Table 4.2.4 shows how I have distributed each of these keywords into my broader classes of customers.

Table 4.2.4 Linking Consumer Load Types to Google Maps Keywords

Load Type	Keywords	Description	Load Type	Keywords	Description
Mechanically -Operated Business	Car_repair; Casino; Electrician; Fire_station; Gas_station; Gym; Laundry; Movie_Theater; Theater;	Any business that relys heavily on some form of machinery.	Retail & Services Businesses	Beauty_salon; Car_dealer; Clothing_store; Convenience_store; Department_ store; Electronics_store; Funeral_home; Furniture_store; Hair_ care; Hardware_store; Home_goods_store; Jewlery_store; Spa; Store;	Any stores or service-providing establishments (e.g. salon).
Health Center	Clinic; Doctor; Health; Hoispital; Dentist; Veterinary_care;	Any establishment providing medical care.	Business Office	Accounting; ATM; Bank; Car_rental; Insurance_agency; Police; Post_office; Travel_agency; Cell_office;	Any office operational business.
Restaurant	Bakery; Bar; Café; Night_club; Restaurant;	Any food/ beverage or server- based business	Specialty Park or Tourist Destination	Amusement_park; Aquarium; Museum; Stadium; Zoo;	Any high-powered recreational place.
Transit Station	Bus_station; Subway_station; Train_station; Transit_station; Taxi_stand;	Transportation Hubs.	Large or Small Market	Drugstore; Liquor_store; Pharmacy; Shopping_mall; Supermarket;	Food/ beverage or medicinal retail.
Facotry or Similar	Airport; Factory;	HV 24/7 powered loads.	Hotel	Hotel; Lodging;	Hotels/Motels/ Lodging.
Places of Worship	Church; Temple; Mosque; Synagogue;	Religious establishments.	Community/ Government Halls	City_Hall; Courthouse; Local_government_office; Embassy;	Government offices.
Mining Company	Mines; Mining_facility;	Mining locations.	Primary School	Primary_school;	Elementary schools.
Nursery School	Nursery_school; Child_care;	Young children schools/ care.	Secondary School	Secondary_school;	Middle and high schools.
University & Institutes	Library; University; Technical_institute; Training_facilitie;	Resembling a technical school.	Telecomunication Tower	Telecome; Cell_tower;	Cellular and network towers.

Final Implementation

To compile a DataFrame of the location of all Places that encompass a single archetype, I search a radial area for each of its associated keyword one term at a time, loop through all cells and all keywords, then remove Place ID duplicates. As mentioned earlier, each Load Types will have an optimal search radius, whereby the protocol for setting it is: narrower radii for more frequent Places, and wider radii for less frequent Places. Users are expected to repeat the process laid out in Figure 4.2.5 for each of the customer archetypes.

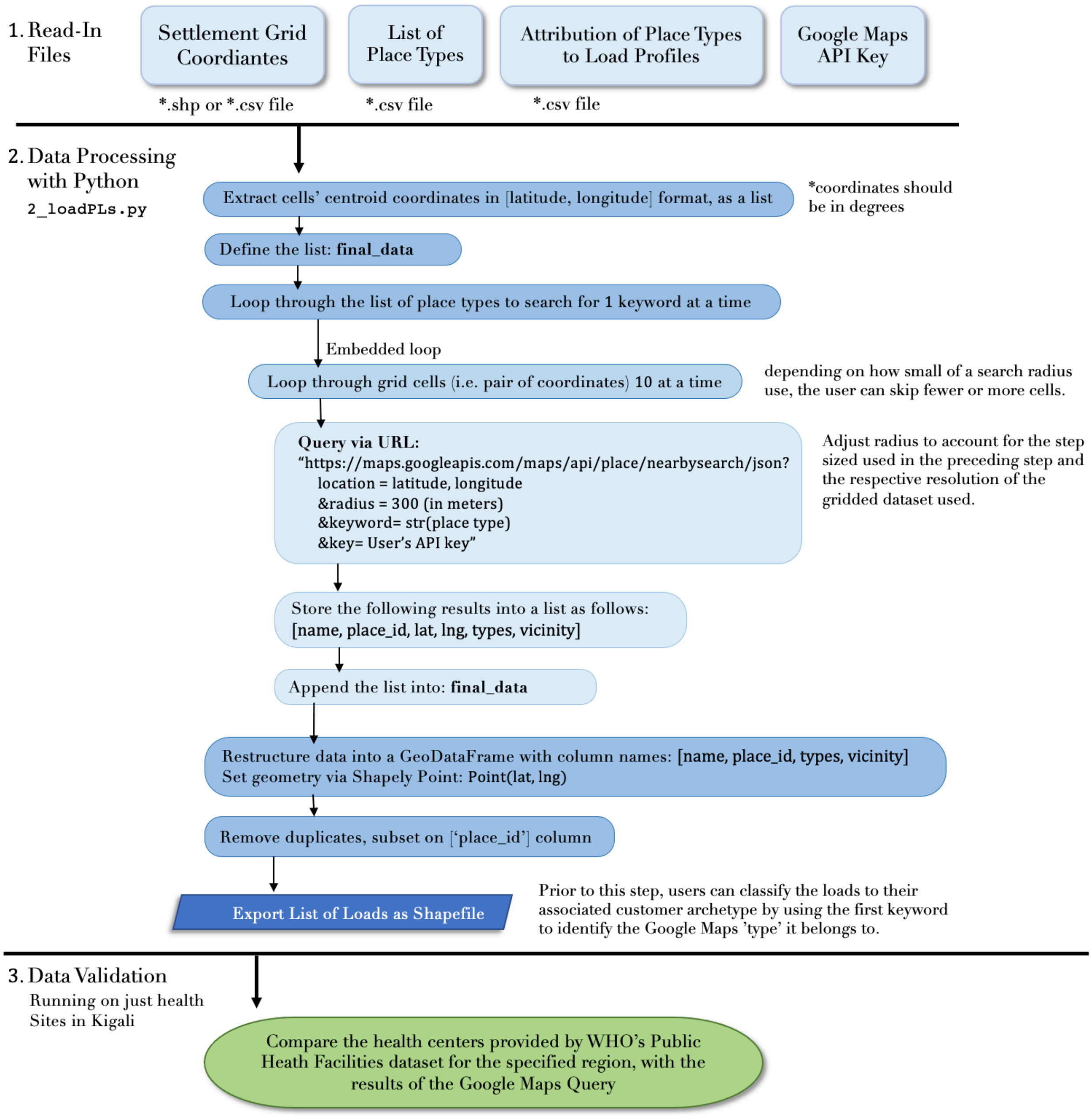

1. Read-In Files
Settlement Grid Coordiantes
*.shp or *.csv file
List of Place Types
*.csv file
Attribution of Place Types to Load Profiles
*.csv file
Google Maps API Key
2. Data Processing with Python
2_loadPLs.py
Extract cells' centroid coordinates in [latitude, longitude] format, as a list
*coordinates should be in degrees
Define the list: final_data
Loop through the list of place types to search for 1 keyword at a time
Embedded loop
Loop through grid cells (i.e. pair of coordinates) 10 at a time
depending on how small of a search radius use, the user can skip fewer or more cells.
Query via URL:
"https://maps.googleapis.com/maps/api/place/nearbysearch/json?
location = latitude, longitude
&radius = 300 (in meters)
&keyword= str(place type)
&key= User's API key"
Adjust radius to account for the step sized used in the preceding step and the respective resolution of the gridded dataset used.
Store the following results into a list as follows: [name, place_id, lat, lng, types, vicinity]
Append the list into: final_data
Restructure data into a GeoDataFrame with column names: [name, place_id, types, vicinity]
Set geometry via Shapely Point: Point(lat, lng)
Remove duplicates, subset on ['place_id'] column
Export List of Loads as Shapefile
Prior to this step, users can classify the loads to their associated customer archetype by using the first keyword to identify the Google Maps 'type' it belongs to.
3. Data Validation
Running on just health Sites in Kigali
Compare the health centers provided by WHO's Public Heath Facilities dataset for the specified region, with the results of the Google Maps Query

Figure 4.2.5- Stepwise Process of Estimating Productive Load Locations. The first row lists the necessary datasets. The round-edged modules represent subsequent steps in the method. Indentations and shading signify embedded and groupings of code, respectively. The final blue parallelogram is the output of the process, and the final step (in green) signifies the validation test for the proposed method. The process was adapted from: https://towardsdatascience.com/tagged/google-places-api?p=588986c63db3

4.2.5 Experiment: Extracting Health Sites in Rwanda from Google Maps API

To see how accurate using Google Places is, I conducted a small experiment comparing the WHO Public Health Facilities dataset with a Google Places API search for health centers. The keywords used to run the API search were: ['*health*', '*hospital*', '*clinic*', '*doctor*', '*medical*', '*health center*']. The search was limited to a small region of Kigali, Rwanda. Upon inspecting the results a few of the matches were unrelated (e.g., a grocery store that sold health products). This is easily remedied by filtering results where the first element in the returned "types" array does not match any of the keywords. I then plotted the results (see Figure 4.2.6) and found that some facilities appeared twice under slightly different spellings. We see this with the Cor Unum Health Center on the map in Figure 4.2.6. Resolving this issue is slightly more difficult since there's no automatic way to extract repeated Places with different field characteristics (i.e., names). However, if instances of repetition are infrequent, then it is unlikely that it will drastically impact the results of REM as an electrification planning tool, or to replicate the MV distribution network. To verify if these repetitions are infrequent, future studies could search different areas for different load types, then plot and inspect the results.

Altogether, the experiment shows promising results by accurately predicting the location, and name of all WHO health facilities. In addition, the API also returns all other health clinics and private practices in the area. This level of load-type-information granularity is uncommon for publicly available data sources and should be taken advantage of. Nonetheless, the proposed model must undergo further testing to discern how practical it is for large-scale use.

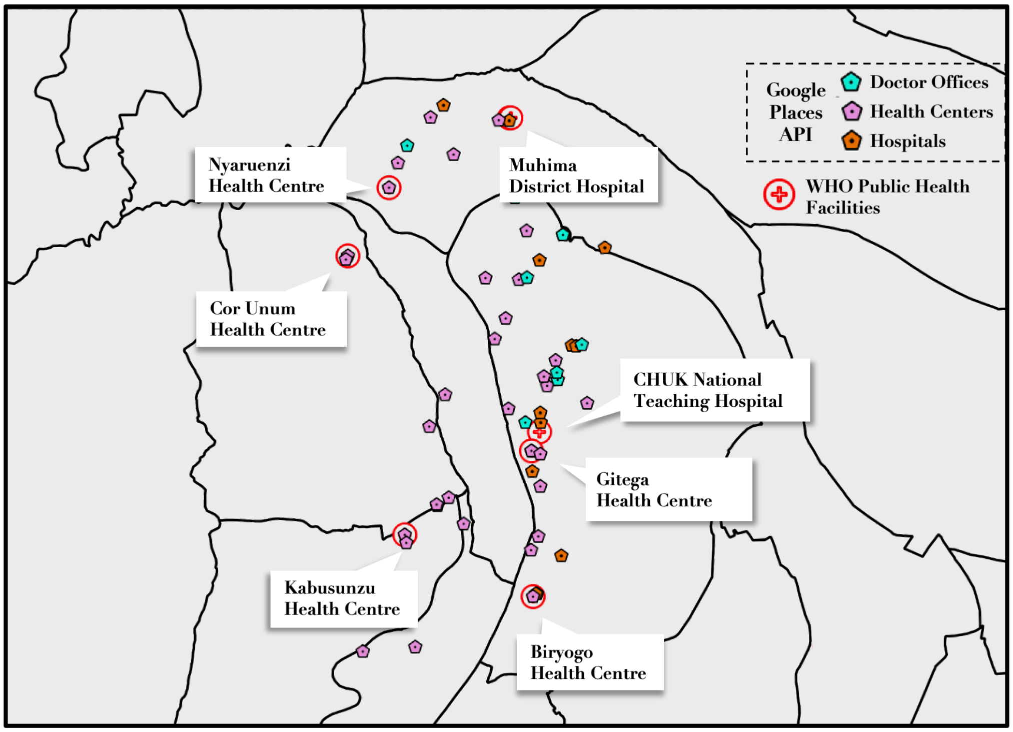


Figure 4.2.6- Health Centers identified by WHO and API within Kigali, Rwanda. Key in upper-right corner explains what each marker on the map is. Black lines represent Administrative Level-3 Boundaries for Rwanda in the Kigali region. The white tags call out all instances in which the WHO and API data aligned**.**

4.3 Delineating Electrification Status

Once the locations of productive and residential loads have been determined, the loads must be classified as 'already electrified' or 'without electricity'. The location and characteristics of electrified loads will be used as an input to REM during the estimation of the existing MV grid layout. Unelectrified load information is used for the actual task of planning how to achieve universal energy access under prescribed techno-economic and reliability requirements. The size and distribution of all unelectrified customers in a study region informs REM how to optimally cluster and delineate loads between the off-grid technologies (MGs or SHS) and grid extension. Furthermore, accurately tracking the state of electricity access is crucial to assessing progress and prioritizing investments towards universal electrification. Despite this, there are only a few recorded approaches for electrification status estimation. The following section will provide an overview of these estimation methods, describe their relative advantages and disadvantages, and propose a scalable, open-access approach to delineating the electricity status of buildings.

4.3.1 Data Sources & Limitations

Nighttime Light Composites

The single most-employed proxy for estimating electrification from openly available sources is nighttime light data (NTL). NTL information provide global daily measurements of nocturnal visible and near infrared (NIR) light; or in other words describes the distribution and relative brightness of all synthetically lit-up areas. It takes a full year of nightly observations to produce a global composite of nighttime lights, free of extraneous features unrelated to electric lighting (Elvidge et al., 2021). The annual composites are filtered against cloudy, sunlit, or moonlit data and exclude outlier pixels representing biomass burning, natural gas flaring, snow, or the aurora. The data is either available through the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership satellite, at a resolution of 450-m, for the year 2012 onwards, or its predecessor the Defense Meteorological Satellite Program - Operational Linescan System (DMSP-OLS), at a spatial resolution of 1-km, for the years 1992-2013.

Previous studies have shown that NTL intensities can be used to assess residential energy consumption, detect power outages, monitor population fluctuations and migration, as well as regional GDP and income inequality (Albert et al., 2017; Elvidge et al., 2020; Falchetta et al., 2019). The strengths of using NTL data for electricity status estimation is its global coverage, frequent collection (daily), consistency of measurement across administrative borders, and long historical record (the VIIRS satellite has reported data since 2012, and its predecessor DMSP-OLS satellite has reported data since 1992-2013) (Correa et al., 2021). However, the data exhibit substantial noise, tend to overrepresent streetlights, suffer from stray light effects, can be difficult to distinguish from lunar illuminance, and are powerless in the face of clouds. Moreover, deep rural areas, even when electricity access is present, may not register any signature because of extremely low levels of external lighting.

Despite these shortcomings, several studies have worked to improve NTL's prediction of electrification in developing areas. Correa (2021) showed that daily NTL composites combined with open-source population counts improves the detection of electrified regions and the generation of more accurate electricity access maps. Elvidge (2020) focuses on low-income regions, typically the subject of most electrification planning tasks, and provides evidence that indices such as the mean, variance, and lift of NTL irradiance can be used to assess long-term power supply growth and stability. Li (2020) harmonize inter-calibrated observations from the DMSP-OLS data with VIIRS data to produce an NTL time-series dataset. The authors believed that studying the temporal trends of nighttime lights could support various studies related to electricity consumption and urban extent dynamics. Finally, researchers within the UEAL have experimented with applying probabilistic graphical models (PGMs) to delineate the electrification status of buildings (Lee, 2018). The PGMs combined multimodal and multiscale features (e.g., census statistics, or satellite image features) with NTL intensity levels. Altogether, the body of research suggests that NTL data can be a viable method for electrification status estimation.

Falchetta (2019) High-Resolution Gridded Dataset for Electrification in SSA

Falchetta (2019) presents and validates a 1-km^2 resolution gridded electricity access dataset that covers all Sub-Saharan Africa. As it stands, the Falchetta dataset is the most complete publicly available map of electricity access in the region; and therefore, the Data Type I apply to the framework in subsequent sections. The dataset is in netCDF format and characterizes energy access in one of two formats. The first gives the local count of people without access to electricity per grid cell (i.e., pixel), for the years between 2014 and 2018. The second netCDF classifies pixels where people have access to electricity in 2018 as 1 of 4 'tiers of consumption', whereby each tier ranks residential electricity consumers by their expected daily energy usage. The continuous consumption values clustered into these tiers follow the thresholds defined by the World Bank Multi-

Tier Framework[8]; where 4 signifies the highest energy usage, and 1 the least (Falchetta et al., 2019).

Given the importance of the Falchetta data to my estimation of the 'Delineation of Electrification Status' Input Module, I will provide a detailed run-through of how they compile their gridded dataset. To classify each 1-km^2 grid as electrified or unelectrified, for each year, the authors use the 2014-2018 VIIRS-DNB (*Visible Infrared Imaging Radiometer Suite, Day-Night Band)* stray-light corrected NTL monthly composites. The median value of radiance within each pixel of the VIIRS composites was calculated per grid cell. If the average radiance was above a minimum threshold value ($0.25/0.35\ \mu W \cdot cm^{-2} \cdot sr^{-1}$) then the grid was considered 'lit' and the people within it 'electrified'. The inverse applied to grids with average radiance equal to- or below- the threshold value. Population counts for each grid, and electricity status, was imported from the 1-km LandScan 2014-2017 Gridded Population Distribution dataset. Falchetta (2019) gives a snapshot of the estimated density of people without electricity access over Uganda in 2018 in Figure 4.3.1.

To classify each cell to a consumption tier, the authors first identified them as containing either rural or urban settlements. Both the 2017 MCD12Q1 V6 MODIS Land Cover Type dataset (500-m resolution) and 2017 LandScan Gridded Population dataset (1-km resolution) were used to distinguish between settlement types. Cells with land-cover type 13 (indicating the existence of urban and built-up lands) and with a LandScan population density higher than a specific ($inhab \cdot km^{-2}$) threshold are classified as 'urban'. If below these thresholds, the cell is identified as 'rural'. Based on the distribution of quartile values of non-zero light radiance across electrified cells, four tiers of residential consumption were defined, with thresholds set at the median value of each quartile distribution. Threshold values were set separately for urban and rural cells to account for the strong discontinuity between the average radiances of each. For exact radiance values, see Table 4.3.2. Finally, the tier data is mapped to the gridded

[8] World Bank Multi-Tier Framework (MTF): Formalized method to define and measure access to energy.

population counts to derive the share of the population with access to electricity in each consumption tiers. The authors retrieve household survey data from the WorldBank Microdata Library to validate the consumption tiers and national electricity rates from the ESMAP/World Bank Tracking SDG portal to validate their counts of unelectrified population. Overall, their model performs well on most-recent electrification data (2018) with an $R^2 = 0.81$ and the main discrepancies in consumption tier data are attributed to poor recording of residences connected to decentralized technologies.

