Data-Driven Decision Support for Low Electricity Access Settings.

Sally Simone Rose Flore Lylie Fobi Nsutezo

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Data-Driven Decision Support for Low Electricity Access Settings

Sally Simone Rose Flore Lylie Fobi Nsutezo

Universal, affordable and reliable electricity remains a key pillar towards achieving Sustainable Development Goals. It is low income countries that find bridging gaps in electricity access particularly challenging. Making judicious financial investments is critical in a low income setting as there are multiple competing compelling areas in which to make resource allocations. A data driven approach that can leverage prior data from electricity service providers can guide decision making.

This dissertation presents approaches that leverage such data, to assist utilities and national bodies with insights that could be useful. There are five unique contributions made. These are in the form of key results about electricity consumption patterns, novel methodologies for electricity demand prediction and relevant metrics for estimating the cost of a grid connection.

First, this thesis, through in-depth analysis of electricity data from thousands of households, sheds light on electricity consumption patterns in Rwanda and Kenya. This work revealed that utilities are increasingly connecting low consuming households whose consumption peaks sooner and plateaus lower than their peers who were connected earlier. While the previous focus of research has been on addressing electricity supply-side constraints, this work is the first of it's kind to show that electricity consumption for the newly electrified is very low, thereby making capital cost recovery of a grid connection even harder to achieve. This mismatch between supply and demand emphasizes the need for utilities to better quantify expected demand upon connection.

Secondly, this thesis makes methodological contributions that support electricity demand prediction for the yet-to-be grid-connected households. Specifically, Convolutional Neural Network (CNN) models were designed to take as inputs pre-grid-access daytime satellite image patches and output electricity consumption levels. Results from this work show that the proposed methodologies perform better than utility based estimates of anticipated demand. This methodology shows that rapid large scale evaluation of latent demand can be effectively performed using daytime satel-

lite imagery, thereby giving guidance on which sites or regions are more suitable for grid versus off-grid technologies. Outputs from the models have been utilized by energy planners in Kenya.

The third unique contribution made in this dissertation is in the development of key metrics to estimate the cost of grid-access. Complementary to the evaluation of electricity demand, this thesis also develops an electricity grid network optimization model, connecting 9.2 million structures in Kenya. Given transformer placement and the estimates for low and medium voltage line, an approximation for the per household wire requirement is obtained. The work shows that traditional rural/urban classification based on population density may not be enough and is often deceiving in estimating the cost of grid-access and a new categorization based on our proposed per household wire requirement metrics provides more relevant estimates on the total cost.

Fourthly, this dissertation also demonstrates methods to re-purpose electricity data in order to provide insights to new domains such as household wealth. This work illustrates how household overall expenditure can be obtained from electricity usage data and how electricity usage can be obtained from daytime satellite imagery. This methodological contribution provides a pathway for stakeholders to estimate household overall expenditure from daytime satellite imagery. The work shows the value of electricity data in answering other questions in new domains without the deployment of additional surveys or hardware.

The final research contribution discussed in this thesis focuses on methods to make smart modifications to existing machine learning models to support analysis in settings where label availability is small and label quality is poor. This concept is illustrated with a building segmentation task given misaligned and omitted building footprints. Our proposed end-to-end learning pipeline demonstrates how data constrained regions can learn about building characteristics despite having incomplete and noisy labels. In addition, this work is used to provide explanatory features to the CNNs used for prediction in the earlier parts of the work.

While the focus of the research was on Kenya and Rwanda, this work transcends multiple domains such as water and internet access and can be extending to countries seeking evidence-based approaches to inform sustainable development.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I am extremely grateful to God, who has abundantly blessed me with opportunities to pursue my dreams and has strategically placed the right people along my PhD journey to support that quest.

My PhD journey has been a time of growth as a researcher, learning how to scope new questions and developing them into full scale research ideas and projects. This growth would not have been possible without the stimulating research environment created by my primary advisor Vijay Modi. I want to thank my advisor for challenging me to develop methodologically robust work, while also prioritizing challenges and solutions that are useful for sustainable development.

Throughout this journey, multiple others have poured immensely into my research career starting off with Jay Taneja. I would like to express my sincere gratitude to Jay who has been a personal mentor and friend, who has helped me navigate tough research question and cleared pathways to developing innovative solutions.

Special thanks to Terry Conlon, Bob Muhwezi, Joel Mugyenyi, Kiki Civian, Zeal Shah, Aggrey Muhebwa. Thanks for joining me to work on topics, many of which were not clear from the start.

Thanks to the Rockerfeller Foundation & to Columbia University for funding my PhD. Your support was critical to giving me access to a top-notch education and to allowing me pursue my research interests.

A very very special thank you to my parents, Grace and Simon Fobi, who instilled and nurtured my curious mind, always made my dreams seem achievable and who never failed to encourage and support my goals. To my siblings, thanks for always being available when ever I needed you.

To my best friend and husband, Iretiayo Akinola who always serves as a sounding board for all my pursuits. Thanks for your patience, feedback and for always helping to bring my goals to fruition. I trust we will have many more pursuits together.

Thanks to my lifelong friends Ngebi Fobi, Atinuke Ademola-Idowu, Faith Amadu, Funmi Ekundayo, Shubuka Mainsah and Mildred Narh. Our sisterhood was a much needed breath of fresh air during the tough times.

Dedication

To my parents (Grace and Simon Fobi), my husband (Iretiayo Akinola) and my daughter Tomi who will do above and exceedingly more than that which was begun in this work.

Preface

Every year, billions of dollars are invested in varying services including but not limited to water, broadband internet, agriculture and energy. In 2021, annual global energy investments were expected to hit $1.9 trillion, where approximately 29 % of that investment ($544 billion) was estimated to go to energy infrastructure [1]. For countries yet to attain 100 % electrification, a significant portion of the energy infrastructure budget goes towards increasing electricity access through grid extension or off-grid systems. However, in 2019, finance for electricity access dropped to $12.9 billion from $32 billion in 2018, while an estimated $41 billion in investments is needed annually to meet universal access by 2030 [2]. Given the large gulf between required and actual investments, approaches that support optimized resource allocation can enable electrification of more people using the same limited resources. In the case of electrification, some opportunities for optimal resource allocation include asset placement such as where the grid should be extended to versus which places are better off with off-grid systems, given the anticipated consumption. Poor allocation of already limited funds implies that fewer people can use electricity given the same investment. Of equal importance to optimized resource allocation, is quantifying the socio-economic impacts of these investments to ensure the intended economic growth is achieved. Data from the International Energy Agency has established the relationship between electricity usage and economic growth, where higher incomes and economic activity are usually correlated with more consumption of electricity. Approaches that quantify both the amount of electricity used given investments and the corresponding socio-economic activity are needed to justify the investment and for finding opportunities for improved resource allocation.

Advances in artificial intelligence and deep learning have improved the ability of computers to find optimal solutions. In recent years, these deep learning optimization algorithms have become ubiquitous and can be easily developed and used. The goal of deep learning models is to uncover patterns given the data by minimizing the error between the model and the data. Once a high performing model is established, the model can be used to infer the behavior for new and unseen examples. A key metric for deep learning models is their ability to generalize to these new and unseen examples. This ability to generalize, makes deep learning approaches very desirable within the energy context as energy providers can use their existing data to learn relevant patterns in electricity consumption and use insights from the extracted patterns to answer questions across the large swath of future customers, thereby better allocating their limited resources.

Critical to the performance of deep learning models is the availability of large amounts of model input data that capture the underlying dynamics of the problem of interest and corresponding groundtruth labels to confirm the predictions from the model. To that effect, the availability of remote sensed input data from varying satellites has vastly improved the ability of deep learning algorithms to answer questions around wealth, infrastructure and more, at the sub-national level. High resolution 50 cm visible 3-band imagery products are available through Maxar though these products tend to have low temporal cadence on the scale of every few years. Medium resolution 10m 13-band imagery such as free Sentinel2 products are collected at a higher temporal cadence of every 10-days. While lower resolution (15 arcseconds) products such as VIIRS Nighttime Lights are available on NOAA's website for every single day. The variety and volume of remote sensed products provides an opportunity to leverage deep learning algorithms in order to answer energy related questions at scale for multiple countries still seeking to offer affordable and reliable electricity [3].

This five chapter thesis presents a convergence of deep learning and remote sensed data applied to the electricity sector. In this body of work, I highlight data-driven approaches to understand electricity usage and growth, and methods for supporting decision making within the electricity sector.

Chapter 1: A Data-Driven Descriptive Analysis of Electricity Consumption & Growth in Kenya and Rwanda - presents an analysis of electricity usage and growth using electricity consumption data from Kenya and Rwanda. This data was obtained from the main providers of electricity covering over 100,000 utility residential customers in each country. In this chapter, electricity consumption patterns are teased out revealing how consumption evolves upon grid-access. From this analysis, a key observation emerged: as more people are electrified, utilities are increasingly adding households that consume less electricity, peak sooner, plateau lower and are more difficult to connect via grid. This insight is relevant to energy providers as they can better understand the kinds of technologies that can support the seen growth and also quantify the impact of their investment in electricity access. An analysis of utility revenue is also presented in this chapter with the aim of understanding the utility's policies and its impact on electricity consumption.

Given a better understanding of longitudinal electricity consumption amongst varying customer cohorts, ***Chapter 2: Predicting Levels of Household Electricity Consumption in Low-Access Settings*** - presents a method of predicting levels of electricity consumption for future customers given features present in satellite imagery. This chapter demonstrates that satellite imagery obtained prior to a building being electrified holds relevant information about the expected consumption that the household will have within the first few years of an electricity connection. In this chapter a Convolutional Neural Network (CNN) is developed to learn relevant household characteristics from images and the learnt features are used to predict levels of household electricity consumption in Kenya. The CNN model is validated with utility data from Kenya. Results from this work shows that using non-linear image-based models such as CNNs provides a better approach to estimating levels of consumption compared to typical methods that an energy provider might rely on. Predictions from the proposed models have been used by energy providers (through an API) to support electricity access planning in Kenya.

Having provided an approach to predicting consumption levels for future connections, ***Chapter 3: A scalable framework to measure the impact of spatial heterogeneity on electrification*** - presents a methodology to plan large scale grid-extension networks at the resolution of individual

buildings. Network planning with millions of household or structure nodes remains an NP-hard problem. However network planning is critical for energy providers to estimate the cost of a grid connection. In this chapter, we present a methodology to planning a grid extension network for over 9 million nodes. The nodes reflect locations of buildings obtained via satellite imagery. This chapter demonstrates how household settlement patterns can heavily influence the cost of a grid connection and where opportunities for different electrification technologies lay within the landscape of Kenya. One such observation from this work is that regions where households are close to each other but communities are far away from each other, tend to have higher medium voltage line costs and smaller low voltage line costs. These regions become good candidates for local generation and distribution (e.g. MiniGrid). This work is complementary to the electricity consumption prediction problem and methodology discussed in Chapter 2. Combining an understanding of grid connection costs with expected consumption can help energy planners better determine how to allocate resources to meet the goal of universal electricity access.

The remaining two chapters present approaches for remote measurement of other indicators such as household wealth and household characteristics. Electricity usage is often correlated with measurements of wealth and economic development. Energy provides have access to large amounts of passively collected electricity usage data from their already electrified customers.

Chapter 4: High resolution estimates of household electricity usage as a proxy for household overall expenditure - explores the effectiveness of combining electricity usage data and high resolution 50 cm daytime satellite imagery in Rwanda to estimate overall household expenditure. Here, we show that electricity consumption data can be repurposed to estimate other indicators such as wealth, without the deployment of additional resources such as surveys. This chapter also presents an approach to estimating actual kiloWatthours (kWh) of household electricity consumption from daytime imagery. These estimates are useful in understanding the value of electricity investments and also providing business insights for future investments. In addition to kWh estimates, this chapter explores the advantages of varying estimation metrics at different levels of aggregation, highlighting the implications on performance and privacy. Results from this work show that at 1

sqkm resolution, the model estimates with high fidelity, the average electricity consumption, while preserving the privacy of households. This chapter provides an approach for governments to evaluate the impact of varying investments on the socio-economic well-being of its people, where a specific resolution that best supports effective decision making can be selected.

Machine learning models learn very well in the presence of large amounts of clean labels. Building footprints are an example of labels that are important to understanding household characteristics and that lend themselves well to machine learning models. However, in resource constraint settings, label availability and cleanliness remains a challenge. ***Chapter 5: Learning to Segment from misaligned and partial labels***- presents a method to leverage machine learning for building roof segmentation, when labels are noisy and incomplete. This chapter demonstrates that relevant machine learning models can be developed by making smart modifications to existing machine learning approaches, thereby making the algorithms suitable for the relevant context. In this chapter, we first present an approach to realign misaligned building footprint labels. Next we demonstrate how a machine learning model can be trained when there are omitted building instances within an image patch. Combining both approaches, we show that building footprints can be obtained given noisy and incomplete labels. Beyond the usefulness of the work in learning with noisy labels, outputs from this chapter (building footprints) were used to provide insights to the CNN models presented in Chapter 2 and 4, highlighting the relevant features learnt during model training.

On the whole, this thesis applies deep learning and geospatial analysis to the energy domain to answer resource allocation questions and remote monitoring of socio-economic indicators. This work is particularly relevant for resource-constraint regions who could benefit from sub-national high resolution guidance about how to invest and the impact of their investments. While this work demonstrates the approach within the domain of electricity access and usage, the developed and proposed methods are of relevance to other sectors that may also require decision support.

Chapter 1: A Data-Driven Descriptive Analysis of Electricity Consumption & Growth in Kenya and Rwanda

Access to reliable and affordable electricity is a primary goal for policymakers, governments and development organizations. Developing economies regularly make critical decisions on how to allocate precious public-sector resources to increase electricity access, often with little evidence. Governments, financial institutions, and entrepreneurs are exploring new pathways for electrification such as solar home systems and mini-grids, as well as redoubling investments in traditional grid extension, all in an effort to build sustainable institutions for delivering electricity services. Despite these efforts, as of 2019 nearly 800 million people still live without access to electricity [4]. This estimate could potentially be higher as the Covid-19 pandemic has slowed progress toward universal electricity access [5]. Thus, electricity access remains a primary pillar for sustained growth and development.

Providing electricity access is an age-old challenge that multiple countries have and continue to address. In the 1930s, the United States saw a huge push to boost electrification rates (especially in rural areas) through programs led by the Rural Electrification Administration. This initiative sought to provide both electricity connections and appliances for farmers in rural areas. While the cost of rural electrification was estimated to be very high (sometimes up to $2000 /mile to extend power lines), appliance financing and credit extensions to farmers equally ensured high demand for power upon electrification. This combination made the rural electrification story in the United States a huge success and has served as an example of electrification for the rest of the world.[6]

Unlike the U.S. rural electrification program, government-led electrification initiatives such as those in India and throughout Sub-Saharan Africa have directly targeted households for grid connections instead of rural farmers. These electrification schemes spear-headed by country-led

agencies have emerged across the world with the goal of fast-tracking electrification. In India, the Rajiv Gandhu Grameen Vidyutikaran Yojana (RGGVY) scheme is an example of a government-led initiatives geared at increasing electricity connections at small to no connection fees. In Kenya, the Last Mile Connectivity Program (LMCP) and Rural Electrification Authority (REA) programs also aim to electrify households and rural public facilities, respectively[7]. In Rwanda, large scale electricity access initiatives such as Electricity Access Rollout Program (EARP) have significantly increased electricity access in households.