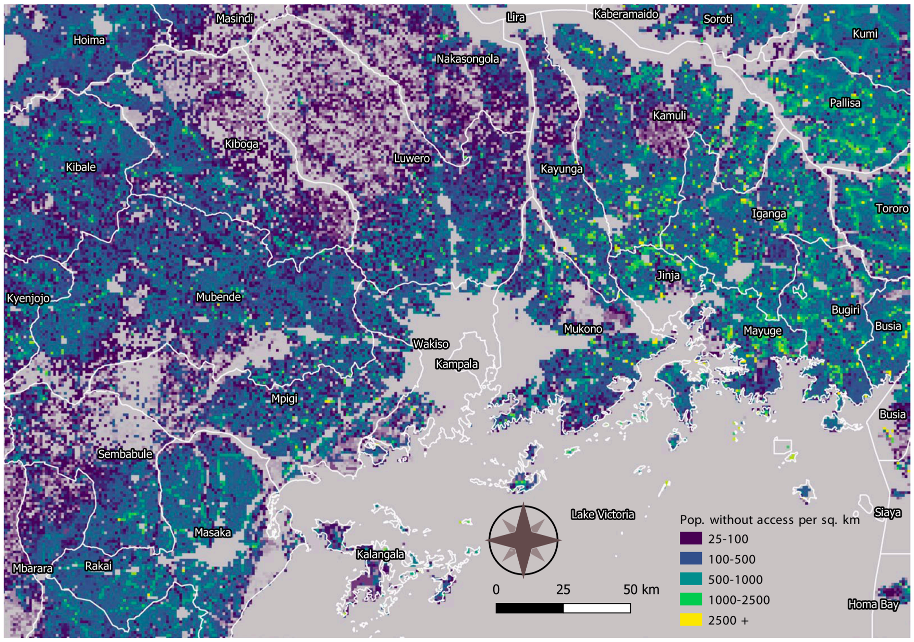


Figure 4.3.1- (from Falchetta et al., 2019) Distribution of People Without Access over Uganda in 2018. Colors represent the density of people without electricity access in each 1-km^2 cell. Administrative boundaries correspond to GADM Level-1.

Table 4.3.2- (from Falchetta et al., 2019) Definition of Radiance Tiers ($\mu W \cdot cm^{-2} \cdot sr^{-1}$) for Electrified Areas Used to Estimate Consumption Levels.

	Urban		**Rural**	
Tier	**Lower-bound**	**Upper-bound**	**Lower-bound**	**Upper-bound**
1	>0	<0.40	>0	<0.38
2	≥0.40	<0.48	>0.38	<0.45
3	≥0.48	<0.88	>0.45	<0.68
4	≥0.88	—	>0.68	—

Machine Learning (ML) Powered Nighttime Lights Data

A potential augmentation to NTL composites or the Falchetta dataset is to incorporate additional publicly available features that might signal the use of electricity in an area. With new advancements in computer vision, some work has been done on applying new machine learning techniques to detect economic activity and poverty from Nighttime Lights (Jean et al., 2016; Xie et al., 2015). Jean (2016) demonstrates an inexpensive and scalable method to estimate consumption expenditure and asset wealth from high-resolution satellite imagery. The study assumes NTL intensities are a reasonable proxy for economic productivity and apply a three-step transfer learning model. First, they fine-tune a pre-trained CNN to predict the nighttime light intensity corresponding to a daytime satellite image. Second, the model learns a non-linear mapping of how features of satellite images can predict the variation in nighttime lights. Finally, using the features extracted and survey data, the authors train a ridge-regression model to estimates cluster-level expenditure and assets. Whereby the subset of features that explain variation in nightlights are also predictive of economic outcomes. Jean (2016) was able to explain up to 75% of the variations in local economic outcomes, emphasizing the heterogeneity captured in NTL data.

Inspired by this work I experimented with using transfer learning to predict local electrification access. As mentioned earlier, NTLs are poor indicators of (or proxy for) the availability of electricity in more remote, rural areas. This is due to customers in those areas consuming less power with fewer streetlights in use, therefore radiance from those regions is more difficult to detect. I employ Jean (2016)'s assumption that high-level features leaned from daytime images are indicative of nighttime lights and therefore economic development (or vice versa), minimizing the crucialness of NTL detectability. Instead, the newly inferred NTL intensities are used as better signals of electricity access and Falchetta (2019) radiance cut-offs can be utilized to re-classify areas as electrified or not.

I train and fine-tune a convolutional neural net (specifically, a ResNet50) to classify a 1-km by 1-km daytime satellite images as one with high, medium, or low nightlight intensity.

Broad classes were used in lieu of precise prediction of radiance values because initial experiments showed that the ResNet50 model performed worse on regression tasks (lower overall validation accuracy). Daytime images classified as *low intensity* corresponded to areas below the electrified-radiance threshold defined by (Falchetta et. al, 2019), *medium intensity* images corresponded to electrified areas with rural or peri-urban settlements (lower tiered consumption), and *high intensity* images represented densely populated, electrified urban regions (consumption tier 4). I trained and tested the model on a subset of north-west India, which was characterized by widely different terrains and population densities. The advantage of this region is its inclusion of Delhi, a city known to be electrified. This is important because my experiment lacked any labels on whether the daytime images were electrified or not, I only used NTL radiance thresholds to make that assertion. Therefore, it is crucial to make this assumption on an area I have enough contextual information on to be able to verify whether it is or is not electrified. In this case, daytime images from Delhi are appropriately labelled as 'electrified' by means of high NTL intensity levels.

Daytime images were obtained from the Google Maps API at a Zoom Level of 16 (spatial resolution of 2.5-m) and with size 400 by 400 pixels, which corresponds to 1 NTL pixel. To extract the NTL data for India only, the image was clipped by overlaying India's administrative boundaries provided by GADM in QGIS. The model performed well with a validation accuracy of 88.9%. I inspected the final activation layer of the ResNet50 model to guarantee that the neural network was learning semantically meaningful features in the images when classifying it as one of the three NTL intensity classes. Figure 4.3.3 plots these activations on an example of a *high intensity* daytime image. Activations refer to the features of an image the CNN uses to classify it as one of the three classes. Figure 4.3.3 demonstrates that the model successfully activates only regions with buildings, ignoring surrounding vegetation. In fact, it does this so well that it precisely activates the footprint of the building in the middle of this green patch. Suggesting the model is using the existence and extent of detected built infrastructure in a daytime image, to classify high-intensity pixel areas (i.e., electrified areas).

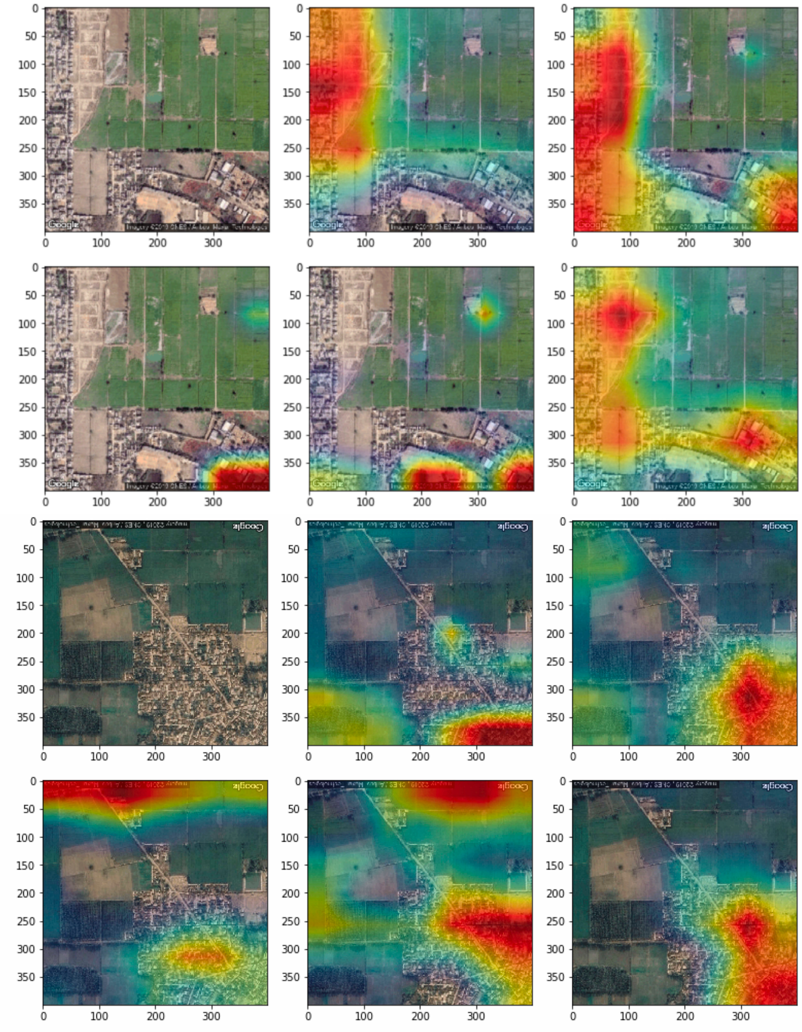
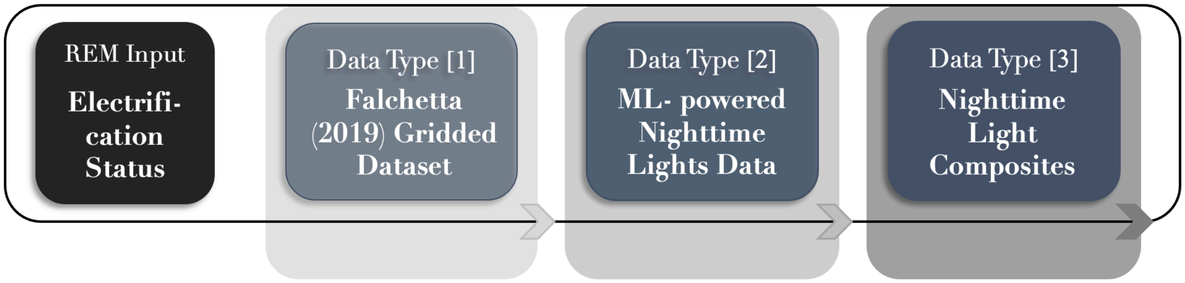


Figure 4.3.3- Activation Features of Correctly Classified High Intensity Daytime image, in North-West India.

The experiment confirms that CNN models can successfully and inexpensively learn useful infrastructure information from publicly available satellite data, then abstract this information to predict broad NTL intensity classes. However, its outputs are not verified against any ground-truth survey data and therefore I cannot verify that the inferences made on intensity classes corresponds accurately to electricity statuses. Therefore, additional validation of the model's outputs is needed, as well as the incorporation of more diverse terrains to test the robustness of the model. Nonetheless, enhancing NTL data with ML-feature learning does address some of the limitations of only using NTLs as a proxy for electricity status (especially for more remote, rural areas).

4.3.2 Applying the Framework: Rank Order Data Types


REM Input
Electrification Status
Data Type [1]
Falchetta (2019) Gridded Dataset
Data Type [2]
ML- powered Nighttime Lights Data
Data Type [3]
Nighttime Light Composites

In the direction of **decreasing** data type quality & validation accuracy.

Figure 4.3.4- Rank Ordered List for Electrification Status Delineation Data Types. Ranking of validation accuracy is presumed and not quantitatively measured.

When deciding how to rank data sources on electrification status, sources that reached the widest number of countries, were easy to implement, and were already validated were ranked highest. Of all the data sources described in Section 4.3.1, only the Falchetta gridded electricity status dataset meets each of these requirements. Additionally, the layering of the NTL data with consumption tiers helps segment residential loads into different load-profiles. Rural residences are necessarily going to have different daily usage trends, different appliances, and different reliabilities than residences located in dense, urban cities. Consumption tiers eliminate the guesswork of which of our residential buildings from Section 4.1 belong to the two profile types.

Data Type [2] has promise at improving the detection of lit-up areas (i.e., nighttime light radiances) by learning additional features that indicate electrical (or economic) activity. However, it is not out-of-the-box ready to overlay on building locations and classify each as electrified. Using nighttime light composites, or Data Type [3], is equivalent to the method used by Falchetta for its binary classification of electrified areas. The only difference is that the authors of Falchetta (2019) already pre-processed the NTL images to remove noise and non-electrical lights, then re-labelled the intensities as with or without electricity. Therefore, unless a user prefers to implement their own threshold values, defaulting to one of the two 1-km^2 gridded NetCDF files is preferred. The only shortcoming to Data Type [1], or Falchetta (2019), is its annual count of "people without electricity between 2014-2018" is only useful if we know how many people in the cell *do* have access, otherwise we cannot know if the count represents 100% of the people who are settled in that cell. Therefore, if a user decides to use the five-year count instead of the 2018 consumption tier file, I suggest they pre-compute the total population number per cell and update each cell with proportion of the total population without electricity.

Finally, independent endeavors aimed at producing more precise estimates can leverage the harmonized NTL composites (Elvidge et al., 2021; Li et al., 2020) or ML classification models to improve upon the Falchetta (2019) dataset. Since the Falchetta (2019) dataset has openly published its code online, we can simply substitute their NTL intensity inputs with alternative, more granular or accurate versions.

4.3.3 Applying the Framework: Gridded Falchetta et al. Dataset for Electrification Status

The following section will describe two tasks. The first task is to assign residential or productive loads a consumption tier, simultaneously confirming their electrification status. The second task will estimate the peak-power of each residential load using the World Bank's Multi-Tier Framework (MTF).

Assigning Electrification Status to All Buildings

Fortunately, the Falchetta (2019) dataset is set-up as a grid and thus can be processed like an image (or raster file), with each cell analogous to a pixel. This lets me take advantage of Rasterio, a GDAL (*Geospatial Data Abstraction Library)* and NumPy-based Python library designed to simplify raster-data processing and analysis. For instance, Rasterio has several built-in functions that automate the task of assigning pixel band values to encased points or polygons. I will explicitly refer to these functions as I describe my method for delineating electrification status. The assignment of tiers or the 2018 grid's band value to loads can be broken down into four steps, a schematic of which is available in Figure 4.3.5:

1. Convert the gridded dataset from netCDF format into a GeoTIFF to make the file Python compatible. GeoTIFF files embed geospatial metadata into image files

(.tiff) that describes the actual location in space each pixel in the image represents. The metadata includes information about the spatial extent of the image, the coordinate reference system[9] (CRS) used to store the data, the spatial resolution of the image (i.e., the number of independent pixel values per unit length), among other geographic properties of the captured land-area.

2. Once converted, the raster image is clipped to just the boundaries (or extent) of the study region. For an entire country this entails clipping the image to Level-0 Administrative Boundaries.

3. Read-in the ESRI shape files of each load type you want to classify, be it residential or otherwise, as a separate GeoDataFrame (since they must be handled separately). All GeoDataFrames should be reprojected to the same CRS as the raster image (as should the administrative boundary file) to avoid any misalignment of geospatial information.

4. Apply Rasterio's `Rasterio.sample()` module to sample each building coordinate, check the pixel in the raster image it corresponds to and output the pixel's band value – in this case, the consumption tier. This value can then be appended as a new feature column in the GeoDataFrame. The only drawback is that as the number of coordinates approach multiple millions, the computation time of the module becomes much longer. However, users are unlikely to apply this method to more than 100 million buildings at once.

Loads assigned a band value of 0.0, or no value, correspond to regions with detected nightlight radiances below the minimum cut-off set by the authors of the dataset. Therefore, these loads are assumed to be unelectrified and segmented into a separate GeoDataFrame as loads with tiers of 1.0 or above. To compare the results to country statistics, sum up the number of electrified loads then divide by the total number of

[9] Coordinate Reference Systems (CRS) refer to the way spatial data that represent the Earth's 3-dimensional surface is transformed into a 2-dimensional representation. Each CRS employ their own mathematical framework to perform the transformation, producing different coordinate system grids.

loads, the ratio should approximately match the national or regional electrification rate.

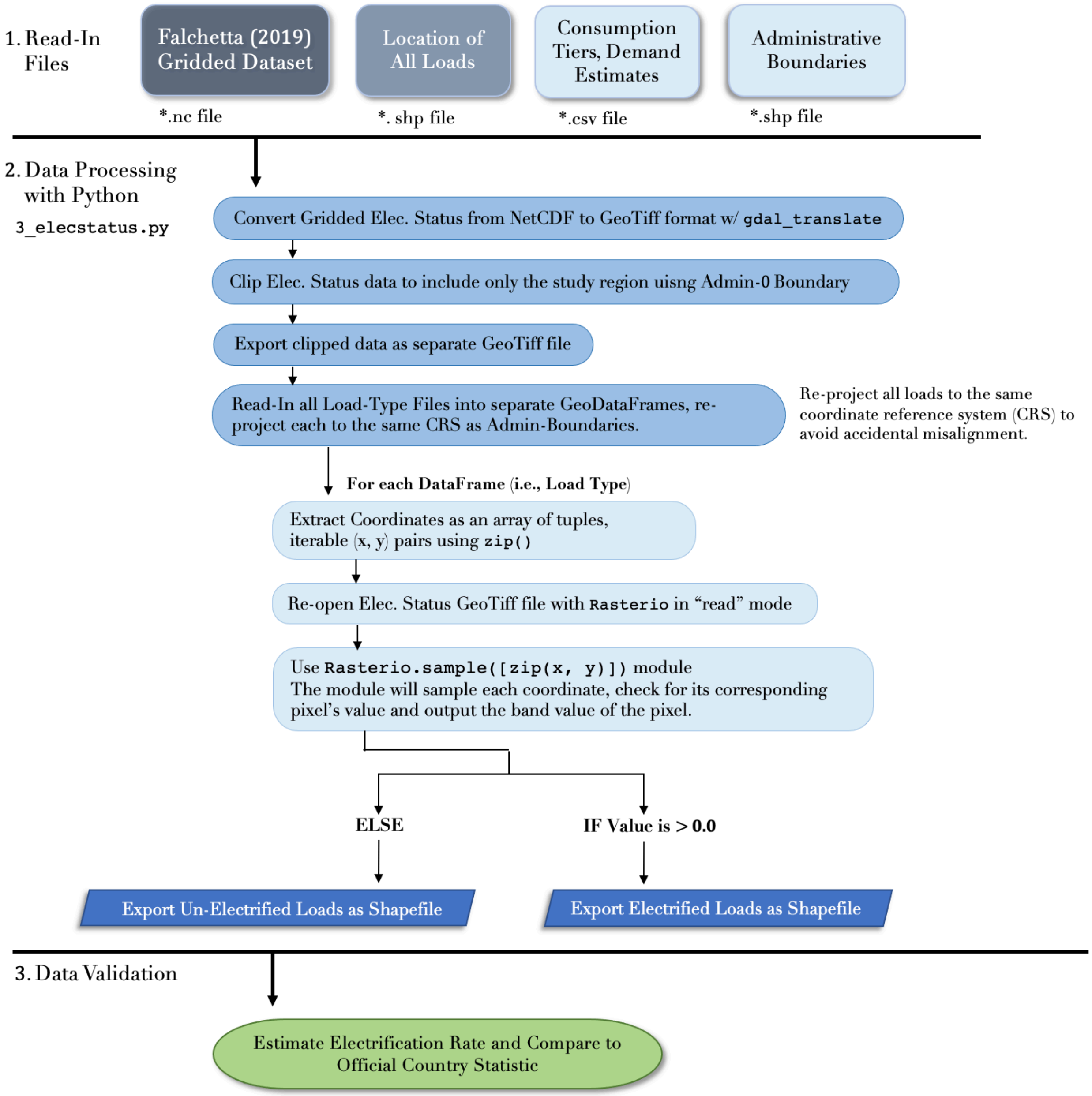

1. Read-In Files
Falchetta (2019) Gridded Dataset
Location of All Loads
Consumption Tiers, Demand Estimates
Administrative Boundaries
*.nc file
*. shp file
*.csv file
*.shp file
2. Data Processing with Python
3_elecstatus.py
Convert Gridded Elec. Status from NetCDF to GeoTiff format w/ gdal_translate
Clip Elec. Status data to include only the study region uisng Admin-0 Boundary
Export clipped data as separate GeoTiff file
Read-In all Load-Type Files into separate GeoDataFrames, re-project each to the same CRS as Admin-Boundaries.
Re-project all loads to the same coordinate reference system (CRS) to avoid accidental misalignment.
For each DataFrame (i.e., Load Type)
Extract Coordinates as an array of tuples, iterable (x, y) pairs using zip()
Re-open Elec. Status GeoTiff file with Rasterio in "read" mode
Use Rasterio.sample([zip(x, y)]) module
The module will sample each coordinate, check for its corresponding pixel's value and output the band value of the pixel.
ELSE
IF Value is > 0.0
Export Un-Electrified Loads as Shapefile
Export Electrified Loads as Shapefile
3. Data Validation
Estimate Electrification Rate and Compare to Official Country Statistic

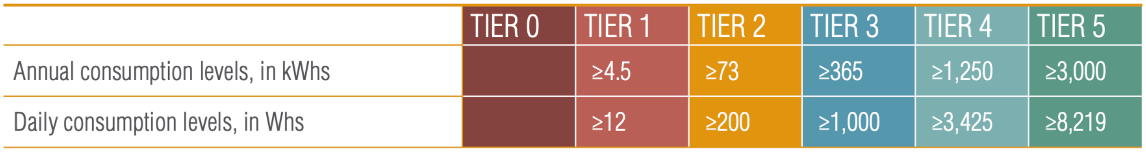
Figure 4.3.5- Stepwise Process of Delineating Electrification Status to Geolocated Loads. The first row lists the necessary datasets. The round-edged modules represent subsequent steps in the method. Indentations and shading signify embedded and groupings of code, respectively. The final blue parallelogram is the output of the process, and the final step (in green) signifies the validation test for the proposed methods.

Assigning Estimates of Peak Demand to Buildings by Consumption Tiers

Section 4.2.1 outlined the role of the network planning model, RNM, in REM's cost-assessment of different electrification modes (GE, MG, SHS). One of RNM's required inputs is an estimation of a customer's peak power demand. When energy usage data is unavailable for a country, UEAL researchers have settled for approximations of different load profiles (or demanded appliances) from prior studies or similar customer types in nearby regions. The advantage of the Falchetta dataset is that its consumption tiers are broadly linked to the results of the World Bank's MTF – specifically the *Multi-Tier Matrix for Measure of Household Electricity Consumption* (see Table 4.3.6). The MTF measures access to electricity based on technology-neutral multi-tiered standards where successive thresholds for supply attributes allow increased use of electricity appliances (Bhatia & Angelou, 2015). The data is acquired through multiple long-term studies conducted in several countries in SSA, Rwanda included. Supplementary Table A.1.1 summarizes these key supply attributes, which include capacity, duration, reliability, quality, affordability, legality, and safety characteristics. Table A.1.2 describes the minimum capacity requirements for each tier and Table A.1.3 links this information to average household load per tier and their expected appliance use. I map these minimum supply requirements to Falchetta consumption tiers as follows:

- Falchetta and MTF Tier 0 translates to no access to electricity.
- Falchetta and MTF Tiers 1 and 2 corresponds to households or low-voltage productive and community facilities that are electrified by small decentralized systems, SHS or Micro-Grids.
- Falchetta and MTF Tier 3 are households or any-sized productive and community facilities that are supplied by larger Mini-Grids or Grid Extension.
- Falchetta Tier 4 corresponds to MTF Tier 4 and 5, all loads are supplied by GE.

Table 4.3.6- Multi-Tier Matrix for Measuring Household Electricity Consumption. Extracted from *Beyond Connections: Energy Access Redefined* (Bhatia & Angelou, 2015).



	TIER 0	TIER 1	TIER 2	TIER 3	TIER 4	TIER 5
Annual consumption levels, in kWhs		≥4.5	≥73	≥365	≥1,250	≥3,000
Daily consumption levels, in Whs		≥12	≥200	≥1,000	≥3,425	≥8,219

For Residential Loads:

I apply my mapping of Falchetta consumption tiers to MTF access tiers and extract the cut-off consumption values from Table 4.3.6 and A.1.3. I use these cut-off values to set a minimum and maximum power demand for households in each respective tier. I assume the cut-offs represent the lower and upper limits of the 95% confidence level of a normal distribution. I estimate the shape of the distribution by computing the mean and standard deviation, then sample randomly from the distribution to assign each household, in each tier, a unique peak-power value. For implementations that would prefer to use coarser peak power estimates, a fixed number of samples can be obtained from the distribution (e.g., 10) and households in that tier-class can have 1 of those X (10) sample values. Figure 4.3.7 are the histograms of sampling and assigning power estimates to each household in my Rwanda case-study.

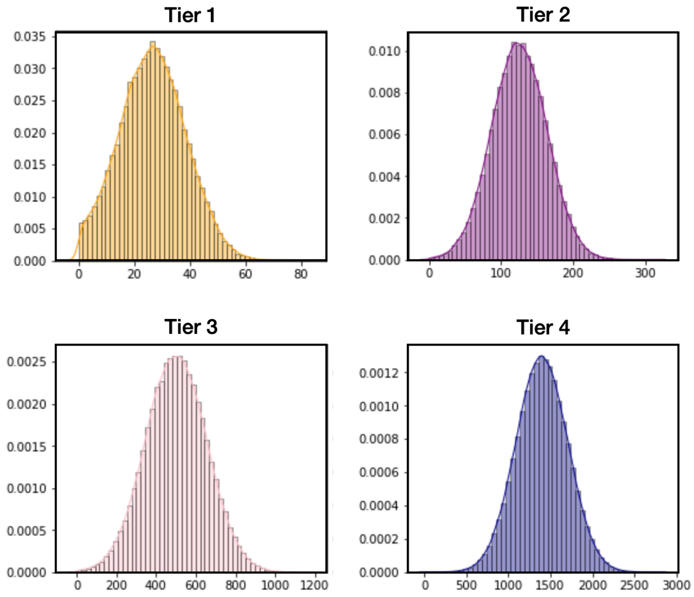


Figure 4.3.7- Normally Distributed Power Values per Tier.

Productive and Community Facility Loads:

As mentioned, urban and rural productive and community facilities consume power very differently. We can use Falchetta's tiers to downscale or upscale baseline load profiles according to which tiers these loads are classified into. Doing so would accommodate for the loads different supply requirements under different tiers. I downscale productive and community loads in Falchetta Tiers 1 and 2 by the same multiplier. The value of this multiplier is set to the ratio of the maximum daily consumption value in MTF Tier 3 to MTF Tier 2, which is equal to 1/5 (or 0.2). In keeping with this scaling method, I set Tier 3 loads as equal to the baseline consumption profile, and the Falchetta Tier 4 multiplier is equal to the ratio of the average maximum daily consumption of MTF Tier 4 and Tier 5 to MTF Tier 3, which is equal to 5.8. Recall that I assume users of the framework have pre-specified the baseline daily and annual consumption trends of each archetype of productive and community loads. While this is a strong assumption, demand estimation and prediction deserve independent study and as of yet no publicly accessible data can substitute ground-data information on how a country's urban, peri-urban, and rural regions consume power.

Case Study: Rwanda Residential Building Data

I apply my method to the residential buildings extracted in Section 4.1.4. The method generated an underestimation of Rwanda's electrification rate by approximately 30% (see Table 4.3.8). This is likely due to Falchetta's use of NTLs as the primary proxy for electrification status in a country with a large rural population. Rural areas are generally less detectable because they are often without street lighting. In a future iteration of the framework, I will experiment then employ one of the many NTL augmentations reviewed earlier in this section to try and improve their prediction of electricity.

Table 4.3.8- Comparison of Estimated and Reported. Rwanda Electrification Rate

Estimated Electrification Rate	Reported Electrification Rate	Error (%)
26.3	37.8	-30.5%

For example, future work should scale the ML method of feature learning from daytime imagery to countries in SSA, or layer NTL radiances with additional indicators of electrification (e.g., transformer or substation location). Both are likely to improve electrification status estimation. Future studies should also include more precise validation methods. For example, researchers should obtain ground-confirmed gridded electrification status data for rural villages in multiple countries and conduct a one-to-one comparison between it and each cell or pixel of their predicted dataset.

In the end, the remote detection of electrification status is one of the most crucial components of monitoring global progress on SDG7 or achieving universal access to affordable electricity. Chapter 5 further motivates the critical nature of better electrification status delineation by evaluating its implication on the prediction of the extent of the MV distribution network.

Chapter 5

Data Sources and Processing: Distribution Network Module

This chapter describes available datasets and tools capable of estimating information that may be missing from the medium voltage (MV) network layout of a study-region. The second half of the chapter introduces a small-scope sensitivity analysis of Rwanda comparing the MV network layout that has been estimated using REM to the existing layout. This case example uses validated data from the Universal Energy Access Lab (UEAL)'s recent collaboration with the country to produce a 2030 plan for universal electrification (referenced in Chapter 2).

5.1 Estimating the Medium Voltage Network Layout

The characterization of the extent, geolocation, and voltage of the MV distribution network is a critical and invariable component across all tools and frameworks introduced in Chapter 2. For context, electric grids are characterized by three primary types of infrastructure: the transmission network, the distribution network, and substations or transformers. The transmission network consists of high voltage (HV) lines (greater than 70 kV[10]) used for the bulk movement of electricity. The distribution network delineates how medium- ($10 - 70$kV) and low-voltage (LV) (<10 kV) lines distribute electricity from transmission networks to end-users. Substations and transformers are used to convert lines between high, medium, and low voltages. To decide if a consumer should be connected to a grid, a mini-grid (MG), or a solar-home system (SHS), accurate and up-to-date records of existing grid infrastructure are required since it may be more economic to connect unelectrified settlements to the national grid if they are near service transformers or MV lines. Most electrification planning tools have narrowed existing infrastructure to mean, at minimum, the MV distribution network. REM further requires users

[10] This is not an absolute amount since it depends on the network characteristics. In power systems with strong consumption densities and vast territories, like in the European Union or North America, transmission network is defined as the network with a voltage of 220 kV and above.

to define potential connection points (i.e., extensions) to the grid by supplying the location, reliability, and capacity of the existing MV distribution lines. Recall from Chapter 2, that each MV segment is assigned an energy price in $/kWh which incorporates upstream transmission network and generation costs as local wholesale energy costs at the MV distribution level. REM also allows for the hourly specification of reliability and electricity costs for each feeder, or MV segment (Ciller et al., 2019). As such, the final input module to my Framework is an open-source method to extract or replicate the MV network of any country and an option to model the supporting upstream infrastructure (e.g., transmission and generation).

Generation information is commonly sourced from the up-to-date and comprehensive WRI Global Power Plants Database, discussed in Chapter 3. High-voltage transmission and distribution infrastructure data have long been open and available for public access. However, georeferenced information on medium- and low-voltage distribution networks is often unavailable, inconsistent, or incomplete. The low quality (erroneous or inadequate status) of these distribution network datasets has a considerable impact on the results of least-cost electrification models, i.e., REM. The cost of extending the grid is high, and thus the distance between an existing settlement and the grid is a core driver of the cost in optimization models. Additionally, because the errors in the distribution grid datasets are more likely to exclude lines rather than show them, models may underestimate the extent to which the grid connections are the cheapest option, either biasing the results of the model toward off-grid technologies or more expensive buildouts of the existing grid (Morrissey, 2019). This in turns increases uncertainty and reduces the applicability of outputs from REM (or similar models).

5.1.1 Data Sources and Limitations

The existing information on MV network layouts can be broadly classified into two types. The first type includes crowd-sourced information, aggregated data from project archives, or public datasets shared by countries or utilities. The second data type includes inferred maps of the network from existing open data, like remote sensing or satellite imagery. This section focuses on the most complete and granular example of the second data type: *gridfinder*.

The largest global database of crowdsourced information on infrastructure is OpenStreetMap (OSM). Data from OSM give a sufficiently detailed picture of high-voltage transmission lines. In some areas, they are comprehensively traced from satellite imagery and ground-truth data, while in others (particularly developing countries) accuracy and completeness is variable, depending on whether someone has already mapped an area (Arderne et al., 2020). Alternatives to OSM are the compiled databases: Africa Electricity Grids Explorer and ECREE (Centre for Renewable Energy and Energy Efficiency) GeoNetwork Catalog. The Africa Electricity Grids Explorer is an online data explorer that records existing and planned transmission and distribution lines for the entire African continent and the Middle East. The data explorer serves as an updated and improved replacement for the Africa Infrastructure Country Diagnostic (AICD) published in 2007. The primary sources of the database are AICD, OSM, country utilities, the West African Power Pool (WAPP) GIS database, World Bank project archives, and International Bank for Reconstruction and Development (IBRD) maps. However, like OSM, the completeness and quality of the data varies country to country. Additionally, data from the AICD and the World Bank is often out of date – many new lines may not be included, and planned lines may have completely changed. Where possible, this is improved using data from other sources (The World Bank, 2019). The ECREE data catalog provides the existing and planned distribution grid network (between 11 – 33 kV) in the ECOWAS (Economic Community of Western Africa) region. Sources included national utilities and electrification agencies in the region. Whether the information for the ECOWAS region is complete and up to date is unclear from available documentation.

The variable quality of these datasets is attributed to the large number of distribution infrastructure operators and owners required to contribute data, compared to transmission operators which are often single national-scale utilities. Grid infrastructure data may also not exist in GIS-ready formats or may be restricted from public access for security reasons. Finally, digitizing grid data can be prohibitively difficult. For example, due to underground distribution grids or small overhead line structures aerial imagery, cannot be used to detect or delineate on-the-ground infrastructure (Arderne et al., 2020).

Gridfinder is an open-source dataset, based on grid mapping work at Facebook, that uses monthly and annual NTL data in concert with other widely available infrastructure data to produce the first composite map of the global power grid (Arderne et al., 2020; Gershenson et al., 2019). To build this dataset, Arderne et al. (2020) begin with a custom image filtering process of the monthly NTL satellite imagery to identify locations with consistent illumination, and thus locations more likely to be producing light from electricity (referred to as target cells). The authors assume these locations are grid powered and output a global two-dimensional array of target coordinates (target cells) that must be connected by network lines. To estimate the layout of the electricity network, they apply a many-to-many variant of Dijkstra's algorithm (shortest distance between two points) on each target cell. The algorithm is weighted by a cost function based on existing roads from OpenStreetMap. Existing grid lines from OSM are assigned a cost of 0.00; the largest roads (motorways) are assigned a cost of 0.10, the smaller roads (tertiary) a cost of 0.17, and the others in between; their threshold values are set arbitrarily as a "reasonable" strategy of preferring larger roads. Areas with no roads are assigned a cost of 1.00. This results in a second two-dimensional array that represents the cost of traversing each cell between the target cells. The algorithm attempts to connect every location with the shortest possible distance while following roads (preferring larger) where possible (Arderne et al. 2020). Note, to remove predicted lines in remote areas, which are unlikely to be connected to the main network, Arderne et al. (2020) use a buffer of 100 km on the OSM data and remove all lines beyond that. The final output is a raster estimate of whether cells contain a medium-voltage line or not.

To validate their medium-voltage results, Arderne et al. (2020) consider 16 different networks from 14 countries. Each represents a wide range of electricity network development and accessibility. Datasets for these countries were either openly available, accessed under special licenses for the research group, or supplied by the World Bank Group. The countries represented in each income group are: Australia, Netherlands, New Zealand, United Kingdom (High Income), Namibia (Higher Middle Income), Bolivia, Kenya, Nigeria, Zambia (Lower Middle Income), Burundi, Ethiopia, Malawi, Tanzania, and Uganda

(Low Income). To quantify model performance, observed lines were compared to lines produced by *gridfinder* in an equal-area cell grid, fitted to the electricity distribution extent of each validation set.

To evaluate the performance of *gridfinder* I assessed two metrics: precision and accuracy. High precision suggests the model is correctly estimating the presence of lines, and not portraying a false sense of accessibility. Accuracy measures the total correct predictions over the total validation set, which might be skewed in areas where it is easy to predict the absence of lines (e.g., mountain ranges or deserts). Arderne et al. (2020) report their performance metrics at the cell-resolution that produced the best results, which is 26 km at maximum and 2 km at minimum. Validation is therefore conducted at very coarse resolutions, which calls into question how well the data will perform at the granular resolutions needed by REM. Overall, *gridfinder* produced accuracies between 70-80 percent and precision values between 60-80 percent in low-income countries; globally the average values were 75 percent and 76 percent respectively.