Despite the large strides made by these initiatives (sometimes at no cost to households) to increase grid connections, there remains a high capital cost to the utility or energy provider. Lee et al [8] estimate the cost of an electricity connection may range from $1300 to $1600 for households in Kenya. Similarly [9] estimate the rural per connection grid connection cost to be $1100 in Vietnam, $2300 in Tanzania, while those of urban grid connections could be $570 in Vietnam, $1100 in Tanzania and $800 in South Africa. As utilities electrify more households, an increasing proportion of these new connections would stem from rural areas, thereby leading to steeper per household connection costs for the utility. Assuming an average rural connection cost of $1,500 per connection, at current U.S. average household electricity consumptions of 900 kWh/month [10], it would take about 10 years to recoup the connection cost if 10 % of the electricity tariff were dedicated to capital cost recovery [1]. As electricity demand decreases the potential for cost recovery becomes significantly harder. At very low electricity consumptions of around 5 kWh/month, as seen in Rwanda, it would take 2,874 years to recoup the investment cost [2],. At such low consumptions, the economics of grid extension becomes infeasible and other cheaper electricity provision technologies have to be considered.

While grid extension efforts in developing economies have led to an up-tick in the percentage of population that has access to electricity at home; a less well-understood phenomena is the evolution of consumption as the utility increasingly electrifies more rural households. Understanding electricity consumption growth especially for more recently electrified households is critical

[1] Assuming the residential electricity tariff of 13.7 cents/kWh [11]

[2] Assuming a residential electricity tariff in 8.7 cents/kWh, with 10 % of the tariff allocated to cost recovery

for cost-effective planning. However, projecting future electricity consumption is difficult, underscored by the observation that projections tend to understate growth in electricity demand in the developing world [12]. Plausible electrification strategies depend on analyzing existing customer data to predict the behavior of varying customer cohorts such as newly-connected customers.

This chapter presents an in-depth analysis of electricity consumption data for utility based customers in Kenya and Rwanda. Using data from electric meters and analyzing over 100,000 customers in each country this chapter presents a descriptive analysis of electricity consumption patterns for varying connection cohorts. A key finding that emerges is that utilities are increasingly connecting lower consumers who peak sooner and plateau lower, thereby making it harder to recover the costs of grid extension. This finding underscores the importance of consumption growth studies which the energy providers must consider in order to cost-effectively electrify households.

1.1 Related Work

1.1.1 Electricity consumption

Accurate electricity consumption estimates are important in designing electrical generation and delivery systems and meeting reliability requirements. A study in Malawi [13] uses off-grid data from 7 PV and battery systems to show the impact of incorrect load estimation on system cost and reliability. They found that system cost scaled proportionally with errors in consumption estimates, where over estimation led to significant increases in system cost of between USD 1.82 to USD 6.02 per watt-hour, while underestimating consumption eroded system reliability. This dichotomy between system cost and reliability emphasizes the need for data-driven approaches to understanding and predicting consumption, which can in turn yield more optimal system design. In the case of residential electricity consumption, predictions are typically made by using multiple variables including socioeconomic characteristics, appliance ownership, and living conditions. A literature review on the topic suggests that at least 62 variables potentially affect residential electricity usage[14]. Other authors conclude that some important explanatory variables for household electricity consumption include appliance ownership, electricity tariffs, available income, and

number of residents in the household [15, 16, 17]. While these analyses offer a deep-dive into electricity consumption patterns, they depend on expensive and time-consuming household surveys, rendering them difficult to scale with similar resolution to larger areas such as countries or regions. Spatio-temporal analysis can provide insights to electricity consumption over large areas. Socio-economic and demographic variables such as population and income levels can be folded into such methods when studying electricity demand. For example, [18, 19, 20] demonstrate spatio-temporal analyses using satellite imagery to study population and energy dynamics in various regions. Results from these papers show a relationship between spatial dynamics, electricity consumption, and population. To explore the differences in electricity consumption due to urbanization, [19] use a pixel-based method to delineate urban, suburban and rural regions in China. A universal definition for urban regions was difficult to obtain and the Chinese administrative units "prefectural city" are a mix of both urban districts and rural counties. The authors use population adjusted nighttime lights to delineate urban areas. Land cover was then used to determine the optimal nighttime lights threshold for highly dense built-up areas in China. The obtained highly dense regions are labeled as the urban core while the difference between urban regions and urban core gives the suburban region. This definition of urbanization allows them to study differences in electricity patterns by urbanization levels. Chévez et al.[21] propose another approach for obtaining spatially homogeneous areas using k-means clustering algorithm. In this case, rather than using urban, suburban and rural as homogeneous areas, they define k clusters, where each cluster is a spatially homogeneous region. Homogeneity is defined by the authors as regions with similar electricity consumption. The algorithm classifies n users with M features into the k clusters. Given the number of clusters (k) defined a priori, the algorithm finds k clusters which minimize the euclidean distance as defined by sum of least squares. Initially, $k \times M$ values are chosen to represent cluster centroids. The authors compute the euclidean distance of each user from the initial centroids of the clusters and then assign the user to the cluster which yielded the smallest distance from the user. The process is repeated until users do not change

1.2 Electricity Consumption in Kenya

Kenya is an example of a country that has vastly expanded its electrification – from 2010 to 2015, grid penetration has increased by 27%, more than doubling the number of customers on the centralized grid – see Figure 1.1 [22]. In addition to the centralized grid, there are now upwards of 600,000 solar home systems deployed, which contribute another 5–6% in electrification (estimated using census figures [23] and current population estimates([24]). Most of the grid connections from 2010 to 2016 were residential, and nationwide residential electricity consumption has increased at roughly 9% annually over the period. Despite these large gains, little is understood about how much electricity these new customers consume, and even less is known about how their consumption will change with time. This study seeks to address this question: how much electricity do newly-connected electricity customers use, and how will that consumption evolve? To that end, we present a longitudinal study of electricity consumption growth in Kenya. This study is built upon a dataset of billing records from Kenya Power, the sole distribution utility in Kenya. The dataset includes monthly billing records over a six-year period, from 2010 through 2015, for a random sample from Kenya Power's customer database at the end of 2015. After cleaning and meta-data verification, the random sample amounts to roughly 136k residential customers. The scale and extent of the longitudinal dataset is heretofore unseen in the literature on electricity consumption for an African country. Further description of this dataset is provided in Section 3. To identify which customers in our randomly-sampled dataset are rural, we developed an algorithm for determining which areas of the country are urban, peri-urban, and rural based on a constrained clustering method – we describe this method and its relevance in Section 1.2.2 and Appendix 1.5.1. Subsequently we show results for urban and rural consumption, where the urban results are a straightforward combination of both urban and peri-urban customers. In Section 1.2.3, we use the results of this method as well as other customer meta-data in order to segment our sample of customers and identify patterns of consumption growth among various groups. We conclude with implications of this study for policymakers and electricity planners, discussion of the limitations

of our work, and next steps for research in the area.

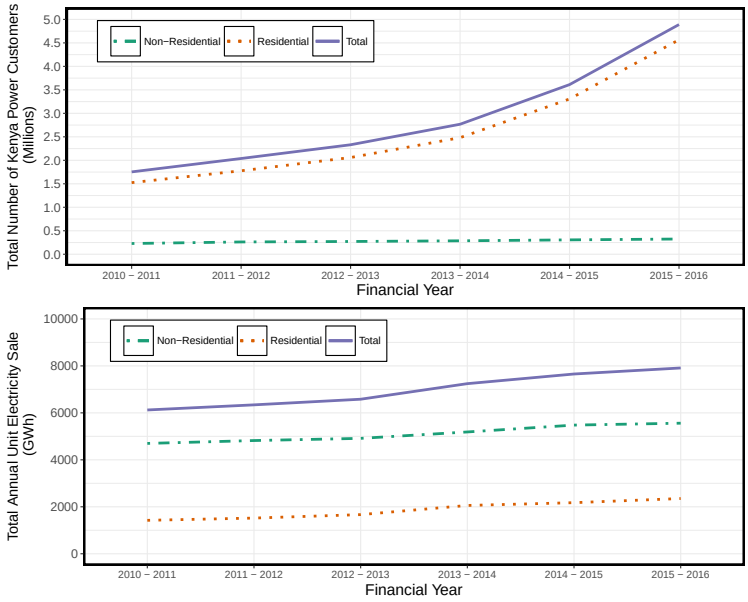


Figure 1.1: Total number of customers and total electricity sales for Kenya Power between 2010 and 2016. Non-Residential includes industrial, commercial, street lighting, and off-peak loads. Customer additions were mainly to the residential sector. Data are from Kenya Power annual reports [22].

1.2.1 Kenya Power Utility Data

Our study analyzes monthly electricity data of historical consumption in Kenya for residential customers from January 2010 through December 2015. The analysis first randomly samples customers from Kenya Power's customer database of about 4 million customers, at the end of 2015 and includes only residential customers with postpaid electricity meters. This random sample consists of 152,752 customers. Using customer meta-data such as the meter GPS location and date of meter installation (connection), we remove customers with missing GPS location or installation date data. After this filtering our study dataset contains 135,579 customers. We use the bills of this study dataset for subsequent computations and analysis. Each bill is provided as a series of components, according to a block tariff structure called the A0 (Residential) tariff. This tariff structure includes a combination of fixed and variable components; a description of these components is provided in Table 1.1.

Component	Fixed/Variable	Description
Fixed Charge	Fixed	150 Ksh
Unit Charge	Variable	1st 0-50 units @ 2.50 Ksh/Unit
	Variable	2nd 51-1,500 units @ 12.75 Ksh/Unit
	Variable	3rd Above 1,500 units @ 20.57 Ksh/Unit
Fuel Cost Charge	Variable	2.51 Ksh/Unit
Forex Fluctuation Adj.	Variable	1 Ksh/Unit
Water Resource Management Authority (WARMA)	Variable	0.05 Ksh/Unit
Inflation Adj.	Variable	0.23 Ksh/Unit
Rural Electrification Program (REP)	Variable	5 % of Unit Charge
Energy Regulatory Committee (ERC)	Variable	0.03 Ksh/Unit
Value Added Tax (VAT)	Variable	16 % of (Unit Charge + Fuel + Forex)

Table 1.1: Kenya Power residential (A0) tariff components. Note that the tariff description is as of the end of our study period; the tariff changed slightly on a couple of occasions during the study period. [25]

In addition to monthly units of electricity consumption (provided in kWh), each component also includes a bill amount (provided in Kenya Shillings – herein, KSh). In this study, we exclusively report on units of electricity consumption (kWh); discussion on the implications of this choice is provided in Section 1.2.3.

While most customers have bill data for all or nearly all months, there are some customers within this study dataset that have missing bill data, creating an unbalanced panel.

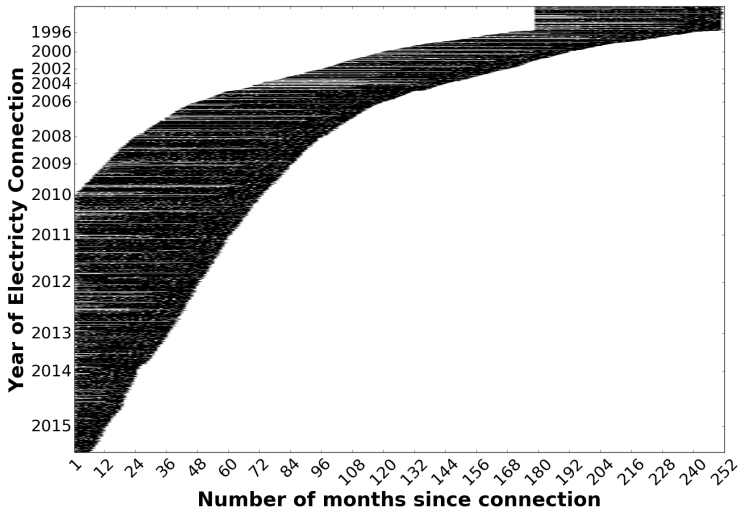


Figure 1.2: Year of electricity connection versus number of months since the electricity connection, for each of the 136k customers. Each horizontal line represents a customer over time, with black indicating the presence of data for that given customer in a given month while white represents the absence of data for the given customer in a given month. This figure shows (i.) the data available for each customer and (ii.) the data available over different durations of access.

Figure 1.2 shows the months for which bill data are available for each customer, where customers are sorted by date of installation (connection). Each horizontal line represents a customer

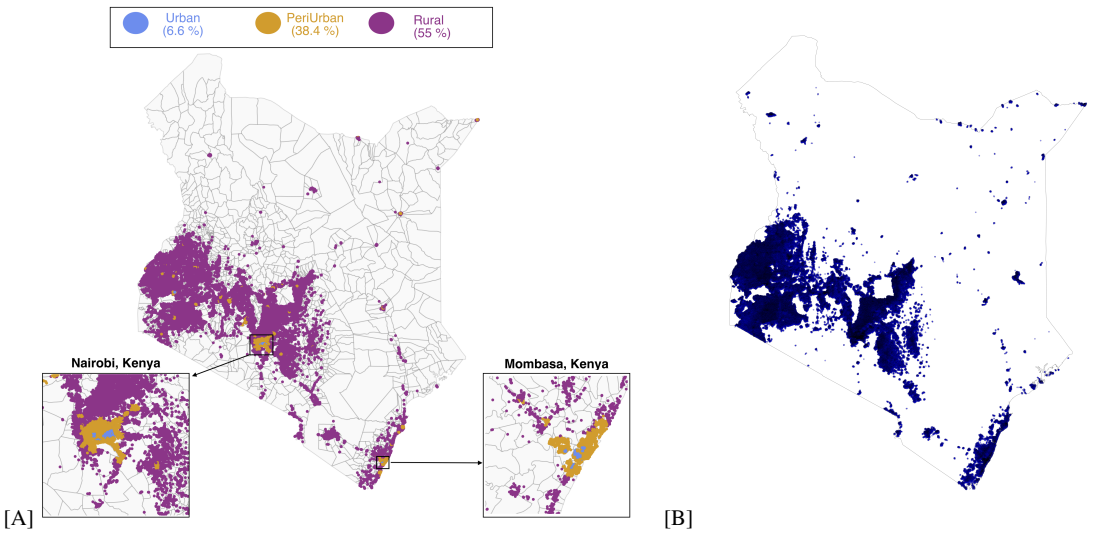


Figure 1.3: This figure compares the locations of electricity customers in our sample with the locations of population in Kenya. The customer locations are also segmented by urbanization level, showing well-defined spatial transitions from urban to peri-urban to rural. **A:**Spatial distribution of 136k customers in sample. **B:** 2015 Kenya Population – each dot represents 100 people.

over time, with black indicating the presence of data for that given customer in a given month. Conversely, white represents the absence of data for the given customer in a given month. Please note that the sample is biased to the rate of growth in Kenya Power's customer base, and that we observe different epochs for each customer based on the relationship between their connection date and our study period (2010–2015). A small number of customers, seen in the topmost rows of the graph, have six years of billing data but are listed as having an installation date of March 1, 1995; we believe these customers originate prior to 1995, but have incorrect installation dates in our dataset. Based on our interaction with Kenya Power, installation dates for customers originating prior to 1995 were not recorded thus these customers were listed as having an installation date of March 1, 1995. We do not use data from these customers for determining customer consumption growth patterns. The customers are spatially distributed across Kenya as seen in Figure 1.3(a), where each dot represents a single electrical connection. For comparison, Figure 1.3(b) shows the population density of Kenya, where each dot represents 100 people. Comparing customer locations to overall population density, there are heavy concentrations of both customers and people in the western, central, and coastal regions of Kenya. The electricity customer dataset is biased

towards higher-density areas; evidence for this claim is available in 1.5.1.