The performance of *gridfinder* is better than low-quality or erroneous public datasets, but not good enough for individual building-level electrification plans. Their design of the network is based only on geometric congruity with roads and OSM data. They do not perform power flow or voltage drop analyses. Furthermore, by identifying electrification targets with only NTL data, the model is hindered by NTL's poor detection of light in rural or peri-urban areas. Correa et al. (2021) improves upon *gridfinder* by proposing a learning model that improves the detection of electrified sites, validated on ground-truth data from Kenya. Their model uses more granular (e.g., daily) composites of NTL data and improves detection accuracy by up to 7% and identifies approximately 78% of electrified sites that were previously missed. Correa et al. (2021) also recommend supplementing NTL data with other publicly available grid infrastructure data, (e.g., generator or substation locations). I try to address each of these limitations by leveraging REM's built-in RNM tool to design what the grid should look like given my estimated building locations, their respective demand, and the location of supporting infrastructure like HV substations and distribution transformers. The proposed method is outlined below, in Section 5.1.2.

5.1.2 Applying REM to Medium Voltage Network Estimation

The RNM module in REM provides an unexplored method for estimating MV network designs. Recall, RNM designs the optimal distribution network to supply georeferenced customers with the electricity they demand at a prescribed level of hourly reliability. RNM exists as a stand-alone model and is available in two formats: Greenfield and Brownfield. Greenfield RNM can plan a network from scratch without any information on existing wires or the area network. Brownfield RNM plans expansions to an existing network, using current grid connections as inputs. In this section I propose leveraging the design capabilities of Greenfield RNM to replicate the existing network in a study region. At minimum, Greenfield RNM requires: 1) the geolocation and characteristics of LV, MV, and HV customers, 2) the geolocation and characteristics of HV-transmission to HV-distribution substations, and 3) a catalog of the technical and economic characteristics of the equipment. In this case, I supply the model with the geolocations of (presumed) electrified residential and productive customers, extracted in Section 4.3, their respective energy demand, and a country-specific network catalog. In addition, I treat any openly available information on the existing grid as a-priori information regarding where the grid network is most likely built-out, feeding the open data as inputs to RNM. An example of this a-priori data is the location and voltage of transformers or substations found on OpenStreetMap.

However, a disadvantage of running RNM as a standalone model is its requirement that each of the inputted infrastructure types are labelled with precise technical and regional characteristics. Obtaining this level of detail from open sources might be infeasible or very uncertain. A work-around is to use RNM via REM. REM generalizes all inputted data regarding the grid to "potential connection points for grid extension"[11], thus enabling users to input open grid data as non-specific supplemental information, despite being incomplete.

[11] MV lines and other supporting infrastructure (e.g., transformers) are represented by a set of short line segments in REM, each defined by two coordinates. REM uses an appropriate subset of these candidate MV connection points for each cluster of consumers and submits them as inputs to RNM. RNM is then free to use any of these points to provide an optimized distribution layout of a grid-extension solution for the respective cluster (Ciller et al., 2019).

Accessing RNM through REM has several other advantages. First, it limits computational troubleshooting to a single interface, REM. RNM is written and initialized with the C programming language which might make it inaccessible to less experienced programmers. Second, REM allows splitting consumers into regions and running the optimization for each region in parallel. This breaks the design problem into several instances of RNM running at once, and thus reduces the overall computation time.

We can estimate the MV network with REM by configuring the input parameters to force an all-grid-extension REM output. I reconfigured the REM parameters most relevant to the network layout as outlined below:

- The Cost of Non-Served Energy (CNSE): Ascribing higher social costs to non-severed energy make it more economical to electrify customers with more reliable technologies, i.e., grid extension. This is only true if grid distribution lines are initialized with high reliability levels for the specific REM case.
- The Per-Customer Cost of Off-Grid Alternatives: Change the per-customer cost of stand-alone systems and mini-grids from $0 to $1,000,000 and $65 to $1,000,000, respectively. $1,000,000 is an arbitrary choice intended to make the electrification mode prohibitively expensive.
- The Annual Management Costs for Mini-Grids:
    - The number of consumers allowed in a mini-grid was increased from 1 to 50 and 100 to 500 for small and medium sized mini-grids, respectively. Off-grid clusters with less than 50 people are eliminated.
    - The annual management costs ($/year) of smaller sized mini-grids were increased to be comparatively more expensive than larger mini-grids (by at least an order of magnitude). This economically favors the construction of larger off-grid clusters (i.e., at least 500 customers in the cluster).

Together these parameters should make off-grid alternatives not cost-competitive and produce large grid-extension clusters. Ideally a single cluster per region will emerge, ensuring that RNM is called only once per region and therefore is considering the optimal distribution network for the whole administrative area at once. Exact values for these parameters will vary by the country and context of a project, therefore they are unspecified in this work. Furthermore, despite artificially reducing the cost of grid extension, REM's cost optimal solution may still produce small clusters electrified via off-grid technologies. These clusters will typically make up no more than 1% of the total customers which is an acceptable margin of error given the uncertainty of using NTLs as a proxy for electrification status (by means of the Falchetta et al. dataset). In the end, I expect RNM's techno-economic objective function to better align with the decision making of distribution utilities who planned and built the existing network; more so than the predominantly geometric design algorithm of *gridfinder*.

As REM is a proprietary software, applying it to estimate the medium voltage distribution network does not constitute an open-access solution for all users of the framework. As such, the notion of how this method would fit within the GitHub open framework, designed in Chapter 3, remains an ongoing discussion. Readers should accordingly interpret this section as an exploration on the efficacy of reproducing a country's medium voltage network with REM. Future work might also consider producing an open dataset of REM's estimates of the local distribution grid for all countries in Sub-Saharan Africa. The remainder of the section will delve into greater detail on the estimation method.

Pre-processing the Input Data

The minimum information needed to successfully run Greenfield RNM by means of REM include the following:

1. **Geolocation and Demand of HV, MV, LV Customers.**

   1.1 Extract Customers and their Demands.

   REM does not allow users to differentiate between HV, MV, and LV consumers. It assumes all consumers are low voltage unless their demand is too high for the LV network wires supplied in the network catalog, in which case it would power them with MV or HV lines. Chapter 4.1 and 4.2 outline how to identify the building coordinates of LV and MV customers. These loads constitute residential, social, and productive customers. Each customer is linked to a customer archetype, from which they are assigned a peak-power-demanded value.

1.2 Extract Electrified Customers

Given that we are only interested in customers connected to the grid, I assume all customers classified as “lit” by nighttime lights are on-the-grid. The assumption holds because nighttime lighting predominantly indicates streetlights, the existence of which is more likely correlated with larger, grid connected areas than smaller mini-grids or stand-alone generators. The Falchetta et al. gridded dataset, described in Chapter 4.3, is used to extract just the electrified, i.e., grid-connected customers from the results of 1.1.

1.3 Cluster Customers

REM can only handle a limited number of customer archetypes (i.e., customer demands). Therefore, to simplify the computational task, consumers of similar peak demands are assigned to the same cluster. Each cluster is then assigned a characteristic demand pattern. I apply a k-means clustering algorithm to cluster the electrified customers. K-means clustering partitions n similar-demand observations into k clusters. Each observation belongs to and takes the value of the cluster with the nearest mean (or centroid). To determine the optimal number of unique demand values, i.e., clusters, I recommend plotting a histogram of all customers’ demands and experimenting with different bin sizes which reduce the number of unique values without eliminating the heterogeneity of demand. In my experiment, elaborated on in Section 5.1.3, 300 clusters were used for residential loads and 300 clusters for social and productive loads.

1.4 Administrative Level

Finally, customers are distributed into the Level-2 Administrative region they are located within. As mentioned, distributing the customers allows for parallel runs of REM and produces results quicker.

2. Open Grid Infrastructure Data.

Overall, any data that signals to RNM where medium voltage lines are likely located is useful input data to REM. Users should begin by searching open databases for existing infrastructure in their area of study. Of this infrastructure, information on distribution transformers (medium- to low- voltage transformers, in particular) is the most useful, as they are necessarily built along the MV lines. Other infrastructure types include substation locations, generation sites, HV distribution lines, and transmission lines. Once collected, distribution or transmission lines should be converted into a series line segments, or infinitesimally short line segments for point data (e.g., generation plants). This can be done with the Shapely library in Python. The coordinates at either end of the line segment should be provided to REM for all infrastructure types in a csv file. Recall, REM will interpret this data as “potential connection points” for grid extension.

To explore how well each grid infrastructure type can predict the extent of a local MV distribution network, I plot data provided to the UEAL on Rwanda’s grid in Figure 5.1.1 and 5.1.2. Since the plotted data comes directly from local stakeholders, I assume it is credible and complete. For both figures I overlay the locations of different infrastructure types on the existing medium voltage network and the locations of Falchetta-Electrified Customers, which are the customers classified as electrified according to Falchetta et al. (2019)’s methodology. I supplement this map of Figure 5.1.1 with ground-truth information on the location of electrified productive loads in Figure 5.1.2.

While the existing MV lines do traverse all the Falchetta-Electrified Customers, they also extend far beyond these customers (areas). Therefore, building data masked by Falchetta’s classification of electrified areas will fail to characterize the expanse of the grid, and should not be used as the single proxy for the grid’s location. On the other hand, the ground-validated productive loads in Figure 5.1.2 exist in most of the areas where the MV grid exists, suggesting that their locations are better signals of the existence of a nearby distribution grid. It also suggests that improvements in electrification status estimation would translate to better estimates of the grid’s extent. Finally, the close proximity of the electrified productive loads to the Falchetta-Electrified Customers suggests that there is a reasonable radius for which the detection of nighttime-lights highly correlates with “lights” existing a radius r away. To test this hypothesis, future researchers should experiment with dilations of the nighttime lights’ raster data. In other words, experiment with “lighting up” the pixels surrounding a “lit up” pixel, then observe if that method does improve nighttime lights’ prediction of electricity distribution.

Comparing the predictive capability of each of the infrastructure types in Figure 5.1.1, the location of generation sites and substations is expectedly less informative than the location of MV to LV transformers. While it is difficult to obtain the location of these transformers from open data, it is easier and faster to request and obtain their location from local stakeholders than precise network line drawings. That said, the generation plants and substations do extend beyond the bounds of the Falchetta-Electrified regions, and thus are still useful for signaling to RNM that it should expand the network beyond those regions.

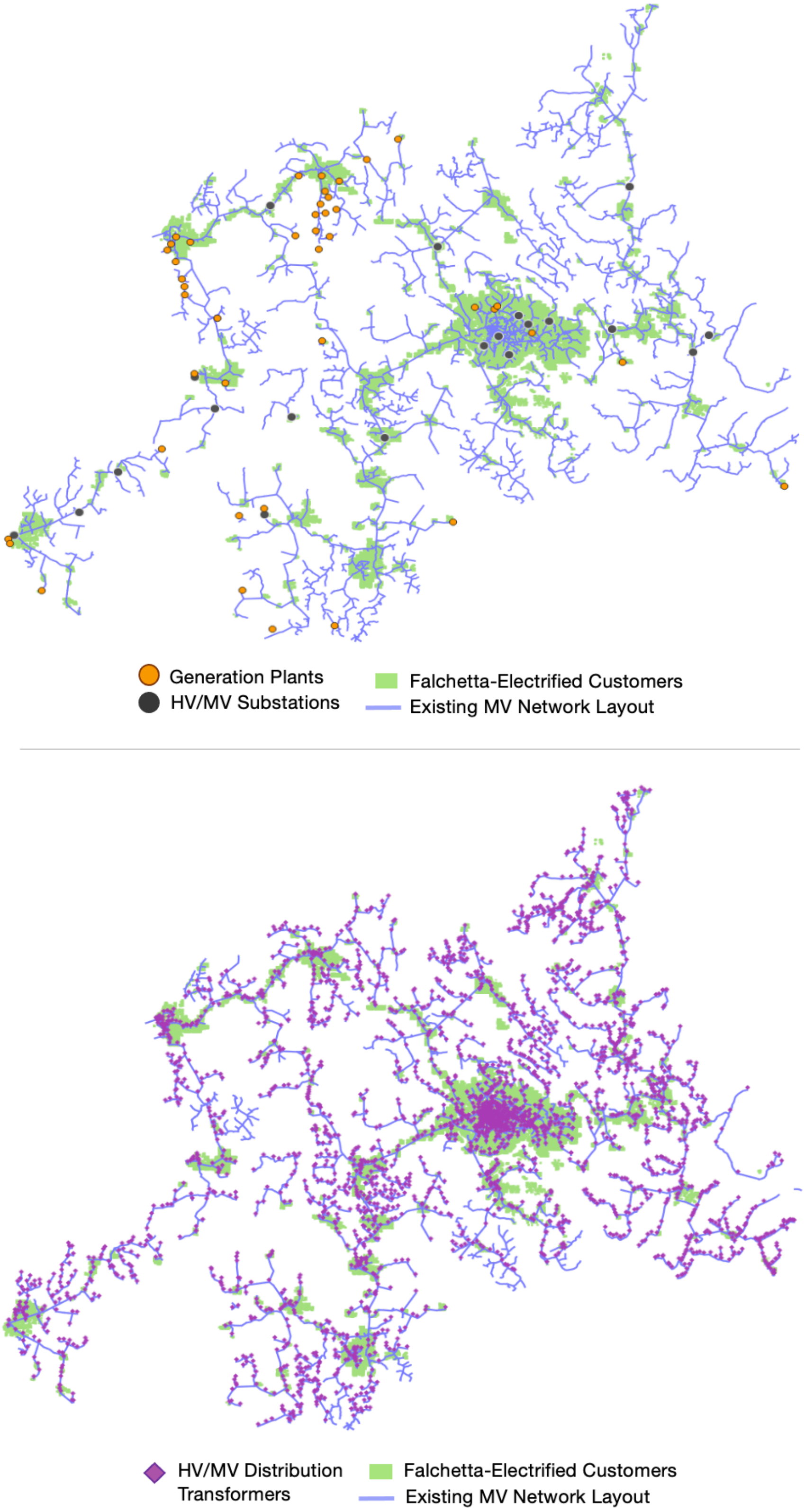

Generation Plants
HV/MV Substations
Falchetta-Electrified Customers
Existing MV Network Layout
HV/MV Distribution Transformers
Falchetta-Electrified Customers
Existing MV Network Layout

Figure 5.1.1- Overlay of Rwanda's Grid Infrastructure on MV Distribution Network and Falchetta-identified Electrified Areas.

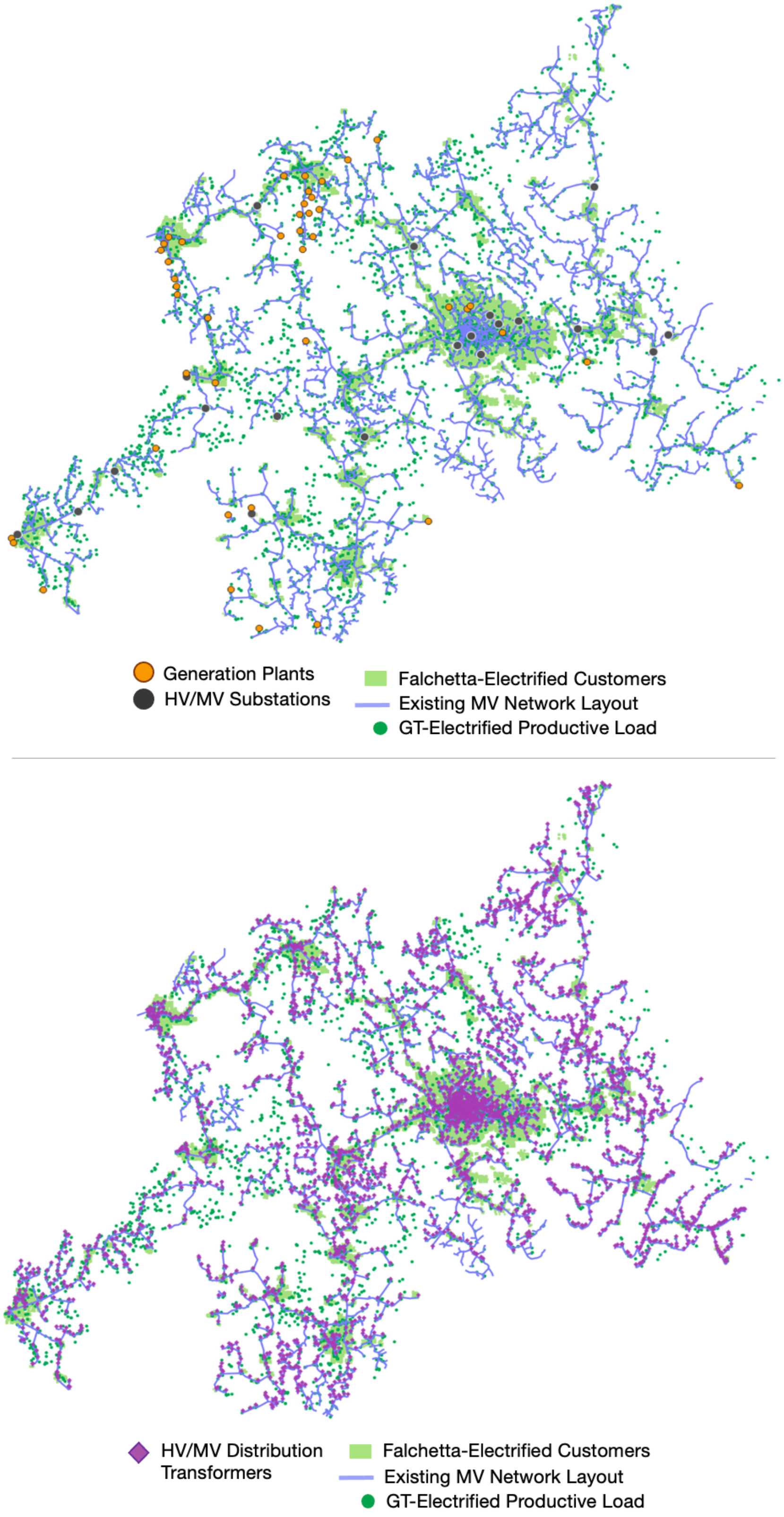

Generation Plants
HV/MV Substations
Falchetta-Electrified Customers
Existing MV Network Layout
GT-Electrified Productive Load
HV/MV Distribution Transformers
Falchetta-Electrified Customers
Existing MV Network Layout
GT-Electrified Productive Load

Figure 5.1.2- Overlay of Rwanda's Grid Infrastructure on its MV Distribution Network, Falchetta-Electrified Loads, and Ground Truth (GT)-Electrified Productive Loads.

3. Network Catalog

To accurately account for the cost and technical constraints associated with the physical assets on the distribution network, users must supply REM with a Network Catalog. The catalog is expanded on in Chapter 2 and 3. Each distribution utility company or country uses a specific catalog of equipment (e.g., specific types of cables, transformers, poles, batteries and solar panels). For the most correct implementations of REM these inter-country differences should be respected and a customized catalog should be supplied per region of interest. To gather this information users are encouraged to consult national reports on the country's energy sector or contextual information from regional grids. For example, documentation by utilities in nearby countries, World Bank archives, or any shared data from the power pool the country might belong to.

Post-Processing the Output

Once the user has successfully configured REM to produce an all-grid-extension case and fed it the necessary inputs, the model will output a series of MV line segments. These lines are produced by RNM and represent the extension to the initial "potential connections" line segments. Together, they form a discontinuous and incomplete representation of a distribution network. To convert these disjointed lines into a continuous network, I utilize the Minimum Spanning Tree (MST) algorithm. MSTs have been previously used by UEAL researchers for designing radial power distribution networks. MST is an edge-weighted undirected graph that connects all the vertices (nodes) together, without any cycles (i.e., no closed loop) and with the minimum possible total edge weight (in this case cost). That is, it is a spanning tree whose sum of edge weights (costs) is as small as possible. In this case, all line segments (translated into nodes, or points) are assigned close-to-zero weights and a continuous network is produced. Those results are explored in the following section.

Final Implementation

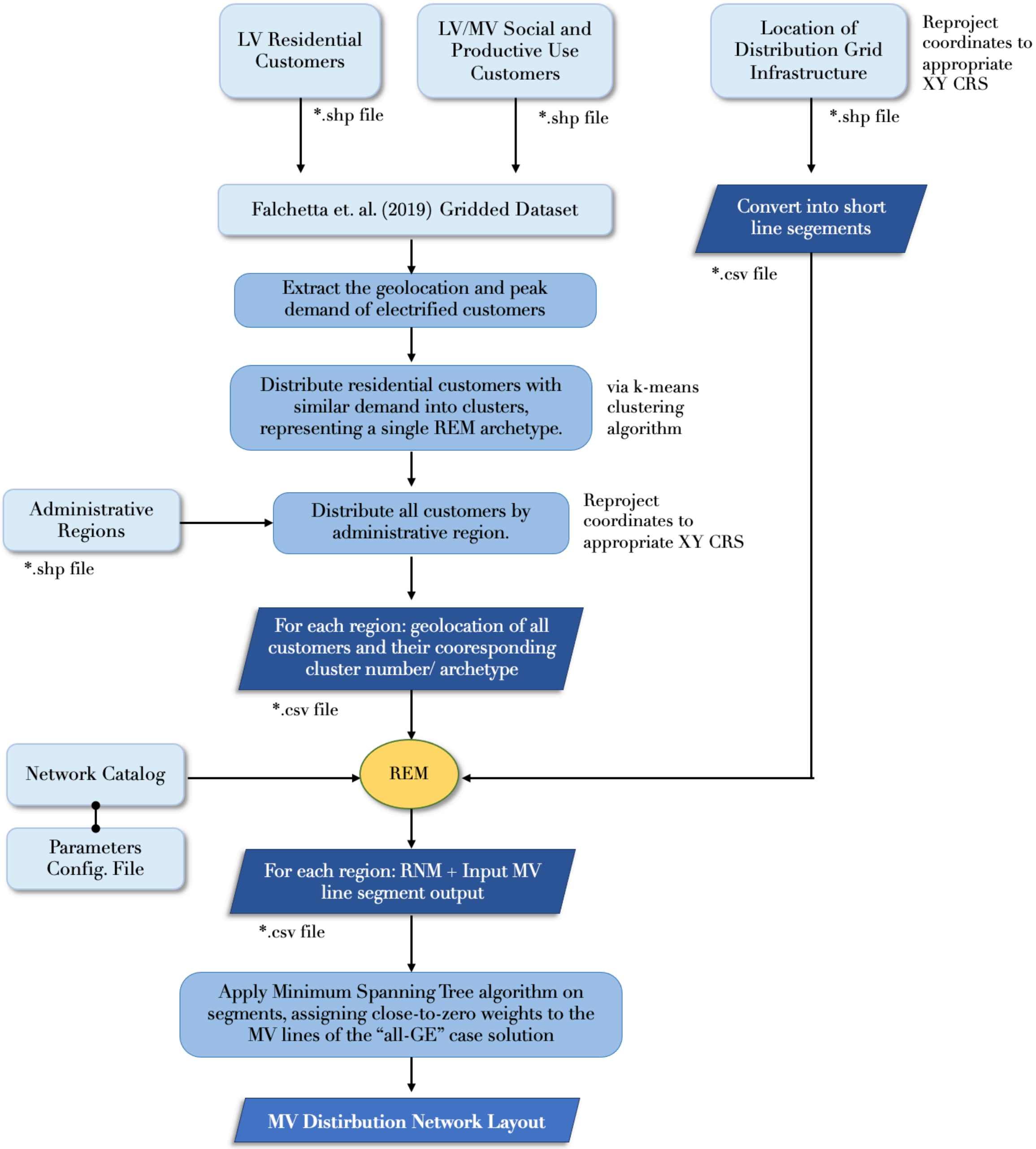

LV Residential Customers
LV/MV Social and Productive Use Customers
Location of Distribution Grid Infrastructure
Reproject coordinates to appropriate XY CRS
*.shp file
*.shp file
*.shp file
Falchetta et. al. (2019) Gridded Dataset
Convert into short line segements
*.csv file
Extract the geolocation and peak demand of electrified customers
Distribute residential customers with similar demand into clusters, representing a single REM archetype.
via k-means clustering algorithm
Administrative Regions
*.shp file
Distribute all customers by administrative region.
Reproject coordinates to appropriate XY CRS
For each region: geolocation of all customers and their coorresponding cluster number/ archetype
*.csv file
Network Catalog
Parameters Config. File
REM
For each region: RNM + Input MV line segment output
*.csv file
Apply Minimum Spanning Tree algorithm on segments, assigning close-to-zero weights to the MV lines of the "all-GE" case solution
MV Distirbution Network Layout

Figure 5.1.3- Stepwise Process of Estimating MV Distribution Network with REM. The light-blue modules represent necessary datasets. The rounded-edged darker modules represent subsequent steps in the method. The final blue parallelogram is the output of each sub-process. Each intermediary output serves as an input to REM. The final parallelogram output is the geo-referenced MV distribution network.

5.1.3 Testing the Model: Rwanda Case-Study

As proof-of-concept, I apply my network estimation model to the country of Rwanda. To run the case, I use the ground-truth data on electrified social and productive loads, as well as the location of medium voltage to low voltage distribution transformers (a total of 2,982 transformers) and high voltage to medium voltage substations (a total of 22 substations). Both the transformers and substations were converted to short MV lines segments such that REM considers them part of the existing distribution network. I used the location, demand, and electrification status of residential loads in Rwanda supplied by my framework (i.e., the outputs of Section 4.1 and 4.3). The network catalog used in my experiment was previously customized for Rwanda by members of the Universal Energy Access Lab. The results of the experiment are plotted in Figures 5.1.4-5.

To test the robustness and accuracy of the model, I ideally would only use open sources as inputs to the case. However, the Google Maps API extraction of productive loads has yet to be scaled to all of Rwanda and therefore was unavailable for the experiment. Similarly, publicly available grid infrastructure for Rwanda only included: transmission lines, power plants, and HV substations. Recalling my analysis on the predictive quality of all grid infrastructure types, distribution transformers are the most informative on the extent of the distribution network. As a result, I chose to use the transformer information supplied by Rwanda, in lieu of the publicly available infrastructure, to observe how REM performed under an 'ideal' grid data scenario – at first. If the initial experiment is successful, future studies can perform a sensitivity analysis on the proposed model and feed REM all possible combinations of publicly available grid data. The analysis should compare the accuracy and precision of each run's (i.e., data combination's) output to the existing MV network layout.

Nonetheless, my experiment showed very promising results. A qualitative comparison between the existing network and REM's outputs in Figures 5.1.4-5 confirms that REM extends the grid to most areas where the existing network is located. An overlay of the prediction and the ground-truth data in the top map of Figure 5.1.4 highlight the low-level discrepancies between the two. Note that REM and other planning tools primarily use information on the existing grid to measure the feasibility of local grid extension, i.e., the proximity of unelectrified buildings to gridlines and the cost effectiveness of building out the grid to supply them with energy. Hence, I expect REM's output to be good enough for the scope of my framework. The bottom map of Figure 5.1.4 shows that the network is built out congruently with the spatial distribution of the transformers. Hence without transformer locations, the best guarantee that REM produces a close-enough approximation of the grid depends on the access to accurate data on where customers are electrified. I also overlay the existing distribution network with *gridfinder's* prediction in Figure 5.1.6. Compared to REM, *gridfinder*, wholly underestimates the scale and extent of the existing grid, which likely translate to more expensive electrification planning models due to the presumed far-away distance between unelectrified customers and grid extension. Looking at the map on the right-hand-side of Figure 5.1.6, the predicted network exactly extends to wherever Falchetta et al., and thus NTLs, assumes customers are electrified. This corroborates my initial concern that *gridfinder's* dependence on roads and NTLs will be determinantal to its accuracy and useability as a Data Type substitute in the framework.

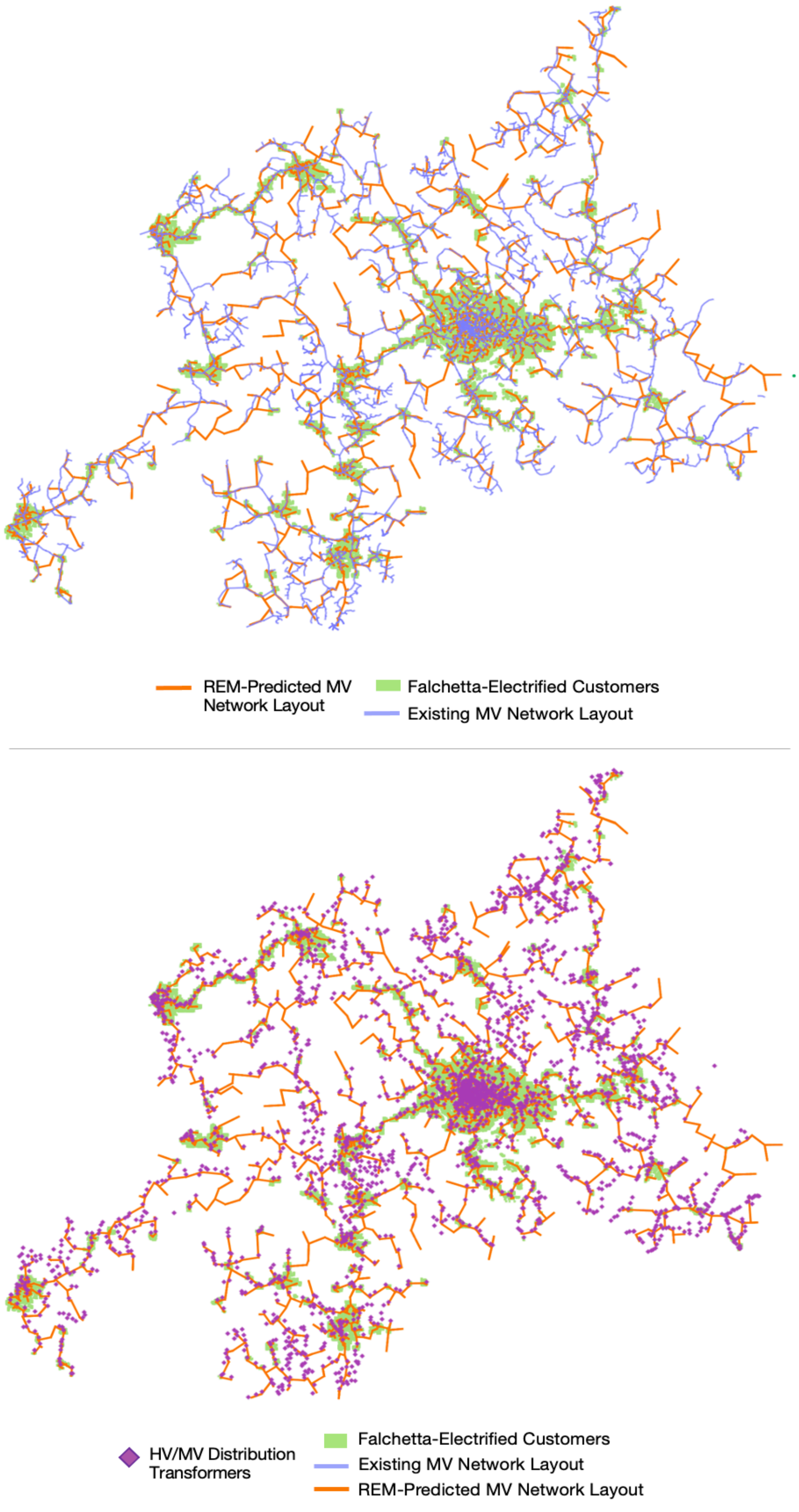

REM-Predicted MV Network Layout
Falchetta-Electrified Customers
Existing MV Network Layout
HV/MV Distribution Transformers
Falchetta-Electrified Customers
Existing MV Network Layout
REM-Predicted MV Network Layout

Figure 5.1.4- Overlay of the REM-Predicted MV Network Layout and the Existing MV Network Layout and Comparison with Transformer Locations.

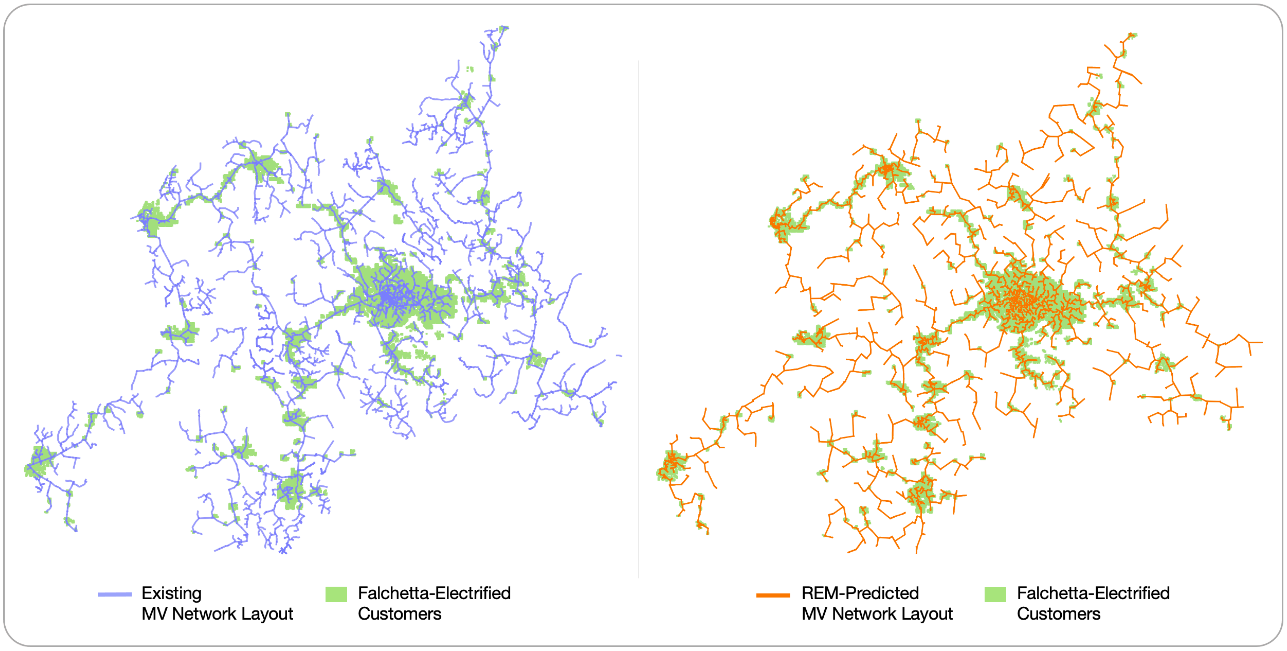


Figure 5.1.5- Side-by-Side Comparison of the Existing MV Network Layout and the REM Predicted Network Layout.

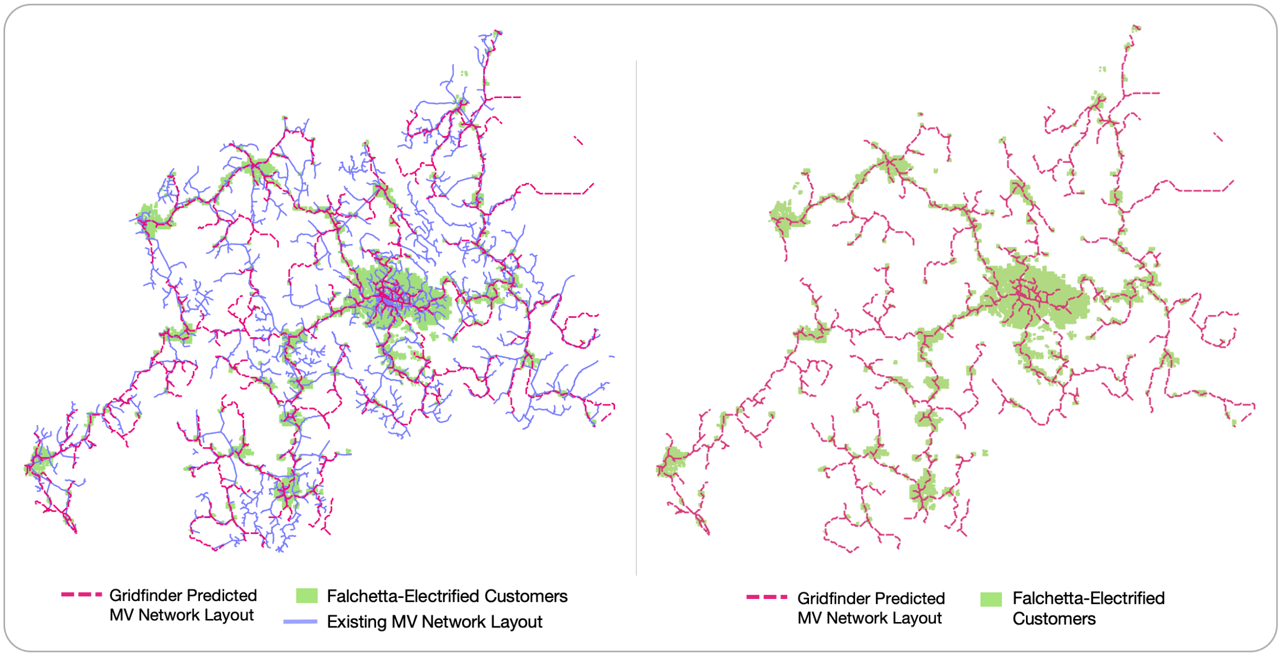


Figure 5.1.6- Overlay of the Gridfinder Predicted MV Network Layout and the Existing MV Network Layout.