1.2.2 Methods

In order to understand consumption among different groups, this study conducts a combination of spatial and temporal segmentation of the study dataset.

Spatial Segmentation

Most newly-connected and unconnected households are in rural areas. In order to identify these households, we classify the customers in our dataset spatially by urbanization level. While it is common to use an urban-rural classification, unfortunately there is no standard definition of these categories [26]). To address this, we developed a new method for identifying urban and rural areas that makes use of high-resolution data on population density, land use classification, and satellite nighttime light intensity. We provide an abbreviated description of our method here, and describe our method in depth in Appendix 1.5.1. Similar to the approach used by Chevez et al., we apply a k-means clustering method, however we apply some constraints to the method [27]. The constraint k-means algorithm partitions predefined pixels of Kenya into k clusters, such that the euclidean distance between the pixel's features and the cluster centroid are minimized. Eq. (1) shows the objective function of the algorithm, where k represents the number of clusters, ki the number of pixels in cluster i, xj a vector of features for pixel j and i is the cluster centroid for cluster i. Unlike Chevez et al. we apply a non-random initialization to the algorithm in the form of constraints as discussed in Appendix 1.5.1. Once the clusters are obtained, we use customer GPS locations to assign each customer to a pixel and by consequence a cluster. From our experience, the clustering algorithm works best with k = 3 clusters, which we identify as our urban, peri-urban, and rural areas. Peri-urban represents areas on the urban fringe whose denizens may access urban services and resources. Figure 1.3(a) shows the clustering results from our constrained k-means algorithm. Three customer clusters are shown in blue (urban, 6.6% of customers), yellow (peri-urban, 38.4% of customers) and violet (rural, 55% of customers). Areas classified as urban are

mainly the cores of Nairobi and Mombasa, the two largest cities in Kenya, although a few urban areas can be seen in the smaller cities of Kisumu and Nakuru. The peri-urban regions generally envelop the urban locations, although other peri-urban locations border regions classified as rural. For this study, we subsequently add the peri-urban cluster to the urban cluster to form a single urban group; justification for this decision is provided in Appendix 1.5.1.

Temporal Segmentation

To tease out underline behaviors, the data was decomposed using two methods: by calendar date and by duration since customer electricity connection. For the former, post-paid billing dates are used to aggregate consumption by calendar month. For the latter, the number of months since a customer established their electricity connection is used to group customers. Most of our analysis uses this latter characterization, which aims to provide insight into growth of consumption by the duration of customers' experience with access to electricity. It is important to note that this method conflates customers from different eras into the same group, where bills from customers grouped by the same duration of experience may come from different months or years. We discuss the implications of this approach in Section 1.2.3.

1.2.3 Consumption patterns in Kenya

Using customer locations and our clustering method, customers were categorized into rural and urban groups. Table 2 shows the number of customers in each category for the entire study dataset, as well as for those who received an electricity connection before 2009 and after 2009. A majority of customers in our dataset are in rural regions (55%) and most received their electricity connection after 2009 (64.5%). Much of the recent increase in connection is due to efforts by Kenya Power, the Rural Electrification Authority (REA), and the Government of Kenya to improve access to electricity in rural areas and slums, especially via the Last Mile Electrification Program (for densification of existing transformers) and the Global Partnership on Output- Based Aid (GPOBA) Program (for formalization of connections in in- formal settlements) (Kenya, 2016).

Consumption of a representative residential customer over time

Initially, we characterize the consumption over time of all customers in our study dataset regardless of the time they obtained a grid connection. A representative residential customer is chosen as one whose consumption is the median consumption of all customers in any calendar month. Note that each month this representative customer (here the customer with median consumption in that month) is not necessarily the same customer.

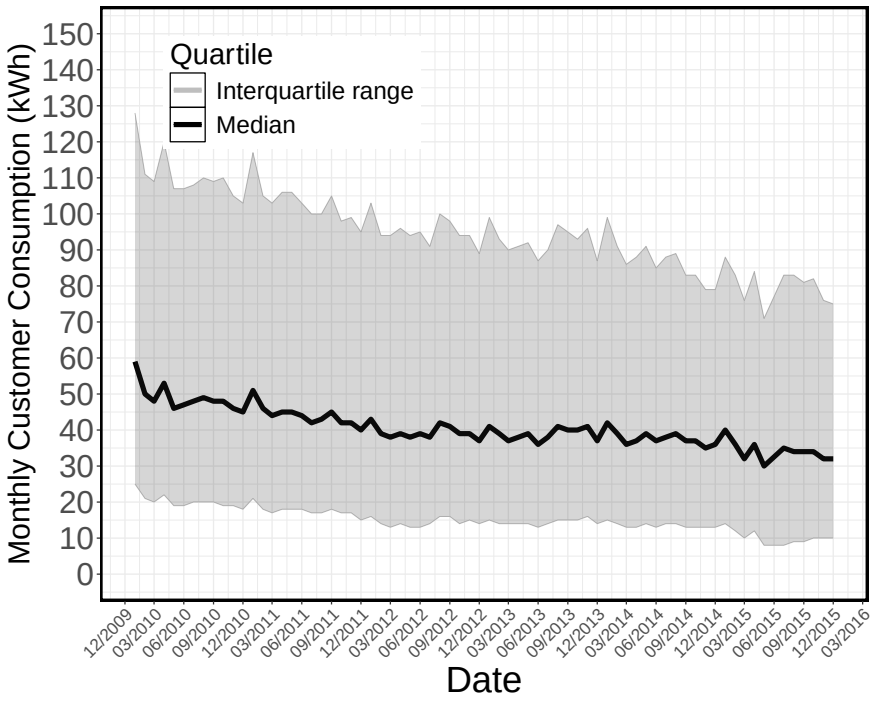


Figure 1.4: Monthly customer electricity consumption for 135,579 customers from 2010 to 2015. The solid line represents the monthly median customer's consumption while the grey area represents the interquartile range of the study dataset. From the utility's perspective, there is an increasing number of lower-consuming customers.

Figure 1.4 shows electricity consumption of the median customer (and the interquartile range of consumption levels) for each calendar month from 2010 through 2015. The Figure shows a declining trend over time for the median customer's electricity consumption (the solid line in the figure). This in itself is indicative that the utility must service an increasing number of customers whose monthly consumption is reducing.

Consumption growth over time since connection

The prior section described the consumption of a representative customer as observed by the utility. We wish to now understand whether the consumption of individual customers actually grows over time and if the growth over time varies between rural and urban customers. We initially examine monthly customer electricity consumption over time as a function of the number of months a customer has had an electricity connection; this draws on the assumption that new electricity customers are similar in their consumption patterns regardless of when they receive their first connection.

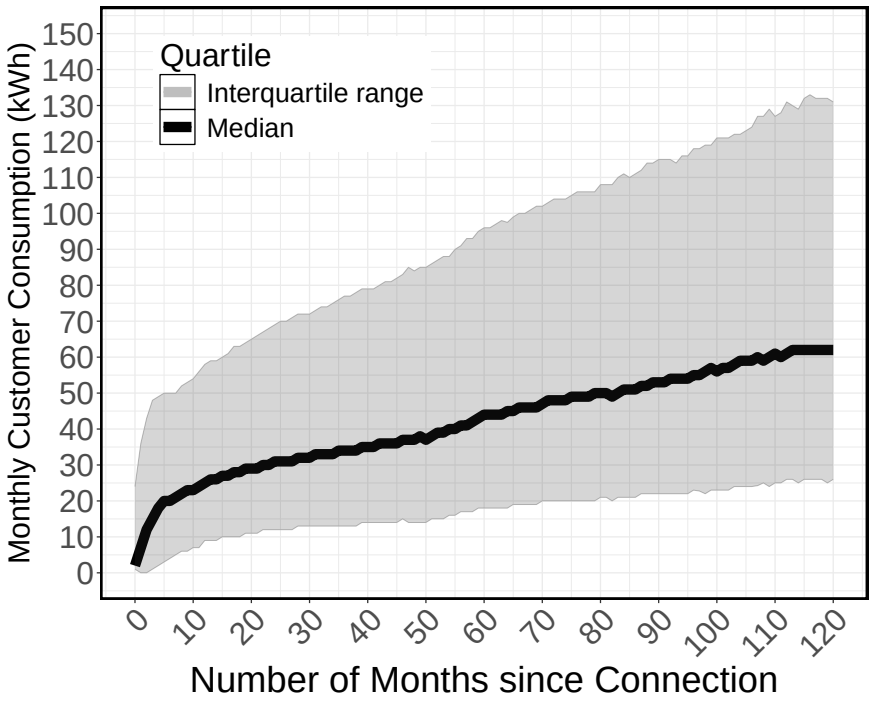


Figure 1.5: Monthly customer electricity consumption for 135,579 customers by duration of customer's electricity connection, for the first ten years of access. The solid line represents the monthly median customer consumption while the grey area represents the interquartile range. Electricity consumption for the whole dataset initially increases sharply followed by continual, though decreasing, growth.

Figure 1.5 shows this organic growth in consumption amongst residential customers in our study dataset. In this figure, the solid line indicates the monthly median electricity consumption, and the grey area shows the interquartile range. From this figure, it is apparent that monthly

electricity consumption for the whole study dataset seems to continually increases upon access. This observed behavior is due to having consumption cohorts contributing to different portions of the figure. For example, customers connected in 2014 would only influence the first 24 months of the graph while those connected in 2007 and prior would not have data in the first 24 months. This figure shows that the older customers consume higher while the newer customers consume less, noting that the customer counts for each connection year are different. We further segment consumption by connection cohort to understand the growth in each cohort. But first, we use the previously-defined customer categories (urban and rural) to further segment the consumption data.

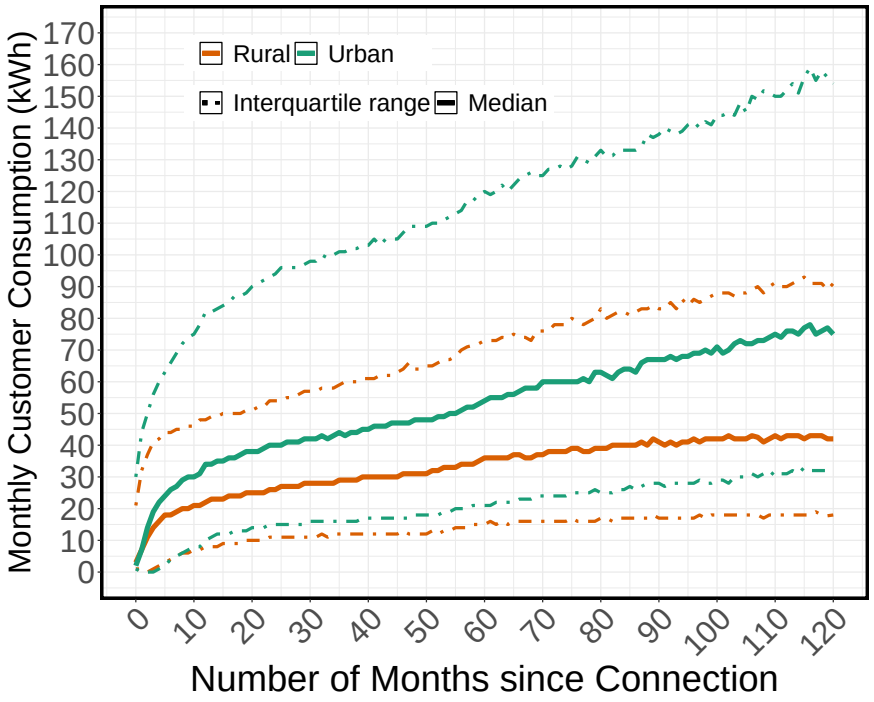


Figure 1.6: Median monthly customer electricity consumption during the first decade of access, by urbanization level. The distribution for rural customers is shown in red and the distribution for urban customers is shown in green. Solid lines are median monthly customer consumption while dashed lines show the interquartile range. Rural customers consistently consume less than urban customers.

Figure 1.6 shows electricity consumption for urban and rural customers. Solid lines represent monthly median customer consumption while dashed lines represent the interquartile range. Across all quartiles, rural customers consumed less electricity during their first decade of access than urban customers. This distinction is most pronounced with high-consuming rural con-

sumers, who use significantly less electricity than their high-consuming counterparts in urban areas. Nonetheless, each group shows the same characteristic pattern of fast initial growth followed by persistent though slowing growth thereafter.

Does the year of connection matter?

So far we have shown that customers grow their consumption upon receiving access, irrespective of their urbanization level. This perspective hides the possibility that customers connected to the grid earlier in calendar time – possibly those who were urban and started out with the means to afford a connection – might have different consumption levels from those who were connected more recently through a wave of subsidized rural electrification. Here we examine the effect of different waves of connection by grouping customers into the year they received an electricity connection.

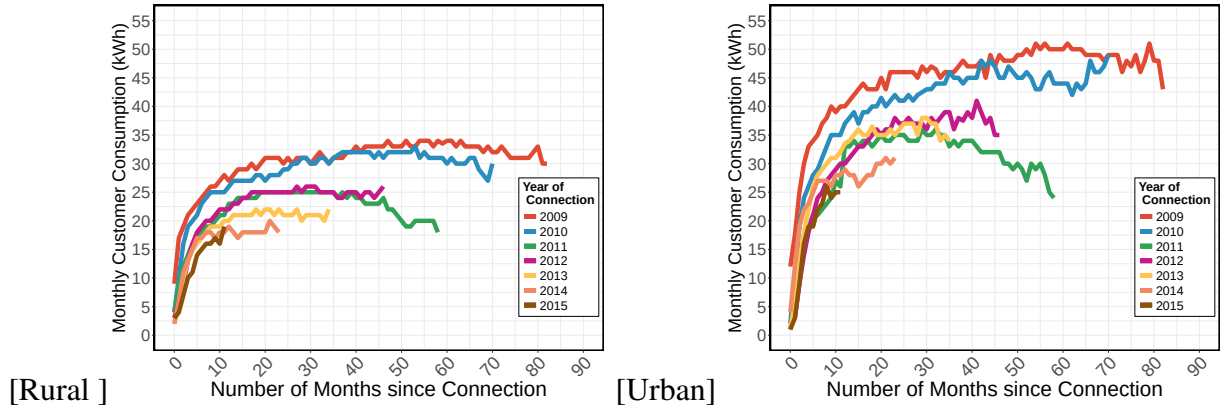


Figure 1.7: Monthly median customer consumptions, separated by the year customers received a connection. The year the median customer received a connection matters, as more recently-connected customers consume less electricity and peak sooner than customers connected at earlier times.