Future Improvements to the Model:

Results in Figure 5.1.4 demonstrate the critical role distribution transformers play in accurately replicating the MV distribution network. Consequently, independent studies to develop a method for identifying distribution transformers should take precedence for improving REM's prediction of the network. UEAL researchers could attempt to predict the location of MV to LV transformers by running Greenfield RNM separately of REM. For example, researchers can attempt using the geolocation and demand characteristics of residential and productive loads with public information on high voltage feeders (high voltage lines and substations) and initialize Greenfield RNM to design the medium-to-low voltage network, which would include the location of medium-to-low voltage transformers. Alternatively, future experiments might simultaneously supply REM with local road-network data and the more widely accessible grid infrastructure (like substations or power plants) and test if the combination is an acceptable substitution for transformer data.

Though the qualitative comparison of REM and *gridfinder's* prediction was sufficient at highlighting which model was a better estimate of the existing grid, the approach is not conducive to running repeated analyses or expanding the area-of-study to multiple countries. Future validation methods could rasterize the files for the whole country and conduct a pixel-to-pixel comparison between each model output. This works by classifying each pixel, of each raster file (i.e., each model output) as containing a MV network line or not. In this way, established pixel-level comparison metrics can be utilized to measure the precision and accuracy of each model compared. Maximizing precision ensures a model is not overestimating the presence of lines and thus providing a false sense of accessibility. Accuracy, on the other hand, is simply a measure of the total correct predictions and can be misleading should there be large areas where it is easy to predict no MV distribution lines, e.g., extreme topographies.

A notable example of these evaluation metrics is Intersection-Over-Union (IOU), also known as the Jaccard Index. To compute the IOU, one begins by counting the number of true-positives (pixels where the model correctly predicts a MV line), false-positives (pixels where the model incorrectly predicts a MV line), and false-negatives (pixels where the model incorrectly predicts no line). The IOU measures how well the model predicts the known presence of distribution lines, disregarding any areas where the model and observed data agree there are no MV lines. An IOU equal to one is the most desirable for object detection models. In the end, quantitate validation metrics are replicable and scalable, thus key to ongoing sensitivity analyses on REM's ability to predict the existing medium voltage network.

5.1.4 Applying the Framework: Rank-Order Data Types

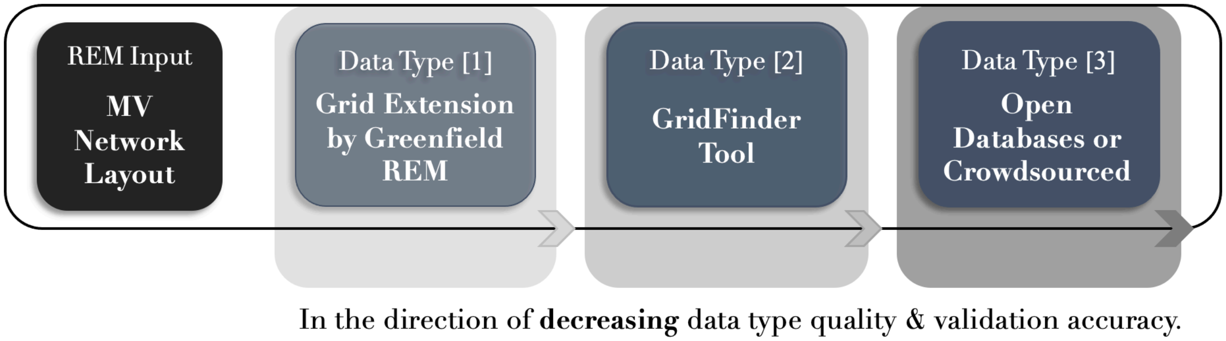

REM Input
MV Network Layout
Data Type [1]
Grid Extension by Greenfield REM
Data Type [2]
GridFinder Tool
Data Type [3]
Open Databases or Crowdsourced
In the direction of **decreasing** data type quality & validation accuracy.

Figure 5.1.7- Rank Ordered List for MV Network Layout Data Types. Ranking of validation accuracy is presumed and not quantitatively measured.

As I rank Data Types, it is important to emphasize that applying REM, a proprietary software, to the estimation of the MV grid does not constitute an open-source solution to the following Input Type. However, if we assume the models in Chapter 4 are expanded upon and used as inputs to REM to reproduce the MV network for all countries in Sub-Saharan Africa, then the results of this Chapter suggest REM will perform better than the Gridfinder tool and, the often incomplete, open databases. Under this assumption I rank the Data Types for the MV Network Layout Input Module in Figure 5.1.7.

Chapter 6

Summary and Conclusions

Today's race for global vaccination against COVID-19 has brought into focus the gaping disparity of energy infrastructure across the developing world. Vast populations remain vulnerable to the disease and others due to their lack of access to reliable electricity, and thus adequate healthcare and vaccine refrigeration. This undeniable hinderance has reinvigorated the UN-SDG7 commitment to achieving affordable, universal electrification as soon as possible. Evidence in literature has shown that achieving this goal is inextricably linked to the uptake and application of spatially explicit computational electrification planning tools (Korkovelos et al., 2019; Szabó et al., 2013; Taneja, 2019). The tools help energy stakeholders in decision-making and budget allocation by automating the delineation of electrification technologies (grid extension, mini-grids, and stand-alone solar systems) which extend to all customers at the lowest cost. However, the effectiveness of these tools is delimited by stakeholders' access to credible and up-to-date records on the local population and present infrastructure. My thesis aims to overcome these data acquisition bottlenecks and expedite universal electrification, by designing a scalable and modular framework on how to leverage publicly available information to produce granular estimates of the data needed.

To guarantee the framework is applicable to any planning tool, I chose to configure its design to the input requirements of the most data-demanding tool in the field, the Reference Electrification Model (REM). REM stands out among other planning tools because it provides system designs and least cost combination of electrification delivery modes at any geographic scale, while working at the individual customer level and respecting the physical laws, reliability, and constraints of power systems (Ciller et al., 2019). To achieve this level of granularity REM depends on extensive input data. Some of these inputs can already be imputed by peer-reviewed public datasets that already exist at the appropriate resolution for all countries in Sub-Saharan Africa (the developing countries subject to this thesis). The remaining inputs might be available through open data but just for a few countries or in coarse and incomplete formats; some might be completely unavailable publicly. These REM inputs are the subjects of my framework and are the following: 1) the geolocation and demand characteristics of residential customers, 2) the geolocation and demand characteristics of social and productive customers, 3) the electrification status of each customer, and 4) the layout of the medium-voltage distribution network. Each module of the framework represents one of these inputs (hereon referred to as Input Types). Each Input Type is then allowed multiple underlying data modules (hereon referred to as Data Types). The Data Types are a tiered list of available datasets per input, and account for scenarios where more accurate and granular data is available for some but not all regions. Users are encouraged to utilize the highest-tiered Data Type as it is available for their area of study, and substitute lowered-tiers Data Types otherwise. At times coarser datatypes might be sufficient for a high-level use-case of the framework and be subsequently easier to download and process. The flexible Data Type modules enables users to make these time-saving decisions as it pertains to their project context.

Since the objective is to expedite global electrification efforts, it is critical for the framework to survive the constantly changing paradigm of data collection and sharing. The modular design of the Input and Data Types ensures the framework can be iteratively updated as more sophisticated imputation methods and accurate public datasets are developed. Each Input module is prescribed a method, or Python model, to process and analyze the Data Types into reasonable estimates of the REM input. Defining what a "reasonable" estimate varied between modules but generally translated to a 10-30% margin of error between the output and the regional statistic it corresponded to. Figure 3.2.1 illustrates the inter-relationships between modules and the Python models. The source code of the Python-based framework is available through GitHub, with directional documents on how to run each module for any country (or region). By making it open-source, the framework is subject to the evaluation of modelers across the field and helps demonstrate adequate due-diligence to the stakeholders who might rely on the outputs for critical decision making. The remainder of this chapter will look separately at each Input Module.

A. Geolocation and Demand Characteristics of Residential Customers

I explore two Data Types for determining the explicit coordinates of buildings in a study region: built-area datasets and automated building footprint extraction models. Built area datasets use remote sensing techniques to accurately describe the extent, location, and characteristics of urban and built-up areas. The most promising dataset of this type was the World Settlement Footprint 2015 (WSF-2015); however, its collection method struggles to discern between congruous or irregular footprints and lacked ancillary information on the distribution of population across detected settlements. Automatic building footprint extraction models apply computer vision techniques to classify each pixel of a high-resolution satellite image as containing a building or not. Two datasets dominate this Data Type, Facebook's High-Resolution Settlement Layer (HRSL) and Google's Open Building Dataset. The HRSL produces binary labeled pixels at 30-m resolution overlaid with information on the population density within the respective census area. The Open Buildings dataset is superior to its competitor by defining the extents of heterogenous building footprints at a resolution of 50-cm, however, was unavailable at the time of this study.

My thesis leverages the HRSL dataset since it is more accessible than WSF-2015 and incorporates information on the density of the local population. Since the HRSL does not include meta-data on the size or building type of detected settlements I could not, with any statistical significance, classify the buildings as either residential or productive (larger loads associate with economic returns, e.g., markets or industrial facilities) customers. As a result, I make the overarching assumption that all settlements in the HRSL dataset are households, or 'smaller' loads. Similar assumptions were made by Fobi et al. (2021) and Korkovelos et al. (2019). Examples of small loads are local commercial businesses, village community centers, or household-operated services in rural or peri-urban areas. These loads usually share similar consumption profiles and are always fed by low-voltage lines. I also claim that the HRSL's census-level population density information closely approximates the population count in a pixel, and that the average size of a household remains the same across a country. Under these assumptions, I propose a method to approximate the number and coordinates of households within all 30-by-30 m pixels in the area-of-interest.

To test my approximations, I run a case-study on the country of Rwanda and compute my model's estimate of the total population to the country's most recently published data on its population. My results showed an overestimation of the total population by 15%, or by approximately 2 million residential customers. This might in part be due to assuming all the detected buildings are households, as well as not accounting for multiple settlements per household. The effects of this overestimation on the final output of REM are explored qualitatively in Chapter 4.1 and suggest that most of the overestimation occurs in denser, urban regions and the model is more likely to perform accurately in rural, sparser regions. I suggest future work should test the quality of the approximation by applying the model on all countries in Sub-Saharan Africa (SSA) and observe if there is a statistically significant trend of over- or under estimation. Moreover, segregation of a country into different land types (e.g., rural, peri-urban, or urban) and accordingly applying different values for the average household size might yield better estimates of the population in dense regions. Finally, Chapter 4.1 links to several studies in literature which describe alternative methods of estimating household locations from building footprint datasets. I suggest a formal comparison between how each of these methods perform and substituting the most accurate model with my prescribed estimation method in future iterations of the framework.

B. Geolocation and Demand Characteristics of Social and Productive Customers

The HRSL dataset approximates the location of residential and smaller loads which share similar consumption patterns. However, social (or community) and productive loads like health centers, education facilities, mines, or large industrial consumers have unique consumption patterns. These loads require larger and more reliable amounts of electricity and might need to be fed by medium or high voltage lines. As such, developing a method for characterizing the spatial variation of these customer types is necessary for REM (and other planning tools) to accurately evaluate the cost optimal electrification plan and distribution network layout for an area-of-study. I expand on REM's pre-defined broad customer archetypes to define 18 non-residential loads the following Input module should explore (refer to Figure 4.2.4). Several publicly available datasets exist for a small subset of these archetypes. The datasets are typically compiled manually or collected through surveys, by provision of non-governmental organizations or development agencies. A notable example is the World Health Organization's (WHO) geospatial inventory of 98,745 public health facilities in SSA. However, these public datasets only span a few load types and even fewer countries. Given that my objective is to produce an open-source and scalable method of estimating all REM inputs, the compiled datasets cannot suffice. To address this, I explore the novel use of the Google Maps API to extract the geolocation of different businesses, commercial services, and industries in any country.

Google Maps currently covers over 200 countries and territories and is updated 50 million times a day. The specificity and ubiquity of Google Maps makes it incredibly powerful at extracting information on a variety of customer types and their locations. I propose using Maps' *Places Nearby Search* API, whereby I specify a set of search parameters and a list of places that match the criteria are returned, with supplemental information like their operational status or latitude and longitude coordinates. While API requests are not free, I expect they are less costly than the time and resources needed to acquire the information through continuous, bureaucratic relationships with local stakeholders or surveyors. The *Places Nearby Search* tool works by looking for places that match a list of user-supplied keywords within a set radial area. The centroid of the area and its extent are determined by the user. I use the centroid coordinates of each 900-m^2 pixel in the HRSL's dataset to iteratively search through my entire area-of-study. Next, I assign each of the 18 non-residential customer archetypes a list of keywords extracted from Google Maps' own library of terms. The terms are used by the API to store and describe the different types of places registered in their repository (e.g., 'clinic' or 'city_hall'). Third, I develop a systematic methodology to infer the optimal search radius for each type of customer. The methodology recommends larger radii for less-frequently encountered customers, like a University of Industrial Factory, and a smaller radius for proliferate place types like markets or health centers. Altogether, my Python model iterates through each HRSL pixel and submits an API request for each customer archetype, one keyword at a time.

To test the accuracy and computational efficiency of the Google Maps API search, I conducted a small experiment comparing the well-documented WHO Public Health Facilities dataset with my model's results for an API search of "Health Center" customers. The experiment was conducted on the pixels delineating a small region of Kigali. The API was successful at locating the WHO facilities and more, returning the location of additional smaller health clinics and private practices in the area. This level of granularity on load-type-information is uncommon for publicly available data sources and should be taken advantage of. Nonetheless, the proposed model must undergo further testing to discern how practical and computationally efficient it is at large-scale use.

C. Electrification Status of Each Customer

Distinguishing between which of the customers are electrified is important for two reasons. First, the location and characteristics of electrified loads is an intermediary input in the implementation of REM for the estimation of the medium voltage distribution network (when this data is not available), described in more detail below. Second, mapping unelectrified loads is necessary for all electrification planning tools to prescribe the cost optimal path to universal energy. Beyond planning tools, accurately and remotely tracking the state of electricity access is a crucial means to monitoring the progress on the UN's seventh sustainable development goal. Despite its ubiquity to the energy access field, there are only few recorded approaches for the estimation of electrification status. Thus far the most employed proxy for electricity access is nighttime light data (NTL). NTL data provide global daily measurements of nocturnal visible and near infrared light, linked mostly to the presence and detection of streetlights. However, the data exhibit substantial noise and may not register any signature of electricity in deep rural areas because of extremely low levels of external lighting. Despite this, multiple studies have employed NTL intensities to asses residential energy consumption, detect power outages, monitor population migration, and income inequality (Albert et al., 2017; Elvidge et al., 2020; Falchetta et al., 2019). My thesis takes a similar stance and utilizes the NTL-based Falchetta et al. (2019) gridded dataset to estimate electricity access.

Falchetta et al. (2019) presents a 1-km^2 resolution gridded dataset in two formats: first as an annual count (for 2014 to 2018) of people without electricity per grid cell and second as a classification of people with electricity into one of four 'tiers of consumption'. The classification of electrified customers is determined by exceeding a minimum threshold value of NTL radiance and is therefore vulnerable to the limitations of NTL predictability of rural energy statuses.

Electrified customers are then ranked into 'tiers of consumption' based on the distribution of quartile values of non-zero light radiance across their cells. The quartiles are set separately for urban and rural cells to account for the strong discontinuity between the average radiances of each. Falchetta et al. (2019) broadly links the consumption tiers to the World Bank's *Multi-Tier Matrix for Measure of Household Electricity Consumption,* or successive thresholds of minimum energy demanded by a household in each tier (Bhatia & Angelou, 2015). I assign both residential and productive loads (from A and B, respectively) a consumption tier, simultaneously confirming their electrification status. I then map the World Bank's tiered minimum demand levels, or supply requirements, to the Falchetta consumption tiers. From which I assign each electrified residential load a unique peak-power-demanded value. I do so by assuming power consumption per tier is distributed normally and use the per-tier cut-offs to define the lower and upper limits of a 95% confidence level of the distribution. Non-residential loads are ascribed demand patterns from their archetype's profile, the specification of which is beyond the scope of this thesis and users should refer to supporting literature on archetypal behaviors of larger loads. The final output of this module is two DataFrames consisting of the location and demand characteristics (i.e., peak power demanded) of either electrified or unelectrified customers.

A small-scale experiment on the Rwanda households extracted via the HRSL dataset showed that my electrification status model underestimates the national electrification rate by 30%. This is a large margin of error and does call into question the efficacy of using NTLs as a single proxy for the existence of electricity. I mostly attribute this error to undetectability of rural areas from this type of data and suggest multiple machine learning or computer vision augmentations to the NTL dataset that should be explored in the next iteration of the framework.

D. Estimation of the Medium Voltage Network Layout

Characterizing the extent, geolocation, and voltage of the medium voltage (MV) distribution network is a critical and invariable component across all tools and frameworks introduced in Chapter 2. The existing information on MV network layouts can be broadly classified into two types. The first Data Type includes crowd-sourced information, aggregated data from project archives, or public datasets shared by countries or utilities. The completeness and quality of these datasets vary country to country and are often out of date, excluding new distribution lines or misrepresenting planned lines which have since changed. The second data type includes inferred maps of the network from existing open data, like remote sensing or satellite imagery. The most notable example, and the primary subject for comparison in my thesis, is the open tool *gridfinder. Gridfinder* is an open-source dataset that uses monthly and annual NTL data with other widely available infrastructure data (e.g., road networks) to produce the first composite map of the global power grid (Arderne et al., 2020). While the performance of *gridfinder* is better than the inconsistent quality of public datasets, it remains not good enough for individual building-level electrification plans (i.e., what REM requires). Arderne et al. (2020) designs the network based on geometric congruity with roads and OpenStreetMap grid data, no power flow or voltage drop analyses are made. Furthermore, by identifying electrification targets with only NTL data, the model is necessarily hindered by NTL's poor detection of light in rural or peri-urban areas. An unexplored method for estimating MV network designs is the Reference Network Model (RNM) built into REM. RNM designs the optimal distribution network to supply georeferenced customers with the electricity they demand at a prescribed level of hourly reliability (Domingo et al., 2011). I propose leveraging the design capabilities of RNM (accessed via REM) to replicate the existing network in a study region.

Estimating the MV network with REM can be done by configuring the input parameters to force an 'all-grid-extension' REM output. The parameters most critical to the network layout are the social cost of non-served energy (CNSE), the per-customer cost for off-grid alternatives, and the annual management costs for mini-grids. Together these parameters should make off-grid alternatives prohibitively expensive and produce large grid-extension clusters. To run the 'all-grid-extension' REM case, I supply it with the geolocations of (presumed) electrified residential and productive customers, their respective energy demand, and a user-provided country-specific network catalog. In addition, I would treat any openly available information on the existing grid as a-priori information on where the grid network is most likely built-out, feeding the open data as inputs to REM. Example of this a-priori data is the location and voltage of transformers or substations found on OpenStreetMap.

As a proof-of-concept I use ground-truth data provided by Rwanda stakeholders to see how the output of REM compares to the existing medium voltage network layout. To run this case, I also used the ground-truth data on electrified social and productive loads, as well as the location of medium voltage to low voltage distribution transformers. The location, demand, and electrification status of residential loads in Rwanda were supplied by my framework (i.e., the HRSL dataset). While it would be a more robust test of my framework to use *just* open sources as inputs to the case, the Google Maps API extraction of productive loads has yet to be scaled to all of Rwanda and therefore was unavailable for the experiment. Similarly, identifying transformer locations from public data is hindered by the same limitations as public sources on distribution layouts. Specifically, the transformer locations are incomplete and lack sufficient meta data on voltages. Nonetheless, my experiment showed very promising results. A qualitative comparison between the existing network, REM's outputs, and *gridfinder* showed that REM extended the grid to most areas the existing network was located, and outperformed gridfinder. Though the lines were not exactly overlaid or identical to Rwanda's layout, REM and other planning tools primarily use information on the existing grid to measure the feasibility of local grid extension, i.e., the proximity of unelectrified buildings to gridlines and the cost effectiveness of building out the grid to supply them with energy. So, I expect REM's output to be good enough for the scope of my framework.

That said, my next step will be to run a cell-by-cell comparison of all of Rwanda to compute how many true-positives and true-negatives are produced by REM's RNM design module. Only after which can I make a final assessment on the efficacy of my model and its advantage over gridfinder. Finally, as REM is a proprietary software, applying it to estimate the medium voltage distribution network does not constitute an open-access solution for all users of the framework. As such, the notion of how this method would fit within the GitHub open framework remains an ongoing discussion.

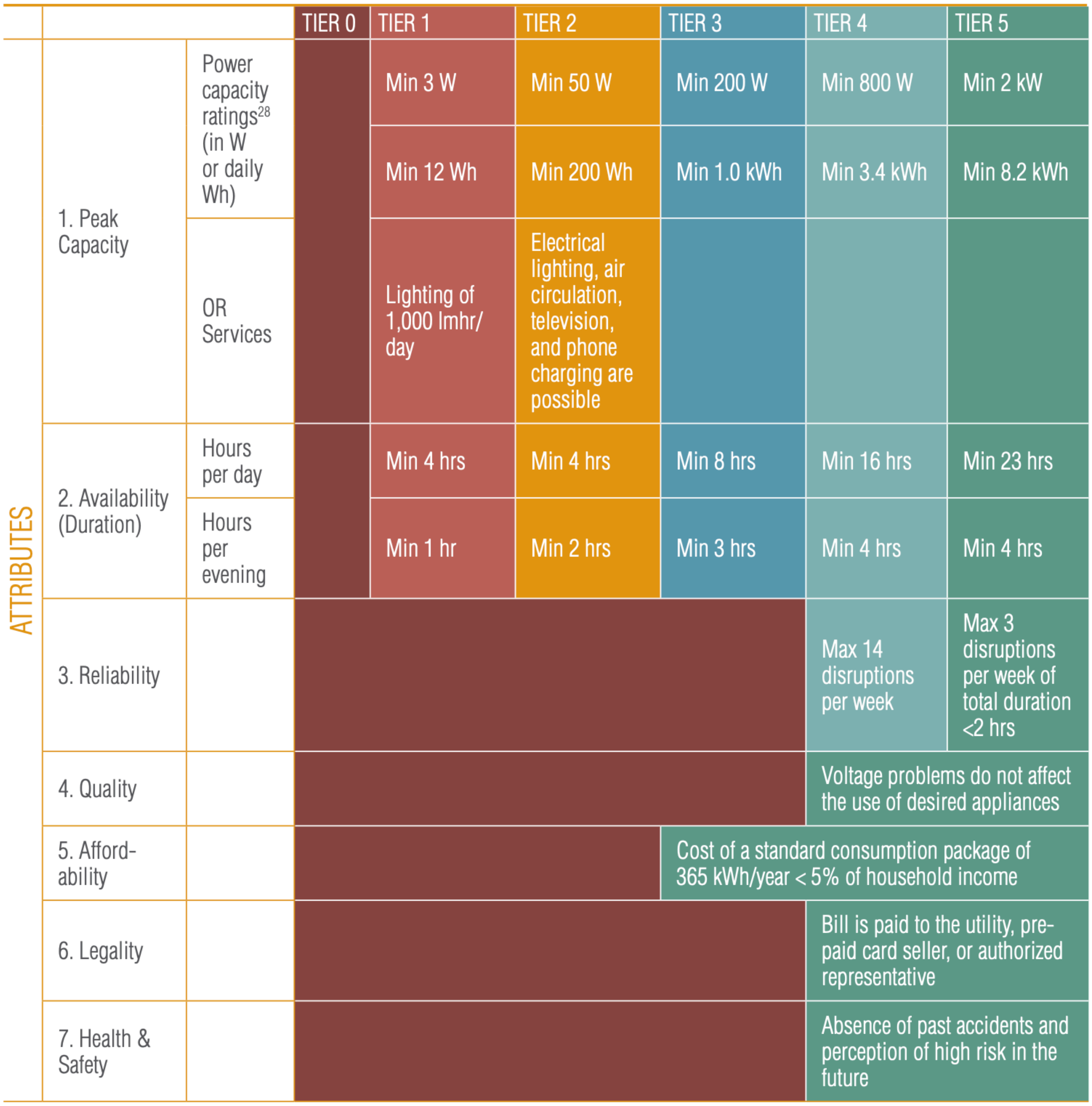
Final Reflections on the Framework

So far, all open-source frameworks in the field have needed to limit their models to coarser inferences, designing optimal electrification pathways at the village-, cell-, or administrative-level. These approaches forgo specificity for quicker implementations and higher levels of uncertainty. My thesis lays the foundation for a complete transition of the energy access sector to granular, open-source modelling. That said, ongoing work is still needed to completely automate the framework. A final iteration of the framework should require minimal engagement by users, except to define the extent of their study area. Furthermore, I have proposed multiple routes to developing more accurate models which estimate each Input Type. Future work should delve into each of these improvements and ongoing research to define the most "accurate" Python model. In the end, my thesis aims to first assist energy planners at fully electrifying the developing world, and next empower researchers with a tool to run synergistic studies on how electricity can change the socio-economic status and air-quality standards of these populations.

Appendix

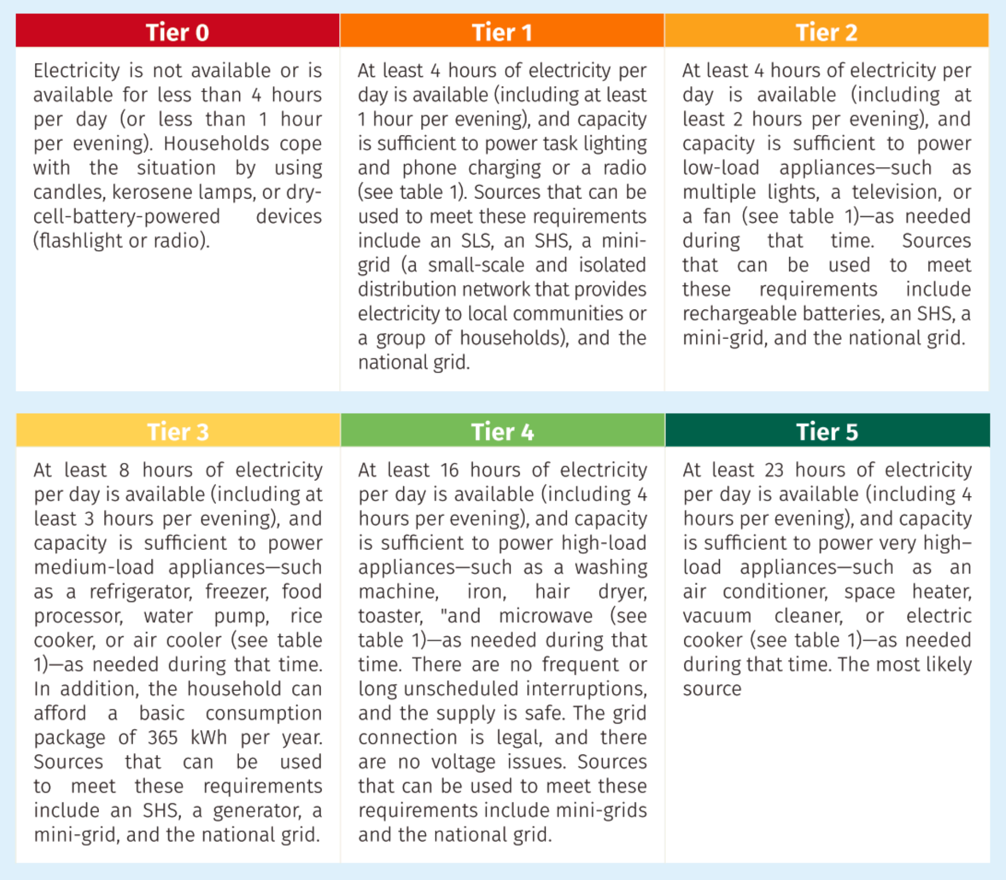
1. Supplementary Figures

Table A.1.1: (Bhatia & Angelou, 2015) Multi-Tier Matrix for Measuring Access to Household Electricity. Extracted from *Beyond Connections: Energy Access Redefined.*



			TIER 0	TIER 1	TIER 2	TIER 3	TIER 4	TIER 5
ATTRIBUTES	1. Peak Capacity	Power capacity ratings[28] (in W or daily Wh)		Min 3 W	Min 50 W	Min 200 W	Min 800 W	Min 2 kW
				Min 12 Wh	Min 200 Wh	Min 1.0 kWh	Min 3.4 kWh	Min 8.2 kWh
		OR Services		Lighting of 1,000 lmhr/ day	Electrical lighting, air circulation, television, and phone charging are possible			
	2. Availability (Duration)	Hours per day		Min 4 hrs	Min 4 hrs	Min 8 hrs	Min 16 hrs	Min 23 hrs
		Hours per evening		Min 1 hr	Min 2 hrs	Min 3 hrs	Min 4 hrs	Min 4 hrs
	3. Reliability						Max 14 disruptions per week	Max 3 disruptions per week of total duration <2 hrs
	4. Quality						Voltage problems do not affect the use of desired appliances	
	5. Affordability					Cost of a standard consumption package of 365 kWh/year < 5% of household income		
	6. Legality						Bill is paid to the utility, prepaid card seller, or authorized representative	
	7. Health & Safety						Absence of past accidents and perception of high risk in the future	

Figure A.1.2: (Bhatia & Angelou, 2015) Minimum Requirements by Tier of Electricity Access. Extracted from *Beyond Connections: Energy Access Redefined*.



Tier 0

Electricity is not available or is available for less than 4 hours per day (or less than 1 hour per evening). Households cope with the situation by using candles, kerosene lamps, or dry-cell-battery-powered devices (flashlight or radio).

Tier 1

At least 4 hours of electricity per day is available (including at least 1 hour per evening), and capacity is sufficient to power task lighting and phone charging or a radio (see table 1). Sources that can be used to meet these requirements include an SLS, an SHS, a mini-grid (a small-scale and isolated distribution network that provides electricity to local communities or a group of households), and the national grid.

Tier 2

At least 4 hours of electricity per day is available (including at least 2 hours per evening), and capacity is sufficient to power low-load appliances—such as multiple lights, a television, or a fan (see table 1)—as needed during that time. Sources that can be used to meet these requirements include rechargeable batteries, an SHS, a mini-grid, and the national grid.

Tier 3

At least 8 hours of electricity per day is available (including at least 3 hours per evening), and capacity is sufficient to power medium-load appliances—such as a refrigerator, freezer, food processor, water pump, rice cooker, or air cooler (see table 1)—as needed during that time. In addition, the household can afford a basic consumption package of 365 kWh per year. Sources that can be used to meet these requirements include an SHS, a generator, a mini-grid, and the national grid.

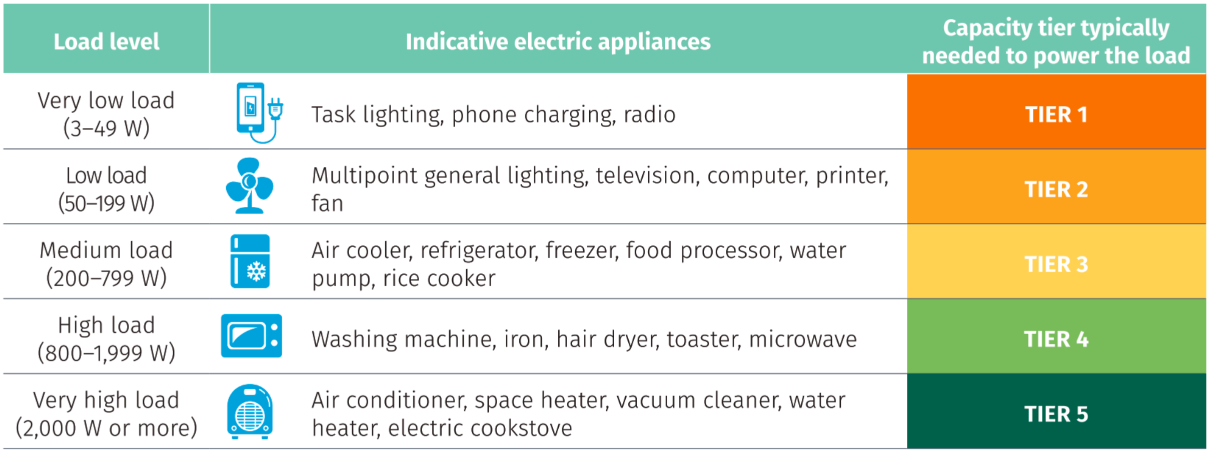
Tier 4

At least 16 hours of electricity per day is available (including 4 hours per evening), and capacity is sufficient to power high-load appliances—such as a washing machine, iron, hair dryer, toaster, "and microwave (see table 1)—as needed during that time. There are no frequent or long unscheduled interruptions, and the supply is safe. The grid connection is legal, and there are no voltage issues. Sources that can be used to meet these requirements include mini-grids and the national grid.

Tier 5

At least 23 hours of electricity per day is available (including 4 hours per evening), and capacity is sufficient to power very high–load appliances—such as an air conditioner, space heater, vacuum cleaner, or electric cooker (see table 1)—as needed during that time. The most likely source

Table A.1.3: (Bhatia & Angelou, 2015) Load levels, Indicative Electric Appliances, and Associated Capacity Tiers. Extracted from *Beyond Connections: Energy Access Redefined*.



Load level	Indicative electric appliances	Capacity tier typically needed to power the load
Very low load (3–49 W)	Task lighting, phone charging, radio	TIER 1
Low load (50–199 W)	Multipoint general lighting, television, computer, printer, fan	TIER 2
Medium load (200–799 W)	Air cooler, refrigerator, freezer, food processor, water pump, rice cooker	TIER 3
High load (800–1,999 W)	Washing machine, iron, hair dryer, toaster, microwave	TIER 4
Very high load (2,000 W or more)	Air conditioner, space heater, vacuum cleaner, water heater, electric cookstove	TIER 5

Bibliography

Adkins, J. E., Modi, V., Sherpa, S., Han, R., Kocaman, A. S., Zhao, N., Natali, C., & Carbajal, J. (2017). A geospatial framework for electrification planning in developing countries. *GHTC 2017 - IEEE Global Humanitarian Technology Conference, Proceedings*, *2017-Janua*, 1–10. https://doi.org/10.1109/GHTC.2017.8239293

Ahmed, A. I., Bryant, R. G., & Edwards, D. P. (2021). Where are mines located in sub Saharan Africa and how have they expanded overtime? *Land Degradation and Development*, *32*(1), 112–122. https://doi.org/10.1002/LDR.3706

Albert, A., Kaur, J., & Gonzalez, M. C. (2017). Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1296*, 1357–1366.

Alova, G., Trotter, P. A., & Money, A. (2021). A machine-learning approach to predicting Africa's electricity mix based on planned power plants and their chances of success. *Nature Energy 2021 6:2*, *6*(2), 158–166. https://doi.org/10.1038/s41560-020-00755-9

Amatya, R., Barbar, M., Borofsky, Y., Brusnahan, M., Ciller, P., Cotterman, T., de Cuadra, F., Drouin, C., Dueñas, P., Ellman, D., González- García, A., Lee, S., Li, V., Mateo, C., Oladeji, O., Palacios, R., Pérez-Arriaga, I., Stoner, R., & Vergara, C. (2018). Computer-aided electrification planning in developing countries: The Reference Electrification Model (REM). *Universal Energy Access Lab*, 1–111.https://www.iit.comillas.edu/publicacion/mostrar_publicacion_working_paper.php.en?id=347 %0AAuthors

Arderne, C., Zorn, C., Nicolas, C., & Koks, E. E. (2020). Predictive mapping of the global power system using open data. *Scientific Data*, *7*(1), 1–12. https://doi.org/10.1038/s41597-019-0347-4

Bazilian, M., Nussbaumer, P., Eibs-Singer, C., Brew-Hammond, A., Modi, V., Sovacool, B., Ramana, V., & Aqrawi, P. K. (2012). Improving Access to Modern Energy Services: Insights from Case Studies. *The Electricity Journal*, *25*(1), 93–114. https://doi.org/10.1016/J.TEJ.2012.01.007

Bertheau, P., Cader, C., & Blechinger, P. (2016). Electrification Modelling for Nigeria. *Energy Procedia*, *93*, 108–112. https://doi.org/10.1016/J.EGYPRO.2016.07.157

Bhatia, M., & Angelou, N. (2015). Beyond Connections Energy Access. *World Bank Group*, 228. https://openknowledge.worldbank.org/handle/10986/24368

Bonafilia, D., Gil, J., Kirsanov, D., & Sundram, J. (2019, April 9). *Mapping the world to help aid workers, with weakly, semi-supervised learning.* Facebook AI. https://ai.facebook.com/blog/mapping-the-world-to-help-aid-workers-with-weakly-semi-supervised-learning

Byers, L., Friedrich, J., Hennig, R., Kressig, A., McCormick, C., & Malaguzzi Valeri, L. (2021). A Global Database of Power Plants. *World Resources Institute*, *x*, 1–18. http://globalenergyobservatory.org\

Ciller, P., Ellman, D., Vergara, C., Gonzalez-Garcia, A., Lee, S. J., Drouin, C., Brusnahan, M., Borofsky, Y., Mateo, C., Amatya, R., Palacios, R., Stoner, R., De Cuadra, F., & Perez-Arriaga, I. (2019).