Figure 1.7(a) and (b) shows median customer electricity consumption for rural and urban customers, respectively. In order to ensure that we can compare consumption of customers with the same age of electricity connection, we consider only customers who received an electricity connection in 2009 or later. Looking at the figure, it is apparent that the year of connection is an important consideration for both the rural and urban cohorts, as earlier connected customers (2009,

2010) tend to peak and level off. Further, it is evident that more recently-connected customers peak sooner and at lower consumption levels than those customers with earlier connections. This pattern is fairly consistent, showing that the most recently-connected customers simply do not consume as much electricity as earlier customers even after their consumption growth has abated. In fact, the median customer whose connection began in 2009 consumes almost twice the electricity of the median 2014 or 2015 customer. Although consumption patterns are similar across urban and rural cohorts, it is clear from Figure 1.7 that median urban customers consume more electricity than median rural customers. To further explore how much more electricity median urban customers consume, we computed the ratios of consumption for each year of connection.

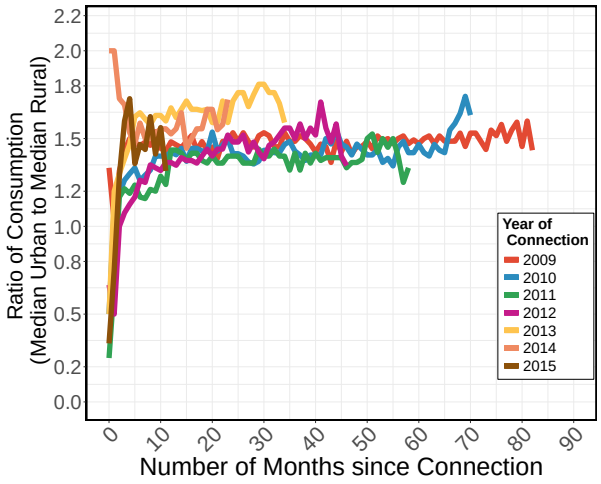


Figure 1.8: Ratio of Monthly Consumption for median urban to median rural customers. Median urban customers consume 50% more electricity than their median rural counterparts.

Figure 1.8 shows these ratios of consumption for median urban to median rural customers, separated by the year customers received an electricity connection. From the figure we see that beyond the stabilization period of 6–12 months the median urban customer consumes 50% more electricity than the median rural customer. This ratio provides a concise way to understand electricity consumption at varying levels of urbanization.

Sample size considerations

Each step of segmentation reduces the sample size of customer bills available in the segment. To ensure that our conclusions are durable, we examine the sample sizes of customer bills for these segments. Figure 1.9 shows the monthly customer sample size for each year of connection.

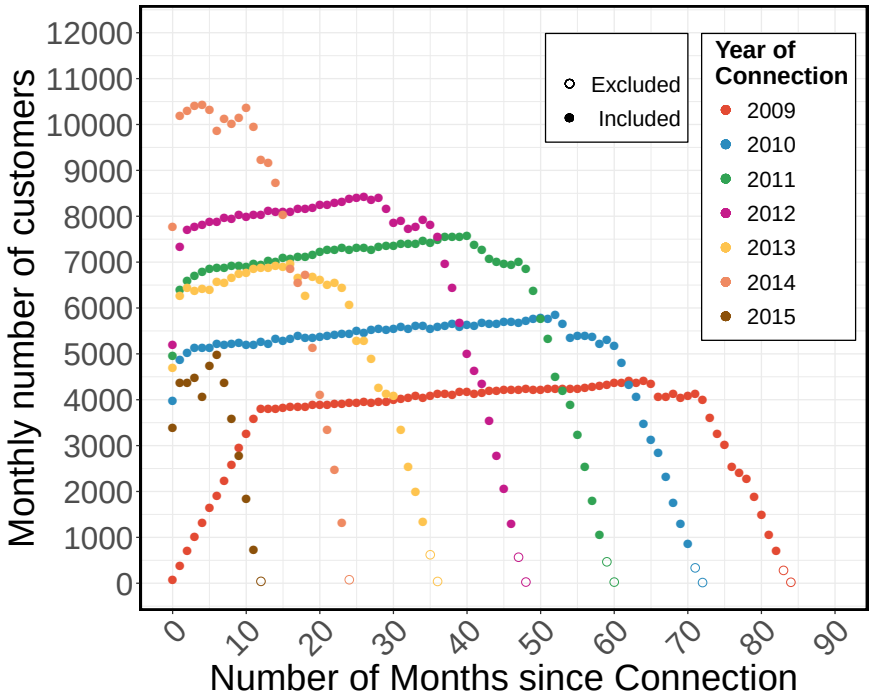


Figure 1.9: Monthly number of customers in the rural category, separated by electricity installation dates. The large number of customers, numbering in the thousands of bills, allows for confidence in the significance of our finding.

To remove points with perhaps too few samples, we filtered out months for which the sample size was less than 10% of the median sample size for a given year of connection. Since each line in Figure 1.9 is comprised of distributions numbering in the thousands of bills, we have confidence in the significance of our finding. We apply the same sample size filtering approach to customers in the urban segment.

Whose consumption is reducing?

To orient our observations towards the implications of increasing electrification, we look specifically at the rural consumers, who will comprise much of the further potential growth in the electricity customer base in Kenya. We must realize that not all rural customers have the same patterns in consumption; Figure 1.7(a) shows a drop in consumption in the later months of access. This pattern stands out for rural customers in 2009, 2010, and 2011 especially, whose consumptions reduce anywhere between 12% and 28%. While these drops appear to be synchronized in calendar dates, their appearance only among customers who started their connections in particular years along with the lack of any known macroeconomic change over the period raises questions about what caused the drop. A drop in the median could be the result of either an equally-distributed "broad" reduction or a deeper reduction focused on a particular group of customers. To investigate this question, we selected the rural customers from 2009 and 2011 and looked at their consumption in two different time periods: all of 2013, when both groups have reached their steady-state peak in consumption, and the last five months of 2015, when the drop in consumption occurs.

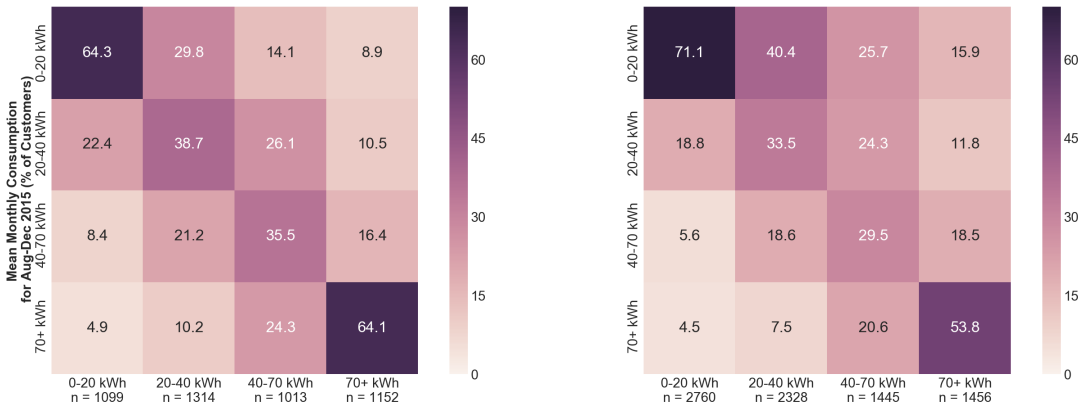


Figure 1.10: Migration within the electricity consumption distribution for (a.) rural customers with start dates during 2009 and (b.) rural customers with start dates during 2011. Horizontal axis shows breakdown of customers by mean monthly consumption for the year 2013 and vertical axis shows breakdown of customers by mean monthly consumption for the last five months of 2015.

Figure 1.10(a) and (b) are migration charts that show the percentage of customers that change their consumption bin from 2013 to 2015. Bin boundaries measured by monthly consumption in

kWh were chosen to be consistent for the 2013 and 2015 groups. The 2013 customer sample sizes of each group (n) are shown at the bottom of the chart. We can see that for both groups more customers reduced consumption than increased it and that reductions in consumption are more concentrated in the lower portion of the distribution. We also see that a larger percent of the 2011 customers dropped to the lowest consumption bin in 2015 than the earlier 2009 customers.

Customer Start Year	% in [0,20] kWh in 2013	% in [0,20] kWh in 2015
2009	24.0	29.3
2011	34.5	43.9

Table 1.2: Comparing the proportion of customers in the lowest consumption bin for two groups of customers: those starting in 2009 and those starting in 2011. For customers who received an electricity connection in 2011, more customers started in the lowest bin and a larger proportion moved there by 2015.

Table 1.2 compares the percentage of customers (2009 and 2011) who were in the lowest consumption bin in 2013 and 2015. For 2009 customers, 29.3% of all customers were in the lowest consumption bin (20 kWh) during the 2015 period compared to 24% during the 2013 period; for 2011 customers, this number is more pronounced, at 43.9% of all customers during the 2015 period compared to 34.5% during the 2013 period. Thus, for the 2011 customers, the reduction in consumption is relatively more concentrated in the lower end of the distribution. Although there is some migration to higher consumption bins, customers at the lower end of the distribution are far more likely to reduce their consumption and sometimes stop consuming entirely. While some customers may actively elect to reduce their consumption by purchasing more efficient lighting and appliances, others may be deprived from enjoying the economic and quality-of-life benefits of electricity consumption due to high electricity costs, poor reliability, lack of access to financing for equipment purchases, damaged equipment, or a combination of factors. We note that only a small proportion of customers in our sample went to zero consumption, which might imply a disconnection or other billing issue. Understanding the motivations for reductions in consumption among these lower-consuming customers, perhaps via surveys and other measurements, is a critical next step for improving customers' experiences and outcomes with electricity access as well as

building more sustainable and durable electricity-providing institutions.

1.2.4 Policy Implications

Examination of grid-connected Kenya Power customers shows that the monthly median electricity consumption of the recently-connected customers is lower than that of grid-connected customers from several years ago, comparing at the same point in time after connection. For example, a median customer in an urban area who received a connection in 2009 consumed 43 kWh per month after 18 months while a median customer in a rural area who received a connection in 2014 consumed 18 kWh per month after 18 months. This result shows that electricity planning based on earlier consumption estimates may be misleading. In this section we consider implications of our results, some limitations, and the sensitivity of our analyses to important methodological choices.

Implications for electricity planning

Countries with low GDP per capita must make critical decisions on how to allocate precious public-sector resources amongst competing priorities, especially when it comes to spending on infrastructure. For example, if Kenya tried to connect 1 million households annually to the grid, the investment in distribution infrastructure alone would exceed 4% of the annual government budget. We are assuming here that investments in generation and transmission can come from private sources. It is equally difficult to recover the investment cost from cross- subsidies applied to industrial customers. Recovering an investment of $1 billion USD from the 3575 presently-connected industrial consumers with an average consumption of 95,000 kWh per month would require an additional tariff of $0.25 USD/kWh levied on industrial customers; this is clearly an unreasonable expectation. Hence a least-cost investment approach suited to anticipated electricity demand is crucial for low-income countries. The results of this study can potentially help Kenya Power to reduce the cost of providing electricity to households. We propose three cost reduction approaches based on our findings: (i) Solar Home Systems (SHS) for low consuming customers; (ii) Reforming technical standards to connect more low-consuming customers within the existing

connection radius; and (iii) Extending the existing connection radius. Median consumption levels below 20 kWh/month for a residential customer may provide a crucial tipping point when compared to planning based on historical estimates of consumption – typically closer to 50 kWh/month. For example, a 20 kWh/month consumption level could possibly be met by an off-grid system that would deliver 500 Wh/ day or a 150 Watt peak SHS costing $500 if such a shift did not limit a customer's anticipated consumption growth. If a 20 kWh/month consumption level were met with a grid connection, the connection cost would be 2–3 times higher. On the other hand, for a 50 kWh/month consumption level, the investment cost of an off-grid system is likely to be higher than that of a grid connection. This simplistic example illustrates how the results of this study impact electrification planning in a resource-constrained economy. The real planning scenario is likely to be much more nuanced and might depend on specifics of sub-populations that are being addressed. Kenya's connection policy states that the utility charges customers who wish to connect a flat fee if those customers reside within 600 m of any transformer on the grid. This fee is 34,980 KSh ($340 USD), or 15,000 KSh ($145 USD) under the subsidized Last Mile Connectivity Program (LMCP). Customers outside of this radius who wish to connect may do so at the full cost of the connection, on average $1200 or more as the distance grows. The reasoning behind this 600m policy is a combination of engineering and cost constraints; the voltage drop experienced as well as the cost of poles and conductors needed both in- crease with a longer distance from the transformer. Knowledge of anticipated demand can shape appropriate engineering requirements of the grid. For example, one could easily and safely reduce the service standard, sized for a peak 3 kW load to perhaps 1 kW for lower-consuming customers. This would in turn lower cost of transformers, conductors, and cables as more customers can be added onto the same transformer. Less stringent yet still sufficient technical standards enable the utility to densify existing transformers at the current connection radius, lowering the per customer transformer cost, as more low consuming customers can be accommodated on the same transformer.

Alternatively, extending the connection radius with the same wire standards would potentially also allow a low-voltage wire to reach customers located further away from the transformer. In

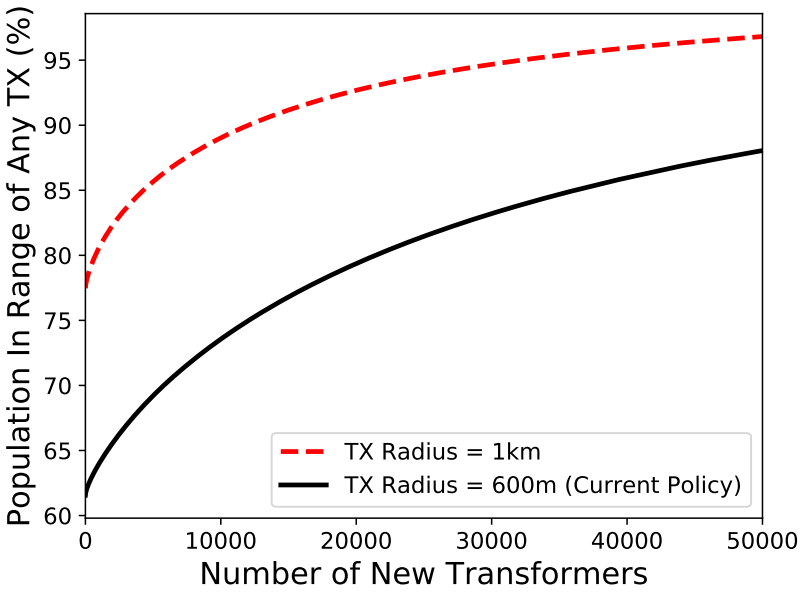


Figure 1.11: The proportion of Kenya's population within range of any of Kenya Power's transformers under two different connection fee policies: (1) customers within 1km of any existing or new transformer can connect for a flat fee (red line) and (2) the existing policy, where customers within 600m of any existing or new transformer can connect for a flat fee (black line). Note that Kenya Power presently has a total of roughly 58k transformers, and those transformers are within range of 62% of the country's population.