Optimal Electrification Planning Incorporating On- And Off-Grid Technologies- And Reference Electrification Model (REM). *Proceedings of the IEEE*, *107*(9), 1872–1905. https://doi.org/10.1109/JPROC.2019.2922543

Correa, S., Shah, Z., & Taneja, J. (2021). This Little Light of Mine: Electricity Access Mapping Using Night-time Light Data. *E-Energy 2021 - Proceedings of the 2021 12th ACM International Conference on Future Energy Systems*, 254–258. https://doi.org/10.1145/3447555.3464871

Domingo, C. M., Gómez, T., Román, S., Member, S., Sánchez-miralles, Á., Pascual, J., González, P., & Martínez, A. C. (2011). *A Reference Network Model for Large-Scale Street Map Generation*. *26*(1), 190–197.

Elvidge, C. D., Hsu, F.-C., Zhizhin, M., Ghosh, T., Taneja, J., & Bazilian, M. (2020). Indicators of Electric Power Instability from Satellite Observed Nighttime Lights. *Remote Sensing 2020, Vol. 12, Page 3194*, *12*(19), 3194. https://doi.org/10.3390/RS12193194

Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F.-C., & Taneja, J. (2021). Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. *Remote Sensing 2021, Vol. 13, Page 922*, *13*(5), 922. https://doi.org/10.3390/RS13050922

Energy Information Administration, U. (2021). *Levelized Costs of New Generation Resources in the Annual Energy Outlook 2021*.

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., Roth, A., Üreyen, S., Zeidler, J., Dech, S., & Gorelick, N. (2018). Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. *Remote Sensing 2018, Vol. 10, Page 895*, *10*(6), 895. https://doi.org/10.3390/RS10060895

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., & Strano, E. (2017). Breaking new ground in mapping human settlements from space – The Global Urban Footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, *134*, 30–42. https://doi.org/10.1016/J.ISPRSJPRS.2017.10.012

Fabini, D., Baridó, D. P. D. L., Omu, A., & Taneja, J. (2014). Mapping induced residential demand for electricity in Kenya. *ACM DEV 2014 - Proceedings of the 2014 Annual Symposium on Computing for Development*, 43–52. https://doi.org/10.1145/2674377.2674390

Falchetta, G., Pachauri, S., Parkinson, S., & Byers, E. (2019). A high-resolution gridded dataset to assess electrification in sub-Saharan Africa. *Scientific Data 2019 6:1*, *6*(1), 1–9. https://doi.org/10.1038/s41597-019-0122-6

Fobi, S., Kocaman, A. S., Taneja, J., & Modi, V. (2021). A scalable framework to measure the impact of spatial heterogeneity on electrification. *Energy for Sustainable Development*, *60*, 67–81. https://doi.org/10.1016/j.esd.2020.12.005

Fuso-Nerini, F., Broad, O., Mentis, D., Welsch, M., Bazilian, M., & Howells, M. (2016). A cost comparison of technology approaches for improving access to electricity services. *Energy*, *95*, 255–265. https://doi.org/10.1016/J.ENERGY.2015.11.068

Gershenson, D., Rohrer, B., & Lerner, A. (2019). *A New Predictive Model for Accurate Electrical Grid Mapping*. Facebook Engineering. https://engineering.fb.com/2019/01/25/connectivity/electrical-grid-mapping/

Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, Enipedia, & World Resources Institute. (2018). *Global Power Plant Database*. Resource Watch and Google Earth Engine. https://datasets.wri.org/dataset/globalpowerplantdatabase

IEA. (2014). *World Energy Outlook 2014* (IEA (ed.)). OECD. https://doi.org/10.1787/weo-2014-en

IEA. (2017). Energy Access Outlook 2017: From Poverty to Prosperity. In *Energy Access Outlook 2017*. https://doi.org/10.1787/9789264285569-en

IEA, IRENA, UNSD, World Bank, & WHO. (2021). Access to Electricity. In *Tracking SDG 7: The Energy Progress Report*. World Bank. www.worldbank.org

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/SCIENCE.AAF7894

Kaijuka, E. (2007). GIS and rural electricity planning in Uganda. *Journal of Cleaner Production*, *15*(2), 203–217. https://doi.org/10.1016/J.JCLEPRO.2005.11.057

Kappen, J. F. (2019). *Project Information Document-Integrated Safeguards Data Sheet-Madagascar-Least-Cost Electricity Access Development Project-LEAD-P163870*. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/281861547039951916/project-information-document-integrated-safeguards-data-sheet-madagascar-least-cost-electricity-access-development-project-lead-p163870

Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Laldaparsad, S., Friedl, L., Held, A., & Chuvieco, E. (2020). Towards delivering on the sustainable development goals using earth observations. In *Remote Sensing of Environment* (Vol. 247, p. 111930). Elsevier. https://doi.org/10.1016/j.rse.2020.111930

Kemausuor, F., Adkins, E., Adu-Poku, I., Brew-Hammond, A., & Modi, V. (2014). Electrification planning using Network Planner tool: The case of Ghana. *Energy for Sustainable Development*, *19*(1), 92–101. https://doi.org/10.1016/J.ESD.2013.12.009

Khavari, B., Korkovelos, A., Sahlberg, A., Howells, M., & Nerini, F. F. (2021). Population cluster data to assess the urban-rural split and electrification in Sub-Saharan Africa. *Scientific Data 2021 8:1*, *8*(1), 1–11. https://doi.org/10.1038/s41597-021-00897-9

Kivaisi, R. T. (2000). Installation and use of a 3 kWp PV plant at Umbuji village in Zanzibar. *Renewable Energy*, *19*(3), 457–472. https://doi.org/10.1016/S0960-1481(99)00053-1

Korkovelos, A., Bazilian, M., Mentis, D., & Howells, M. (2017). *A GIS Approach to Planning Electrification in Afghanistan*. https://energypedia.info/images/d/d6/A_GIS_approach_to_electrification_planning_in_Afghanistan.pdf

Korkovelos, A., Khavari, B., Sahlberg, A., Howells, M., & Arderne, C. (2019). The role of open access data in geospatial electrification planning and the achievement of SDG7. An onsset-based case study for Malawi. *Energies*, *12*(7), 1395. https://doi.org/10.3390/en12071395

Kuffer, M., Thomson, D. R., Boo, G., Mahabir, R., Grippa, T., Vanhuysse, S., Engstrom, R., Ndugwa, R., Makau, J., Darin, E., Albuquerque, J. P. de, & Kabaria, C. (2020). The Role of Earth Observation in an Integrated Deprived Area Mapping "System" for Low-to-Middle Income Countries. *Remote Sensing 2020, Vol. 12, Page 982*, *12*(6), 982. https://doi.org/10.3390/RS12060982

Lee, S. J. (2018). *Adaptive electricity access planning*. Massachusetts Institute of Technology.

Lee, S. J., Sánchez Jacob, E., González García, A., Ciller Cutillas, P., Dueñas Martínez, P., Taneja, J., Cuadra García, F., Lumbreras Martín, J., Daly, H., Stoner, R. J., & Pérez Arriaga, J. I. (2019). *Investigating the necessity of demand characterization and stimulation for geospatial electrification planning in developing countries*. https://repositorio.comillas.edu/xmlui/handle/11531/42497

Lee, S. J., & Taneja, J. (2020, February). Why Energy Demand Demands More Attention. *Energy for Growth Hub.*

Leyk, S., Gaughan, A. E., Adamo, S. B., De Sherbinin, A., Balk, D., Freire, S., Rose, A., Stevens, F. R., Blankespoor, B., Frye, C., Comenetz, J., Sorichetta, A., Macmanus, K., Pistolesi, L., Levy, M., Tatem, A. J., & Pesaresi, M. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, *11*(3), 1385–1409. https://doi.org/10.5194/ESSD-11-1385-2019

Li, X., Zhou, Y., Zhao, M., & Zhao, X. (2020). A harmonized global nighttime light dataset 1992–2018. *Scientific Data*, *7*(1), 1–9. https://doi.org/10.1038/s41597-020-0510-y

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. *PLOS ONE*, *7*(2), e31743. https://doi.org/10.1371/JOURNAL.PONE.0031743

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., & Tatem, A. J. (2019). Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data*, *3*(2), 108–139. https://doi.org/10.1080/20964471.2019.1625151

Louie, H., & Dauenhauer, P. (2016). Effects of load estimation error on small-scale off-grid photovoltaic system design, cost and reliability. *Energy for Sustainable Development*, *34*, 30–43. https://doi.org/10.1016/J.ESD.2016.08.002

Maina, J., Ouma, P. O., Macharia, P. M., Alegana, V. A., Mitto, B., Fall, I. S., Noor, A. M., Snow, R. W., & Okiro, E. A. (2019). A spatial database of health facilities managed by the public health sector in sub Saharan Africa. *Scientific Data*, *6*(1). https://doi.org/10.1038/S41597-019-0142-2

Mandelli, S., Brivio, C., Colombo, E., & Merlo, M. (2016). Effect of load profile uncertainty on the optimum sizing of off-grid PV systems for rural electrification. *Sustainable Energy Technologies and Assessments*, *18*, 34–47. https://doi.org/10.1016/J.SETA.2016.09.010

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., & Strano, E. (2020). Outlining where humans live, the World Settlement Footprint 2015. *Scientific Data 2020 7:1*, *7*(1), 1–14. https://doi.org/10.1038/s41597-020-00580-5

Mentis, D., Andersson, M., Howells, M., Rogner, H., Siyal, S., Broad, O., Korkovelos, A., & Bazilian, M. (2016). The benefits of geospatial planning in energy access – A case study on Ethiopia. *Applied Geography*, *72*, 1–13. https://doi.org/10.1016/J.APGEOG.2016.04.009

Mentis, D., Howells, M., Rogner, H., Korkovelos, A., Arderne, C., Zepeda, E., Siyal, S., Taliotis, C., Bazilian, M., De Roo, A., Tanvez, Y., Oudalov, A., & Scholtz, E. (2017). Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa. *Environmental Research Letters*, *12*(8). https://doi.org/10.1088/1748-9326/aa7b29

Mentis, D., Odarno, L., Wood, D., Jendle, F., Mazur, E., Qehaja, A., & Gassert, F. (2019). Energy Access Explorer: Data and Methods. In *Technical Note* (Issue September). www.wri.org/publication/energy-access-explorer.

Mentis, D., Welsch, M., Fuso Nerini, F., Broad, O., Howells, M., Bazilian, M., & Rogner, H. (2015). A GIS-based approach for electrification planning-A case study on Nigeria. *Energy for Sustainable Development*, *29*, 142–150. https://doi.org/10.1016/j.esd.2015.09.007

Modi, V., Adkins, E., Carbajal, J., & Zhao, N. (2013). *Liberia Power Sector Capacity Building and Energy Master Planning Final Report, Phase 4: National Electrification Master Plan.*

Moksnes, N., Korkovelos, A., Mentis, D., & Howells, M. (2017). Electrification pathways for Kenya–linking spatial electrification analysis and medium to long term energy planning. *Environmental Research Letters*, *12*(9), 095008. https://doi.org/10.1088/1748-9326/AA7E18

Morrissey, J. (2019). Achieving universal electricity access at the lowest cost: A comparison of published model results. *Energy for Sustainable Development*, *53*(November), 81–96. https://doi.org/10.1016/j.esd.2019.09.005

Ohiare, S. (2015). Expanding electricity access to all in Nigeria: a spatial planning and cost analysis. *Energy, Sustainability and Society 2015 5:1*, *5*(1), 1–18. https://doi.org/10.1186/S13705-015-0037-9

Ouzounis, G. K., Syrris, V., & Pesaresi, M. (2013). Multiscale quality assessment of global human settlement layer scenes against reference data using statistical learning. *Pattern Recognition Letters*, *34*(14), 1636–1647. https://doi.org/10.1016/j.patrec.2013.04.004

Palacios-Lopez, D., Bachofer, F., Esch, T., Marconcini, M., MacManus, K., Sorichetta, A., Zeidler, J., Dech, S., Tatem, A. J., & Reinartz, P. (2021). High-Resolution Gridded Population Datasets: Exploring the Capabilities of the World Settlement Footprint 2019 Imperviousness Layer for the African Continent. *Remote Sensing 2021, Vol. 13, Page 1142*, *13*(6), 1142. https://doi.org/10.3390/RS13061142

Parshall, L., Pillai, D., Mohan, S., Sanoh, A., & Modi, V. (2009). National electricity planning in settings with low pre-existing grid coverage: Development of a spatial model and case study of Kenya. *Energy Policy*, *37*(6), 2395–2410. https://doi.org/10.1016/J.ENPOL.2009.01.021

Peña Balderrama, J. G., Balderrama Subieta, S., Lombardi, F., Stevanato, N., Sahlberg, A., Howells, M., Colombo, E., & Quoilin, S. (2020). Incorporating high-resolution demand and techno-economic optimization to evaluate micro-grids into the Open Source Spatial Electrification Tool (OnSSET). *Energy for Sustainable Development*, *56*, 98–118. https://doi.org/10.1016/J.ESD.2020.02.009

Pittman, K., Hansen, M. C., Becker-Reshef, I., Potapov, P. V., & Justice, C. O. (2010). Estimating Global Cropland Extent with Multi-year MODIS Data. *Remote Sensing 2010, Vol. 2, Pages 1844-1863*, *2*(7), 1844–1863. https://doi.org/10.3390/RS2071844

Qiu, Y., Zhao, X., Fan, D., & Li, S. (2019). Geospatial Disaggregation of Population Data in Supporting SDG Assessments: A Case Study from Deqing County, China. *ISPRS International Journal of Geo-Information 2019, Vol. 8, Page 356*, *8*(8), 356. https://doi.org/10.3390/IJGI8080356

Reed, F. J., Gaughan, A. E., Stevens, F. R., Yetman, G., Sorichetta, A., & Tatem, A. J. (2018). Gridded Population Maps Informed by Different Built Settlement Products. *Data 2018, Vol. 3, Page 33*, *3*(3), 33. https://doi.org/10.3390/DATA3030033

René Saameli, Dikolela Kalubi, Mark Herringer, Tim Sutton, & Eric de Roodenbeke. (2016). Healthsites.io: The Global Healthsites Mapping Project . In *UNESCO Chair Conference on Technologies for Development* (pp. 53–59). Springer.

Riva, F., Gardumi, F., Tognollo, A., & Colombo, E. (2019). Soft-linking energy demand and optimisation models for local long-term electricity planning: An application to rural India. *Energy*, *166*, 32–46. https://doi.org/10.1016/j.energy.2018.10.067

Sanoh, A., Parshall, L., Sarr, O. F., Kum, S., & Modi, V. (2012). Local and national electricity planning in Senegal: Scenarios and policies. *Energy for Sustainable Development*, *16*(1), 13–25. https://doi.org/10.1016/J.ESD.2011.12.005

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). *Continental-Scale Building Detection from High Resolution Satellite Imagery*. https://arxiv.org/abs/2107.12283v1

Szabó, S., Bódis, K., Huld, T., & Moner-Girona, M. (2011). Energy solutions in rural Africa: mapping electrification costs of distributed solar and dieselgeneration versus grid extension*. *Environmental Research Letters*, *6*(3), 034002. https://doi.org/10.1088/1748-9326/6/3/034002

Taneja, J. (2019). If You Build It, Will They Consume? Key Challenges for Universal, Reliable, and Low-Cost Electricity Delivery in Kenya. *SSRN Electronic Journal*, *July 2018*. https://doi.org/10.2139/ssrn.3310479

The World Bank. (n.d.). *Electrification Paths for Nigeria, Tanzania and Zambia*. Retrieved August 30, 2021, from http://electrification.energydata.info/presentation/

The World Bank. (2019). *Africa - Electricity Transmission and Distribution Grid Map | Data Catalog*. https://datacatalog.worldbank.org/dataset/africa-electricity-transmission-and-distribution-grid-map-2017

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E. B., & Dang, H.-A. H. (2017a). Mapping the world population one building at a time. *Mapping the World Population One Building at a Time*. https://arxiv.org/abs/1712.05839v1

United Nations. (2017). *Big Data for Sustainable Development*. Global Pulse. https://www.un.org/en/global-issues/big-data-for-sustainable-development

Williams, N. J., Jaramillo, P., Cornell, B., Lyons-Galante, I., & Wynn, E. (2017). Load characteristics of East African microgrids. *Proceedings - 2017 IEEE PES-IAS PowerAfrica Conference: Harnessing Energy, Information and Communications Technology (ICT) for Affordable Electrification of Africa, PowerAfrica 2017*, 236–241. https://doi.org/10.1109/PowerAfrica.2017.7991230

World Bank. (2019). *Chad - Global Electrification Platform (GEP)*. Energydata.Info. https://energydata.info/dataset/chad-global-electrification-platform-gep

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2015). Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 3929–3935. https://arxiv.org/abs/1510.00098v2

Xiong, J., Thenkabail, P. S., Tilton, J. C., Gumma, M. K., Teluguntla, P., Oliphant, A., Congalton, R. G., Yadav, K., & Gorelick, N. (2017). Nominal 30-m Cropland Extent Map of Continental Africa by Integrating Pixel-Based and Object-Based Algorithms Using Sentinel-2 and Landsat-8 Data on Google Earth Engine. *Remote Sensing 2017, Vol. 9, Page 1065*, *9*(10), 1065. https://doi.org/10.3390/RS9101065

Zeyringer, M., Pachauri, S., Schmid, E., Schmidt, J., Worrell, E., & Morawetz, U. B. (2015). Analyzing grid extension and stand-alone photovoltaic systems for the cost-effective electrification of Kenya. *Energy for Sustainable Development*, *25*, 75–86. https://doi.org/10.1016/J.ESD.2015.01.003