Figure 1.11, we show the implications for Kenya Power if the connection radius were increased. For this analysis, we use a greedy algorithm that places new transformers in the locations that maximize the population covered. At present, 62% of Kenya's population lives within 600 m of Kenya Power's roughly 58,000 transformers, and the Government of Kenya has a stated goal of providing access to electricity to 100% of the population by 2020. According to the figure, maintaining the same connection policy and attempting to reach 85% of Kenya's population with the grid would require an additional 35,000 transformers. However, newer transformers are in rural areas where customers are further apart, but voltage drops are lessened due to lower consumption per customer. Thus, relaxing the 600 m constraint no longer poses as much of an engineering challenge and would enable the grid to reach more customers with existing or fewer additional transformers. If the policy were changed to allow any customer within 1km of any transformer to connect for a flat fee, it would take fewer than 5000 additional transformers to reach the same 85%

of the population. While the cost of connections is still a heavy burden achieving those connections by extending existing low-voltage infrastructure, as opposed to deploying new transformers, may present a lower-cost option. Further, this strategy would align well with the LMCP, which aims to densify existing underused transformers using a budget of roughly $450 million USD. It is important to note that existing plans for the three phases of LMCP (an investment of roughly $450 million USD) include only 1400 additional transformers, challenging the Government of Kenya's stated goals of reaching 70% electrification by the end of 2017 and universal electrification by 2020. Without a significant change of direction on alternative means of electrification, massive reductions in connection costs, or unexpectedly high growth in electricity consumption, the utility model faces severe challenges in meeting the dual mandate of universal electrification and investor profitability. Sustained low consumption levels will hinder the financial viability of utilities whose goal is to increase electricity access. It may be possible to boost consumption and by consequence financial viability via targeted programs such as appliance financing and tariff subsidies. These can create more growth in electricity consumption, support higher quality-of-life and have potential income benefits for customers while supporting the dual mandate of electricity providers. Although our discussions have focused on Kenya, we believe that Kenya Power's experience can highlight broader lessons that are relevant for utilities in other developing countries.

- Customer consumption may not grow at a constant percentage over time.
- Performing better customers analytics, prior to deciding how to connect these customers can result in fewer underutilized grid connections, allowing more customers to be reached at a cheaper cost.
- The assumption that everyone must be connected in the same manner has both benefits and costs, and it is important to quantify the costs to design evidence-based policy.

Additional considerations

Urban/Rural Sensitivity: Urbanization levels were defined using a combination of datasets. However, we recognize that there are a range of classification methods for determining urbanization level, and that our results are sensitive to the method we used. Additionally, not all rural regions are similar – localized economic effects will not be captured by this approach, but we attempt to deal with this by primarily considering medians as well as interquartile ranges, so as to not be affected by extremes in the distribution. Further, definitions of urbanizations are hardly static as captured in our clustering analysis. These definitions change with time and are influenced by changing socio- economic factors and migration. Thus our definition of urbanization levels only capture one snapshot, which is at the start of the analysis period (2010). Future work on this topic is to examine how consumption evolves in areas that experience slower or faster changes in urbanization levels.

Other Temporal Effects: Analyzing customer growth on a calendar basis conflates the effects of a growing customer base with those of an evolving customer base, a typical situation for grids in sub-Saharan Africa. In an effort to disaggregate these two, we spend the majority of our analysis analyzing customers via the lens of time since electricity connection. While transforming the temporal axis from calendar dates to time since electricity connections reveals relevant information for electricity access, there are also adverse effects to consider. This approach obscures the effects of cyclic and seasonal changes, macroeconomic shocks, and, as we show in this work, differences among newer and older customers. While we acknowledge that these exogenous events occurred during our study period, we believe that a six-year duration to our study should allow examination of larger trends in growth of consumption among these customers.

Tariff and Meter: We use kilowatt-hours as the measurement of consumption over time, with limited consideration of the various tariff structures in place for these customers. Some of these tariff components changed during the course of the study period; for example, in mid- 2014, the fixed tariff increased from 120 KSh per month to 150 KSh per month. Some of the variable tariff components also had small changes during the study period, and others, such as the Foreign Exchange

(Forex) and Fuel Cost Charges changed on a monthly basis to reflect market conditions. While many of these changes were seemingly negligible, more in-depth analysis is needed to estimate the scale of these effects on longitudinal consumption. In particular, our sample consists of customers only on postpaid electricity meters. Initially, we do not have any clear evidence that differentiates these customers from customers with prepaid electricity meters. However, since customers with postpaid meters tended to receive their connections earlier, as a class they are likely more wealthy than their prepaid counterparts, potentially depressing the consumption values reported throughout this paper. We take it as future work to understand the implications of examining only customers with postpaid meters, and seek to compare the consumption patterns among those customers with postpaid and prepaid electricity meters.

Equity: Different electricity delivery technologies within the same community challenge notions of equity in electricity connections and may pose political barriers. Quantifying the costs of equity of connections, though not necessarily equity of service, are worthy of further study, though beyond the scope of this work.

1.3 Electricity Consumption in Rwanda

Over the past decade, Rwanda has seen increased electrification rates from 10% [28] in 2011 to 64.5% [29] as of June 2021. As of June 2021, 47% of Rwanda was connected to the national grid and 18% utilizing off-grid systems primarily solar. The government has set a target of achieving 100% electrification by 2024 with all productive users gaining access by 2022. 100% electrification will be achieved by having 52% of users connected to the grid with the remaining 48% utilizing off-grid systems[30]. Understanding electricity consumption for already electrified users can provide the government with insights on which users might benefit more from grid versus off-grid systems.

Increased electrification efforts have been primarily led by the Rwanda Energy Group (REG), a government owned holding company incorporated in July 2014 with two subsidiaries; the Energy Utility Corporation Limited (EUCL) and the Energy Development Corporation Limited (EDCL). While EDCL focuses on increasing investment in the development of new energy generation and

transmission infrastructure as well as planning and executing energy access projects, EUCL focuses on provision of utility services through operations and maintenance of existing generation plants, transmission and distribution networks and retail of electricity to end-users[31]. In addition to growth in electrification, Rwanda has also experienced a series of tariff changes in the past decade. Table 1.3 shows the electricity tariff structures for residential customers in Rwanda over the past 15 years[32].

Table 1.3: Residential Tariff Structure for Rwanda. A fixed 500 FRW service charge was in place but later removed in 2015.

Year	**Tariff Structure**	**Energy Charge (FRW/kWh)**
2006	Flat Rate	112
2012	Flat Rate	134
2015	Flat Rate	182
2017	0 - 15 kWh	89
	15 - 50 kWh	182
	>50 kWh	189
2018	0 - 15 kWh	89
	15 - 50 kWh	182
	>50 kWh	210

Electricity tariffs prior to 2012 were considered to be below cost of service and therefore a 20% increase in residential tariff was approved by Rwanda Utilities Regulatory Authority (RURA), the agency mandated to regulate the energy sector[32]. Residential tariffs were further increased by 35% in 2015 with authorities citing increased generation costs as a result of the continued reliance on thermal energy[33]. A fixed service charge of 500 FRW was in place prior to 2015 but was removed in 2015. Two years later in 2017, a block tariff structure was introduced by increasing the tariffs for high consumers while introducing *"lifeline"* tariffs for the lowest consumers. Given that most new connections are low consumers, the block structure aimed at improving their ability to afford electricity[34]. There has been an acceleration of new electricity connections to the grid in the latter half of the past decade as Rwanda seeks to achieve universal electrification by 2024[29, 34].

In this work, we present a data-driven descriptive analysis of electricity consumption in Rwanda with the goal of understanding how consumption has evolved for different cohorts. This work also

highlights the changes in electricity consumption given the tariff policies. Our analysis is built on prepaid purchases of 811,541 Rwandan customers (with over 361 Million transactions) between 2012 - 2020. Specifically this work aims to answer these two research questions:

1. How much electricity do newly-electrified customers use, and how has that consumption evolved?

2. How do different tariff policies influence utility revenues and electricity consumption?

By answering these research questions, this work can be relevant to other utilities seeking to electrify more of its population by providing insights on customer behavior and responses to varying tariff polices.

1.3.1 Rwanda Energy Group Utility Data

Our panel dataset obtained from the Rwanda Energy Group consists of 361,029,383 prepaid historical electricity consumption purchases from 2012 to 2020. This dataset covers 811,541 grid connected customers who received an electricity connection between 1996 to 2019. For each customer, historical electricity purchases are available, showing both the quantity of electricity purchased and the corresponding tariff for the purchased units of electricity. While the dataset includes associated taxes such as Value Added Tax (VAT), our analysis only evaluates the amount of electricity purchased (kWh) and the per kWh tariff. We observe that 88 % of customers are residential while the remainder are non-residential (including small commercial, industries, hotels, health facilities). Our dataset does not include large industrial customers. REG has undergone a number of revisions in their customer categorization process over the years resulting in customers belonging to different categories over time. For this study, we take the customer's final category as of 2020 as their customer category. Figure 1.12 shows the number of new electricity connections made by the utility for customers within our dataset [3]. This aligns with REG's reports of higher number of residential customers compared to non-residential customers. We observe that 66 %

[3] Our dataset ends in April 2020, resulting in the dip in connections on the figure for the year 2020.

of electricity connections are made within the time frame of our study period 2012-2020. This analysis focuses on the customers who are receive a grid connection during the study period.

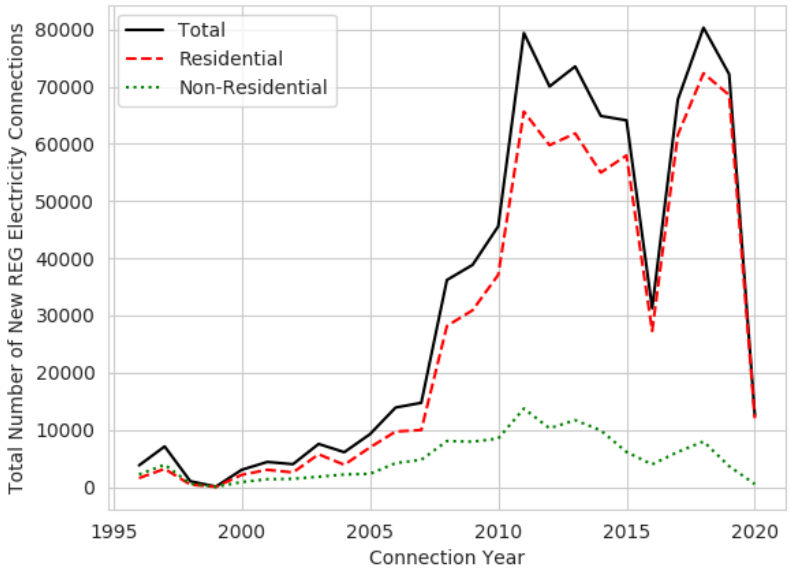


Figure 1.12: Annual number of new electricity connections made by the utility (REG) for both residential and non-residential customers. 66% of new connections were made after 2012 while 88% of customers within the dataset are residential.

The dataset consists of customers from 15 of the 30 districts that make up Rwanda. We observe that 37.4 % of customers within the dataset are from Kigali, the capital of Rwanda, which makes up three of the 15 districts for which we have customer data. The second largest clusters of customers are located in the south western and northwestern districts, followed by the Northeastern and southern districts and the lowest number of connections in the central districts that neighbor the capital Kigali as shown in Figure 1.13.

1.3.2 Methods

From pre-paid purchases to monthly consumption

Our dataset begins with pre-paid purchases made by customers at varying points in time. In Rwanda, grid connected electricity customers tend to have pre-paid meters and can purchase units

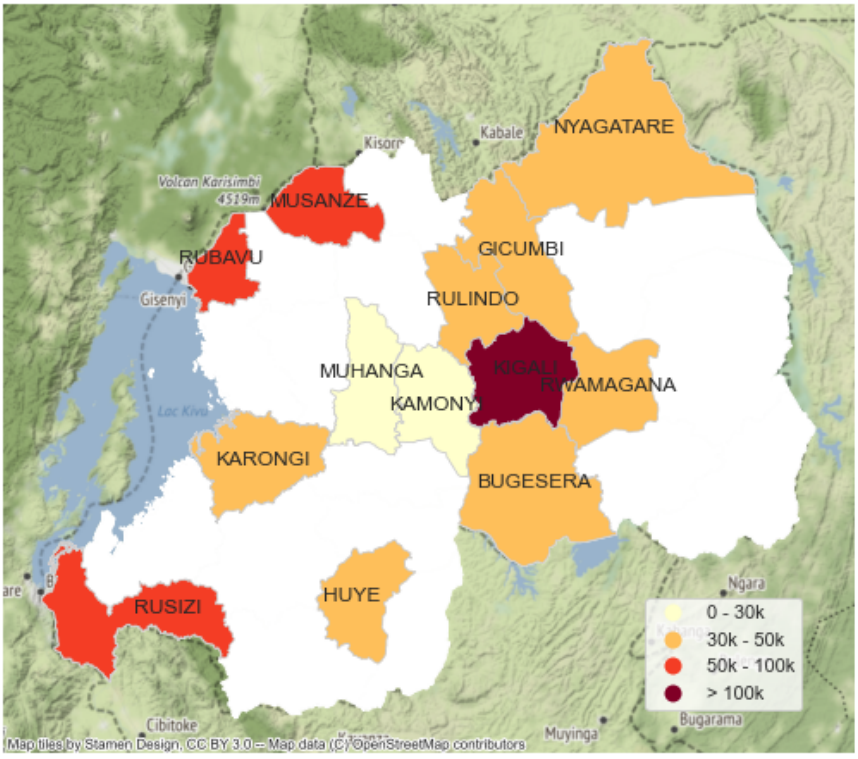


Figure 1.13: Spatial coverage of Rwanda customers (residential non residential) within our 811K dataset

of electricity on an *as-needed* basis. To compare consumption overtime at the same frequency, the following pre-processing strategy is applied. i) Median purchase frequency for each customer i is estimated e.g. every 5 or 7 days. Given any sequentially occurring two purchase periods (t_n and t_{n+1}), consumption is spread between both purchase periods (on a daily basis) if the time delta between both periods is lower than the median purchase frequency. Otherwise the purchased units are spread over the median purchase frequency of the customer. The data smoothing process is performed at a daily resolution, after which the daily consumption is aggregated to monthly consumptions in kiloWatthours (kWh). The smoothing process transforms stochastic electricity purchases to monthly electricity consumption to support further analysis. The smoothing process preserves the aggregate monthly consumption seen across customers, thus giving confidence to the smoothing method.

Temporal Segmentation

To understand how residential electricity demand evolves over time, we dis-aggregate the electricity consumption data temporally (similar to the methodology in 1.2.2) based on the number of months that each customer has been connected to the grid. By doing this, we are able to understand how a typical customer's electricity consumption changes with each additional month of being connected to the grid.

Tariff Analysis for Residential Customers

Our tariff and revenue analysis is performed on three electricity consumption classes (i.e. low, medium and high). The classes are defined to align with RURA's tariff block criteria as defined in [34]. Eqn (1.1) shows the class definitions given the average monthly amount of electricity $\bar{y}_i$ which is used by customer i in a 12-month period (a year prior) to the tariff change.

$$c = \begin{cases} Low, & \text{if } \bar{y}_i <= 15 \text{ kWh.} \\ Medium, & \text{if } 15 < \bar{y}_i <= 50 \text{ kWh} \\ High, & \text{if } \bar{y}_i > 50 \text{ kWh} \end{cases} \tag{1.1}$$

Table 1.4 shows counts within each class for tariff periods 2015 and 2017. High confidence customers by class are defined as customers who have at least 80 % of their data in the assigned class given the class allocation from the $\bar{y}_i$. These customers are used to evaluate the impact of tariff changes on utility revenues and electricity consumption.

Table 1.4: Customer count by class prior and after obtaining the high confidence balanced sets.

	2015 Tariff Period			2017 Tariff Period		
	Low	Medium	High	Low	Medium	High
All	133,735	92,304	37,708	170,209	94,708	34,806
High Confidence	99,512	28,329	16,473	124,701	30,719	16,666

1.3.3 Consumption patterns in Rwanda

Consumption of representative REG customers

First we present the overall consumption behavior for all residential and non-residential customers in our dataset, regardless of their grid connection date. Amongst residential customers, the total monthly residential consumption for customers within the dataset ranged between 7 to 9 GWh/month. Despite a 54 % increase in the number of residential customers between 2016 and 2020, the total residential monthly consumption only increased by about half (28.5 %) during that timeframe. This suggests that while the utility is increasing the number of people with access to electricity, the newly connected customers are consuming less electricity. Thus increased the number of grid connections does not always translate to consumption. For non-residential customers, electricity consumption increased by about 10 % between 2016 and 2019 while the number of connections increased by about 22 %. Even in the non-residential sector, the utility is adding many more smaller commercial customers relative to the initially connected high consumers. While more residential customers are being connected relative to non-residential customers, the total monthly consumption for the non-residential sector remains higher than that of the residential sector. This suggests that the revenues from the fewer non-residential customers may be cross-subsidizing the associated costs for the many more residential customers.

We also observe that between 2013 - 2019, electricity consumption growth was on average 4.6% with the non residential sector recording an average growth of 6.3% and 3% in the residential sector. This falls far below REG targets of 15% demand growth required to avoid excess costs of generation which would in turn necessitate more government subsidies [35]. It is important to note that this study does not include electricity consumption data from large scale industries that would have a significant impact on the aggregate electricity consumption growth.

We also analyzed the purchasing frequency of prepaid electricity tokens. We observe that in Kigali, residential customers purchase on average about 13.9 kWh every 16 days while non-residential customers purchase about 47.5 kWh every 13 days. In comparison, outside Kigali resi-

dential customers purchase about 7.2 kWh every 37 days while non-residential customers purchase 21.2 kWh every 20 days. These observations show that residential customers in Kigali purchase both twice the amount and twice as frequently as residential customers outside Kigali. Or put otherwise, rural household use less electricity and purchase at half the cadence of their urban peers. A similar observation is made amongst non-residential customers who purchase twice as much in Kigali compared to outside Kigali. These observations align with the general expectation that productive electricity demand is highest in urban centers and that urban regions use more electricity than non-urbanized regions.

Consumption growth with time in Rwanda

Previous studies on residential and non-residential grid connected customers in Kenya have shown interesting patterns of how electricity consumption evolves as a function of time spent on the grid [36, 37]. This pattern of fast initial growth (in the first year of receiving a connection), followed by plateaued or slowed growth is identical for both residential and non-residential customers albeit with slight differences for rural vs urban locations. Using a similar method as [36, 37] and all customers (regardless of their data completeness), we kick off our descriptive analysis by assessing how electricity consumption among REG's residential customers evolves as they "mature" on the grid.

Our first observation is that generally, more newly connected customers progressively stabilize at lower levels of consumption. That is, customers connected in 2019 and 2018 have the lowest median consumption when compared with older customers as shown in Figure 1.14.

This pattern is consistent with similar observations by [36] and [37] among Kenya's grid connected residential and small commercial electricity consumers. Our hypothesis regarding this behaviour is that newer customers are mostly low-income households that are connected through the government rural electrification program and as such tend to consume less electricity. Unlike observations in Kenya, consumption levels in Rwanda tend to be much lower. For example, the median residential Kenyan customer electrified in 2013 plateaued around 30 kWh/month, while the

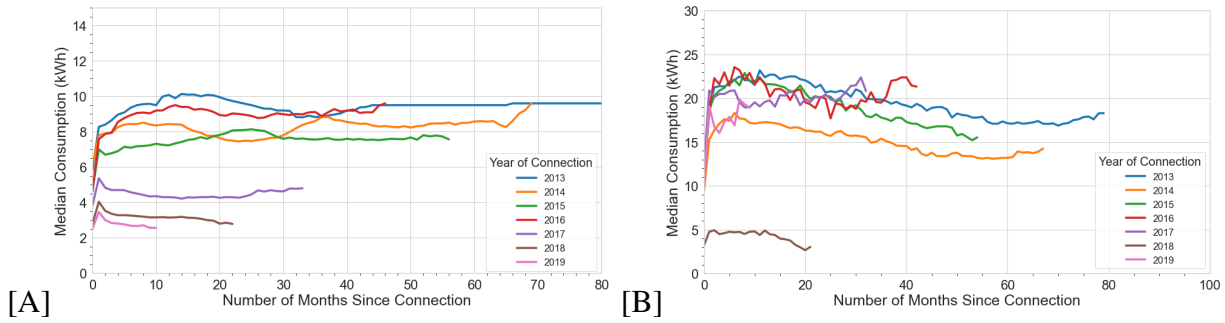


Figure 1.14: Shows median monthly electricity consumption for REG's electricity customers segmented by year in which they got connected to the grid. **A:** Residential customers within the REG dataset. **B:** Non-Residential customers within the REG dataset

median residential Rwandan customer also electrified in 2013 plateaued around 10 kWh/month. Though the absolute difference is only 20 kWh/month, implications for energy providers are significantly different as these customer groups might be better suited for different electrification technologies. One hypothesis around much lower median consumption levels in Rwanda might be due to differences in household income between both countries. It is worth noting that the 2020 per capita GDP in Kenya was $1879 while that of Rwanda was $798 [38]. In depth income analysis (possibly through extensive surveys) are needed to better understand the reason for consumption differences between both customer groups despite being electrified in the same calendar year. The differences in consumption amongst varying connection cohorts in Rwanda suggest that providing access to electricity does not guarantee consumption of electricity and improved welfare, and that other micro-economic factors might influence consumption.

The second observation is that median electricity consumption amongst all cohorts remains fairly flat or may slightly decrease as consumers "mature" on the grid. This suggests that the Rwandan utility is servicing an increasing number of customers whose consumption is either stationary or not changing much, and as such, who might have to be cross subsidized by older customers. Our dataset consists mainly of monthly consumption (kWh), thus additional variables which may help explain the plateauing behavior or slight decrease are not available.

Furthermore, we test the relationships between monthly kWh consumption and number of months since connection under a regression framework to evaluate if the observed pattern of growth

and stabilization is a statistically significant relationship across all members of the same connection year cohort (not only the medians). The regression results show that for residential customers connected prior to 2018, there is a positive relationship between consumption and time. On average, each additional month yielded a small but significant increase in consumption. For residential customers connected 2018 and after, they on average experienced drops in electricity consumption overtime. This result aligns with the general trend in Figure 1.14 showing the behavior of the median customers. As for the non-residential customers, we observe mostly negative correlations between kWh consumption and number of months spent on the grid within the 2013-2018 connection year cohorts, which is also consistent with Figure 1.14 (See Table 1.9 in the Appendix 1.5.2).

The third observation is around the large consumption differences between urban (Kigali) and rural (non-Kigali) districts. Figure 1.15 shows the median and interquartile range of customers by urbanization level. We observe that customers consuming in the upper quartile and residing in non-Kigali districts consume similar amounts of electricity as median customers residing in Kigali. This emphasizes the even lower consumptions (under 20 kWh even after 4 years of receiving an electricity connection) that are experienced outside of the urbanized district of Kigali. Beyond the difference between Kigali and other districts, we also observe a large consumption distribution that exist amongst both Kigali customers and customers located in other districts. This wide distribution highlights the variability in consumption amongst customers. Thus methods that are able to better differentiate between customers would be critical for energy access planning.

1.3.4 Implications of Tariff Changes on Utility Revenues and Electricity Consumption

Having seen how a typical REG customer's electricity demand evolves as a function of time spent on the grid, this section seeks to understand how residential tariff changes (in 2015 and 2017) affected demand for electricity and by extension revenues remitted to the utility. To do this, we present the average residential revenue per customer collected by REG after segmenting customers based on low-medium-high categories described in the methodology section.

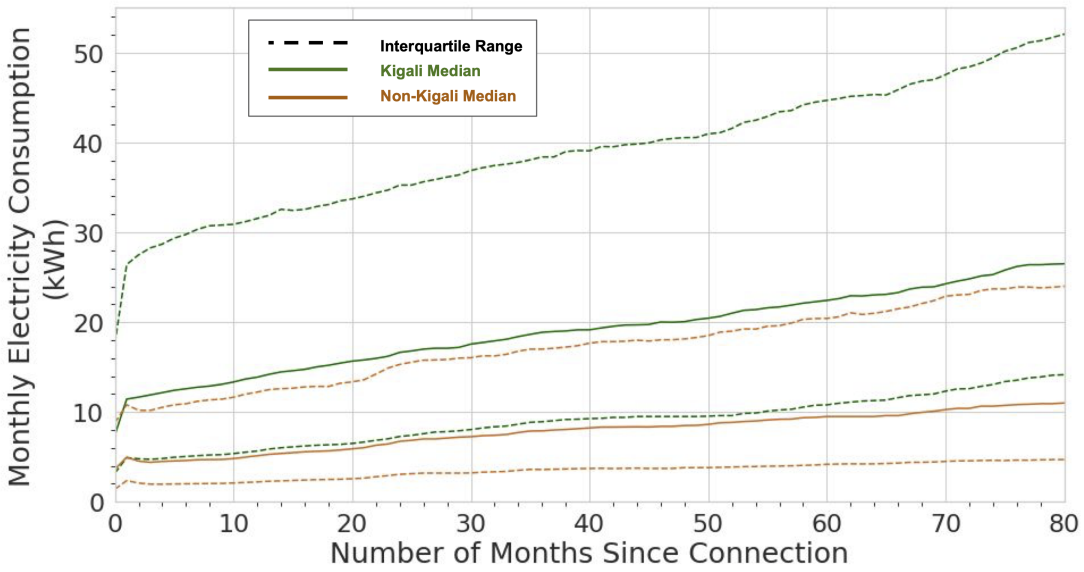
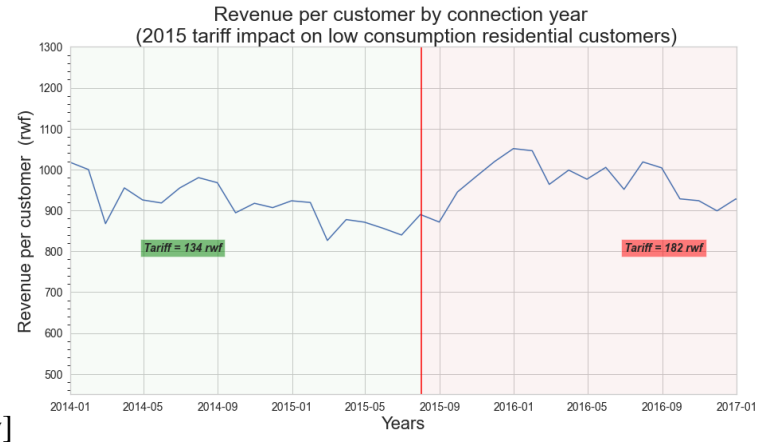


Figure 1.15: Electricity consumption over time for customers in Kigali versus non-Kigali customers.

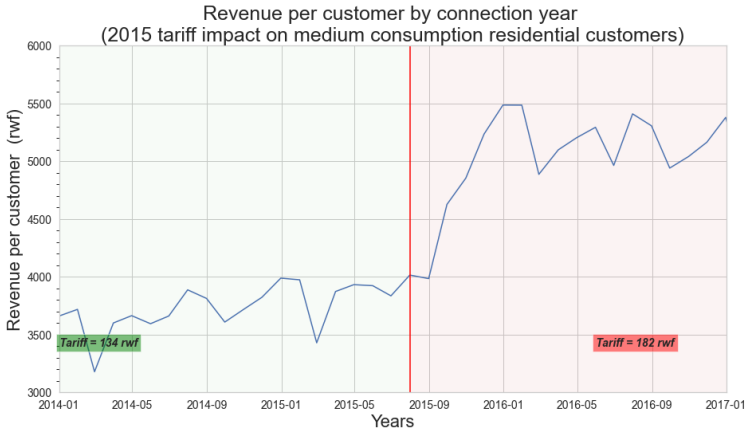
Average revenue per customer after 2015 tariff change

As a result of the increased operating costs associated with running diesel power plants at Jabana 1 and 2, REG increased the tariff for all customers by 35% [33] from 134 FRW to 182 FRW . To see if the policy objective of increasing company revenue was met by this tariff increase, we present the average monthly revenue per customer before and after the policy change for each consumption group as shown in Figure 1.16. We observe from the figure that the highest increase in average monthly revenue per customer occurred for the medium category of residential customers, though smaller increases in average monthly revenue per customer were also observed for the low and high categories. The steepest increase in revenue happens in the first 3 months following the tariff change date, after which the average monthly revenue per customer appears to stabilize. The actual percentage changes in average monthly revenue per customer are reported in Table 1.5.

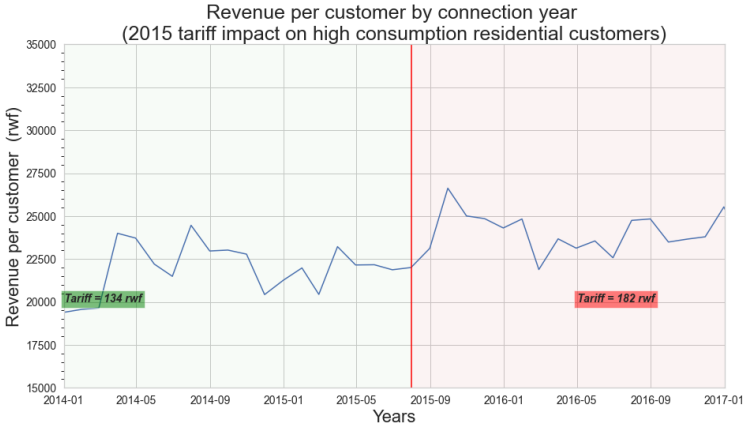
Given the 35% increase in electricity price, the average corresponding per customer increase in revenue remitted to REG one year prior to and after the tariff change ranged between 10% and 33% and a corresponding reduction in kWh consumption ranged between 2% and 18%. The largest per



[Low]



[Medium]



[High]

Figure 1.16: Shows changes in revenue per customer among low, medium and high consumption customers after introduction following the 2015 tariff change from 134 RWF to 182 RWF. The red vertical line indicates the date when the tariff change took effect.

customer increase (33%) in revenue was observed in the medium consuming group which consists of customers consuming between 15 and 50 kWh/month. While the lowest increase in revenue (10%) was observed in the high consuming customers. Although the high consumers displayed

Table 1.5: Shows the average change in average monthly revenue per customer and average change in customer consumption for each of the Low, Medium and High categories of customers one year pre and post the 2015 tariff change.

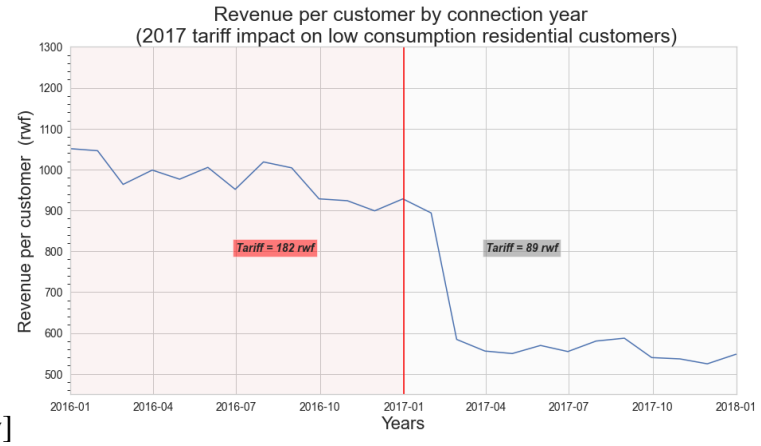
	Low	Medium	High
Pre 2015 Tariff Change (FRW)	825	3,846	22,029
Post 2015 Tariff Change (FRW)	991	5,140	24,131
Avg % Change	+20	+33	+10
Pre 2015 Tariff Change (kWh)	6.2	28.7	164
Post 2015 Tariff Change (kWh)	5.5	28.2	134
Avg % Change	-11	-2	-18

the lowest average per customer percent increase in revenue they had the largest absolute monetary revenue increase of 2,102 FRW. It is noteworthy that the average consumption for medium consuming customers barely changed with the introduction of the new tariff, while the consumption of high consumer dropped by 18 % (the highest observed drop in consumption). On the whole, the consumption group with the highest revenue increase showed the smallest average decrease in consumption, while the consumption group with the smallest revenue increase decreased its consumption the most. By increasing the tariffs by 35% the average low consuming customer paid an additional 166 FRW per month. For households living on a dollar a day, this represents about an additional 0.7% of their income dedicated to electricity consumption. [39] suggests that households spend about 3% of their income on electricity. Thus a marginal increase in income of close to 1% dedicated to electricity consumption might suggest why on average electricity consumption decreased.

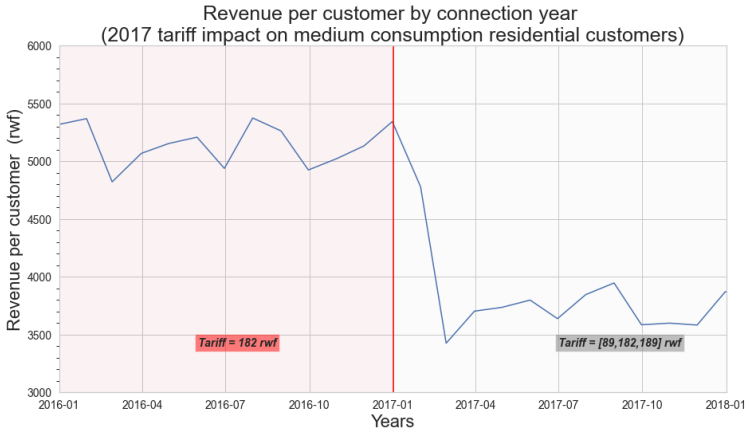
Average revenue per customer after 2017 tariff change

Similar to our analysis of the 2015 tariff change, we examine how introduction of the block tariff in January 2017 impacted the per-customer revenue remitted by residential customers to REG.

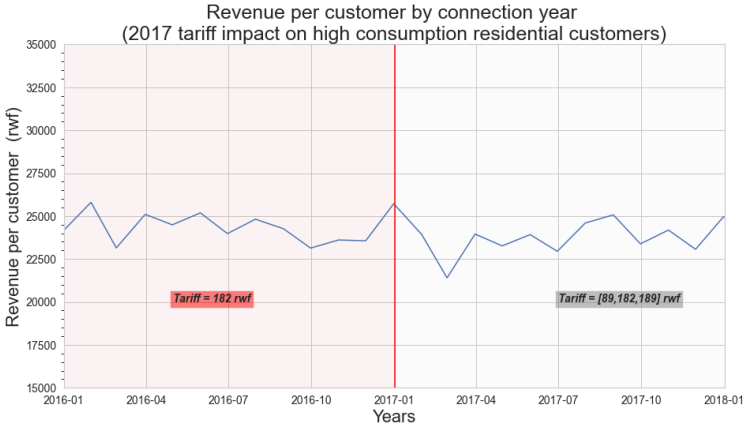
Figure 1.17 shows the average monthly revenue per customer pre and post the 2017 tariff change. We observe significant drops in average monthly revenue per customer for the low and



[Low]



[Medium]



[High]

Figure 1.17: Shows average changes in average monthly revenue per customer among low, medium and high consumption customers after introduction of the block tariff in January 2017. The red vertical line indicates the date when the tariff change took effect.

medium categories, while the high category experienced a very small drop in average monthly revenue per customer. Similarly, this drop is experienced within the first 3 months of the tariff change. Table 1.6 summaries the percentage changes in average monthly revenue per customer. From Ta-

ble 1.6 we observe that for customers in the low category, there was on average a 39% decrease in average monthly revenue per-customer as a result of the 49% reduction in tariff. This correlates with the observed increase in average electricity consumption by low consuming customers of 12% which offset further decrease in average monthly revenue per-customer. In the medium category, we see that despite the tariff remaining unchanged for consumption above 15kWh a month, households on average remitted 26% less revenue to grid after the tariff change. We also observe that consumption remained nearly constant pre and post the tariff change. This suggests that the drop in revenues remitted by this group is as a result of customers paying a lower unit price for their first 15 kWh of consumption. The effect of decreased remitted revenues might be exacerbated if most of their consumption occurs within in the lowest consumption tier of less than 15kWh/month. Furthermore, the block tariff generally leads to a decrease in revenues among the low and medium categories because any increase in monthly kWh consumed (as a result of reduced price) is not enough to make up for the fall in revenue per-household. On the other hand there was very little change in the average monthly revenue per customer for households in the high consuming group.

	Low	Medium	High
Pre 2017 Tariff Change (FRW)	970	5,136	24,400
Post 2017 Tariff Change (FRW)	590	3,795	23,720
Avg. % Change	-39	-26	-3
Pre 2017 Tariff Change (kWh)	5.3	28.1	134.0
Post 2017 Tariff Change (kWh)	6.0	28.0	134.6
Avg. % Change	+12	-0.6	+0.4

Table 1.6: Shows the average revenue per customer collected by REG and average customer consumption for each of the Low, Medium and High categories of customers one year pre and post introduction of the block tariff in January of 2017

1.3.5 Discussion

Decreasing consumption and Electricity planning

Analyzing grid-connected customers in Rwanda reveals that monthly residential electricity consumption is overall low ranging from 4 - 10 kWh/month for the median customer. Further-

more, more recently-electrified households consume less electricity than their counterparts who were connected earlier on. A similar trend is also observed on the non-residential side where small businesses connected in 2019 consume about 5 kWh/month after 12 months, while those connected in 2013 consume about 23 kWh/month after 12 months. These findings are consistent with those made by [40, 37] in Kenya.

For any household in Rwanda, the official electricity connection fee is 56,000 RWF ($54), which is far below the cost to the utility to electrify the home. This connection fee was set up in 2013 around the onset of the Electricity Access Rollout Program (EARP). This low connection fee, despite not reflecting the true connection cost to the utility, allowed many households to receive access to electricity. Vulnerable households can pay this connection fee overtime, further reducing the amount a household has to pay at anytime for an electricity connection. [8] conducted an experiment in Kenya where they offered randomized electricity connection prices to 1,139 households. This experiment showed that demand for electricity connections significantly decreased with price. While up-take was universal under the free grid connection strategy (high subsidy arm), even at a 29 % connection fee subsidy the uptake was significantly lower. This findings in conjunction with the observed very low electricity consumption in Rwanda suggests that while affordability may be a barrier to a grid connection, at such low connection fees of 56,000 RWF, many households who may not utilize grid services are being added to the grid.

Thus far, Rwanda plans to achieve universal electricity access with 48 % of its residential users utilizing off-grid systems. For low consumers such as those observed within the data, off-grid systems such as solar home systems which have a smaller cost of installation would better support their consumption needs. This is consistent with the notion that off-grid systems are necessary technologies in 'modern electricity service ladder' [41]. While grid services may not be fully replaced, these offgrid systems could provide affordable first-access to low consuming homes pending the time their demand for electricity grows.

Barriers to electricity consumption

Following the consumption growth analysis, there may be a number of reasons that explain the decrease in residential electricity consumption that occurs in more recently electrified cohorts. In this section, we introduce these hypotheses and present preliminary evidence for them, though we acknowledge that further surveys and analysis are imperative to validate them. In the first hypothesis, we postulate that low electricity consumption among households is a consequence of lower incomes. The Fifth Integrated Household Living Conditions Survey (EICV5) from Rwanda presents household demographic and socioeconomic variables for 14,580 nationally representative households [42]. Here we analyze the total consumption expenditures of households, which represents the monthly monetary amounts spent on goods and services. We use consumption expenditure as a proxy measure for wealth. Further inspection of grid connected households from the survey reveals that low electricity consuming households spend on average a total of 132,652 FRW per month, while non-low (medium and high) electricity consuming households spend on average a total of 396,595 FRW per month. [4]. The categories for low, medium and high electricity consumers align with those defined in our methodology section. The huge differences in overall household expenditures (which may be an indication of household wealth), suggests that the lower consuming households tend to be poorer. Additional surveys directly measuring household income or wealth, electricity consumption and household location would provide better insights into the electricity consumption constraints due to differences in wealth and affordability.

In the second hypothesis, we posit that lower electricity consumption could be tied to electric appliance ownership and use among households. Evaluating appliance ownership from the EICV5 survey suggest that there may be key differences in the appliances owned by different electricity consumption categories. The five most frequently used appliances in the survey were evaluated in light of the consumption categories. The most owned appliances within the survey include mobile phones, followed by radios, video/DVD players, computers, and refrigerator/freezer, respectively.

[4]The survey reported total expenditure reflects a household's monthly expenditure on goods and services. This accounts for food, rent, electricity, water and more, derived from survey responses.

Whie mobile phones and radios can be considered low-wattage appliances, video/DVD players, computers and refrigerator/freezer can be considered as high-wattage appliances. Overall, appliance ownership amongst medium and high customers is higher than appliance ownership in low consumers. While the ownership of high-wattage appliances is low overall there are significant differences between low and non-low (medium + high) electricity users. Specifically, while combined refrigerator/freezer ownership is 15 % in the medium and high consumers, less than 1% of low consumers own a refrigerator/freezer. A similar observation is made for the ownership of video/dvd players which is 42 % in the non-low group and only 11 % in low consuming customers. Computer ownership was also found to be 4 % and 22 % for the low and non-low groups, respectively. While this initial observation does not provide additional insights to the efficiency of appliances, it is an indication that higher consumers own more high-wattage appliances than low consumers. Affordability of these appliances may also influence their ownership as poorer households have less purchasing power to buy high-wattage appliances. Appliance financing, similar to endeavours carried out in the U.S. in the 1930s, may contribute to lifting the observed electricity demand in low consuming households. We hope to corroborate this initial analysis using more detailed surveys thus helping to better understand the relationship between electricity consumption and barriers to electricity consumption in Rwanda.

Quantifying Subsidies given Tariffs & Revenue

Through our analysis, we have shown that increasing the tariff in 2015 by 35 % resulted in suppressed consumption especially for the lower tier consumers. Nevertheless, there was a boost in revenue for the utility as a result of the tariff increase. To revive consumption, especially for low income households, a block tariff structure was introduced in 2017. This block tariff provided the "lifeline" tariff of 89 RWF/kWh for consumption less than 15 kWh. While the lifeline tariff provided more affordable tariffs to low income households, it also came at a cost to the utility in the form of decreased revenues. Households consuming more than 15 kWh/month also benefited from the "lifeline" tariff, further straining utility revenues. We observe from the customers within

our dataset that by introducing the block tariff, the utility experienced a drop in annual revenue from RWF 2.6 billion to RWF 2.3 billion. This significant drop in revenue further decreases the financial viability of the utility as it is already serving customers who use little to no electricity and pay a small fee to use it. Similar analysis can be carried out to quantify the potential subsidies that governments or institutions can offer utilities to better support utility operations while encouraging electricity consumption amongst poor and economically vulnerable households.

1.4 Summary

Developing economies are experiencing rapid increases in the number of households with grid-access. A decade ago electricity access in Rwanda and Kenya were 10% and 30%, respectively. In 10 years that number has more than doubled to 64.5% in Rwanda and over 70 % in Kenya. This acceleration towards universal electricity access has enabled many households to use electricity and has opened opportunities to improve their welfare. However, it has also revealed some important patterns in electricity consumption given the pace of electrification.

This work analyzes the dynamics of electricity consumption among these newly-connected customers in Kenya and Rwanda. The key finding of newer customers using less electricity and peaking sooner continues to highlight the importance of non-grid electrification technologies to support the initial low demand of newer customers. While reaching the entire population with some form of electricity access is a goal for all countries, it is vital to consider the challenges therein. If, as our results show, the expected consumption plateau is lower for newer customers, then the lowest-cost technology for initially providing electricity access to some customers, at least until the demand grows significantly, may not be grid power.

Beyond electricity consumption trends, this work also shows the impact of tariff changes on utility revenue and the amount of electricity used. The observed increase in electricity consumption for the lower consumers shows that a block tariff structure might be beneficial in making electricity consumption more affordable for the poorer households, however this comes at a significant cost to the utility, further placing strain on its ability to remaining financially viable. While universal

electrification is the goal, tariff structures that preserve consumption especially for the lower tier consumers, can promote the use of electricity but may require large monetary investments in the form of subsidies to the utility.

1.5 Appendix

1.5.1 Kenya Supporting Material

Kenya Classification

We applied a constrained k-means clustering method (Wagstaff et al., 2001) to identify three clusters (urban, peri-urban, and rural). We initially used two clusters, representing urban and rural areas, but discovered that the numerical uniqueness of the urban cores of Nairobi and Mombasa – with high population density and intense nighttime lighting – set those areas apart into their own cluster. The peri-urban surroundings of the cities and the rural areas were quantitatively more similar and therefore grouped into the same cluster. This does not agree with conventional definitions of urban areas. By identifying three clusters, the algorithm is able to separate these "peri-urban" areas from the rural areas, arriving at a much more justifiable classification. We also note that electricity consumption levels across all quartiles were largely similar between the urban and peri-urban categories, so we felt comfortable pairing these two clusters into a single category representing urban consumption. The constrained k-means clustering method works by exploiting accepted characteristics about urban areas, which is used to apply initial constraints on the clustering algorithm. For identifying these initial constraints, we leverage three methods for determining urban and rural locations from the literature. We use the spatial regions of overlap of these three methods to bootstrap our algorithm, effectively identifying consensus-urban regions. The methods include: 1. The Global Rural-Urban Mapping Project (GRUMP) (Socioeconomic Data and Applications Center, SEDAC, 2010), which combines census and satellite data to produce various datasets, including urban masks used in this analysis; 2. LADA Land Use Systems of the World data which provides 40 land-use classes for the world including urban areas (Land Degradation Assessment

in Drylands, 2010); and 3. The UN population estimate (The United Nations Population Divisions World Urbanization Prospects, 2010), which uses a national urbanization level (23.6% in the case of Kenya) to determine a threshold of population density at which to separate urban areas and rural areas. The constraints (consensus regions) describe which items in the dataset must be or cannot be "linked" (appear in the same cluster). These areas provide the initial conditions of the clustering algorithm, effectively bootstrapping the cluster definitions with areas that must appear in the same cluster. With this guidance for initial cluster relationships, the algorithm can then proceed to assign the remaining areas to any of the three clusters. To determine cluster membership, the algorithm uses features obtained from three 2010 data sources, all of which are available publicly and in a raster format at a maximum common resolution of 1 km × 1 km:

- Population Density via WorldPop (AfriPop, 2010);
- Nighttime Lights via the DMSP-OLS satellite imagery dataset (NOAA's, 2010);
- LADA Land Use Systems of the World data which provides 40 land-use classes for the world including urban areas (Land Degradation Assessment in Drylands, 2010)

Various methods for urban-rural classification in the literature employ one or two of these datasets, but we were unable to find any methods that used all three data sources. Prior to applying the clustering algorithm, the features are normalized by their mean and standard deviation. The algorithm is able to classify each 1 km × 1 km grid cell of Kenya as urban, peri-urban, or rural. Based on this classification, customers in our sample can be assigned to an urbanization level using the GPS locations of their electric meters. Table 1.7 compares our method under 2 and 3 clusters to the other urbanization methods, in classifying the total population of Kenya. We show that our method under 3 clusters better allows us to extract the most rural population of Kenya, compared to when we only apply 2 clusters. Although our method performs similar to existing methods when defining urbanization levels, our method offers a robust clustering approach because it leverages regions which existing definitions all agree to be urban, and uses these regions to initialize the clustering thereby providing a more trustworthy definition of urbanization. In Table 1.8 we also

show the performance of our method in classifying our study dataset of about 136k customers. The peri-urban customers defined in our 3 cluster approach tend to be carved from mostly the urban customers in a 2 cluster approach- although there are some from the rural cluster. This result aligns with our decision to group urban and peri-urban customer consumption while understanding the behavior of the most rural customers.

Urbanization Methods	Urban (%)	Rural (%)	PeriUrban (%)
Our Method (2 clusters)	15	85	NA
Our Method (3 clusters)	5.4	81.7	12.9
GRUMP	22.6	77.4	NA
Land Use Systems	11.9	88.1	NA
UN Population Estimate	23	77	NA

Table 1.7: Comparison of our clustering method with other definitions of urbanization, in classifying the total population of Kenya in 2010.

Urbanization Methods	Urban (%)	Rural (%)	PeriUrban (%)
Our Method (2 clusters)	32.9	67.1	NA
Our Method (3 clusters)	6.6	55	38.4
GRUMP	53	47	NA
Land Use Systems	22	78	NA
UN Population Estimate	46	54	NA

Table 1.8: Comparison of our method with other definitions of urbanization, in classifying the 136k customers in our sample, by urbanization level.

1.5.2 Rwanda Supporting Material

Sample Size Considerations

Similar to evaluations in [36], sample size evaluations where performed with the data from Rwanda. Months for which the sample size was less than 10% of the median sample size for a given connection year, were filtered out, resulting in over 10,000 data points at any given point in time for residential customers and over 1,000 data points at any given point for non-residential

customers. Figure 1.18 shows the sample sizes used for the descriptive analysis in Rwanda. These large sample sizes give confidence in the findings.

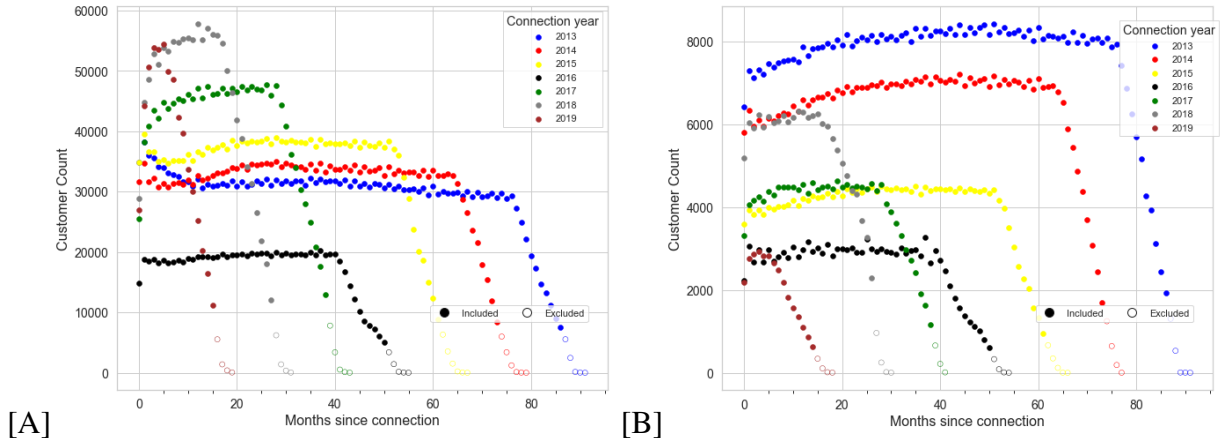


Figure 1.18: Monthly number of residential (**A**) and non-residential (**B**) customers in Rwanda, separated by electricity installation dates. The large number of customers, numbering in the thousands of bills, allows for confidence in the significance of our finding.

Validating the relationship between consumption growth and time

From our analysis, we observe that median electricity consumption initially increases then stabilizes, for each connection cohort. Table 1.9 shows the average monthly change in electricity consumption. We present regression results for each connection cohort when electricity consumption is regressed with time. The years indicate the coefficients for customers electrified in that year. For both residential and non-residential customers, we mostly observe statistically significant relationships, indicating that observations from the median customer figures are consistent across all customers in the cohort.

Table 1.9: Average monthly changes in electricity consumption with time for various connection cohorts, across residential and non-residential customers. Regression coefficients indicate statistically significant changes in consumption over time. Standard errors are reported in brackets.

	2013	2014	2015	2016	2017	2018	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Residential (kWh/month)	0.003^{*}	0.002	0.020^{***}	0.057^{***}	0.035^{***}	-0.041^{***}	-0.169^{***}
	(0.001)	(0.002)	(0.003)	(0.004)	(0.004)	(0.005)	(0.011)
Non-residential (kWh/month)	-3.802^{***}	-0.087	-4.758^{***}	-0.827^{**}	0.671^{***}	-2.022^{***}	0.195
	(0.182)	(0.068)	(0.345)	(0.360)	(0.190)	(0.175)	(0.714)

Note: *p<0.1; **p<0.05; ***p<0.01

Chapter 2: Predicting Levels of Household Electricity Consumption in Low-Access Settings

In low-income settings, the most critical piece of information for electric utilities is the anticipated consumption of a customer. Electricity consumption assessment is difficult to do in settings where a significant fraction of households do not yet have an electricity connection. In such settings the absolute levels of anticipated consumption can range from 5-100 kWh/month, leading to high variability amongst these customers. Precious resources are at stake if a significant fraction of low consumers are grid-connected over those with higher consumption.

Now suppose an energy planner or a utility is interested in predicting the anticipated electricity consumption of an unelectrified household, so as to prioritize areas for grid extension and off-grid systems. One approach might be to ask the household to enumerate the appliances that would be used upon grid-access. While this approach would provide the planner with a good sense of appliance ownership and thus the latent household demand, it requires the planner/utility to perform an extensive survey of all households before determining which areas to prioritize for grid extension versus off-grid systems. Deploying such extensive census is time-consuming and expensive. In reality, household censuses in many countries are performed decennially. Thus to support large-scale rapid and repeatable evaluation, alternative approaches to electricity consumption prediction have to be considered.

One alternative would be to utilize the large amounts of already collected utility consumption data to extract relevant features about electricity consumption. Planners can then develop models that attempt to correlate electricity consumption with features from widely available datasets such as satellite imagery. This approach lends itself very well to machine learning methods and remote sensed imagery. An electricity consumption prediction model that allows the discovery

of relevant household features from satellite imagery can be developed. This model can even be extending to incorporate and combine multiple data sources for feature extraction. An objective or loss function specifying the attribute for error minimization would also be required. In this case, the loss function would minimize the difference between the utility consumption data and the model predictions. Finally, a training scheme, which allows the model to learn the most relevant features from input data supports optimal model tuning. By combining these components, a large-scale, rapid approach to predicting electricity consumption can be developed in order to guide investments for grid extension and off-grid systems.

This chapter presents the first study of it's kind in low-income settings that attempts to predict a building's consumption and not that of an aggregate administrative area. We train a Convolutional Neural Network (CNN) over pre-electrification daytime satellite imagery with a sample of utility bills from 20,000 geo-referenced electricity customers in Kenya (0.01% of Kenya's residential customers). This is made possible with a two-stage approach that uses a novel building segmentation approach to leverage much larger volumes of no-cost satellite imagery to make the most of scarce and expensive customer data. Our method shows that competitive accuracies can be achieved at the building level, addressing the challenge of consumption variability. This work shows that the building's characteristics and it's surrounding context are both important in predicting consumption levels. We also evaluate the addition of lower resolution geospatial datasets into the training process, including nighttime lights and census-derived data. The results are already helping inform site selection and distribution-level planning, through granular predictions at the level of individual structures in Kenya and there is no reason this cannot be extended to other countries.

2.1 Introduction

Improved engineering and new business models for electrification have contributed to increasing access to electricity around the world. However, 840 million people still lack access to electricity services [4], many of them residing in places that are difficult to reach and, as a result, expensive to serve [43, 44]. Energy providers, constrained by limited investment budgets, face a perpetual

trade-off between expanding electricity access and cost recovery. When consumption levels are low, as can occur in low-income settings, utilities struggle to recover the cost of servicing a grid connection, and the government subsidies[40] for initial capital are poorly utilized. Alternatives to grid extension such as Solar Home Systems (SHS) can support smaller loads without the large wire investments, while in some cases clustered homes (with clusters far from each other) can make mini-grids viable[45]. In practice, identifying those likely to become high consumers is critical to the energy provider, as these are critical to revenue generation and system cost recovery. Given the diversity of electrification technologies, planners rely on energy access planning tools (e.g., the *Open Source Spatial Electrification Tool (OnSSET)*[46]) that utilize electricity consumption tiers, to match potential customers with technologies that can cost-effectively meet consumption. Consumption predictions can assist matching areas with cost-effective energy technologies, enabling a country to provide electricity access to a larger population given the same investment.

We make four unique contributions. *First*, we introduce a data-driven method to predict levels of future electricity consumption for individual households, using information prior to the household being electrified. Our approach trains a CNN to predict levels of household consumption using pre-electrificaion daytime satellite images. Although accurate individual household electricity consumption predictions are difficult to achieve [47], we show that high-resolution daytime satellite imagery (0.5 m/pixel) performs better (preserving performance at different levels of consumption) than other approaches (historical consumption, census indicators, and Nighttime Lights) that result in heavily-skewed prediction performance. *Secondly*, our proposed method shows that learning about buildings through a building segmentation task and over a large volume of images improves the downstream task of electricity consumption prediction. *Thirdly*, we demonstrate a method for model interpretation that quantifies the importance of building characteristics relative to the surrounding context. Specifically, we show that building roof sizes and color are relevant to predicting consumption levels. Our approach also shows that learning about the household's surrounding context improves prediction performance between 2-5% depending on the consumption tier.

Finally, we present additional validation of our results using the World Bank's Multi-Tier Framework survey of electricity consumption among households. Sample weighted Pearson correlation scores between the survey and our predictions for 5.3 million residential buildings were 0.82 when excluding the over-sampled and already-electrified capital city of Nairobi, and 0.64 otherwise. Outputs from our model can be used in planning tools such as OnSSET, which utilize electricity demand tiers as an input parameter for spatial electrification planning. Given the potential dependence of consumption on tariffs and policies for recovery of installation costs, the specific results of this data-driven approach apply to Kenya. However, our methods can be extended to other countries, thereby offering insights to electricity planners.

2.2 Related Work

Predicting Electricity consumption: [48, 49] present a comprehensive survey on residential energy consumption prediction. First, we review load forecasting in time for individual households, highlighting how the problem at hand is different. One approach [50] forecasts individual household level electricity loads 24 hours ahead using sequence mining and smart meter data. Another [51] clusters customer smart meter data into behavioral groups and later use supervised techniques such as Random Forest to predict customer clusters given unseen smart meter data. Other residential load forecasting studies [52, 53, 54] also predict the short-term future consumption of houses given their historical data or appliance usage, and deep learning techniques. Past consumption data is essential for such studies and hence not suitable for future consumption prediction where no prior data exists. Some studies have attempted to address the problem of predicting the future consumption of a currently unelectrified household. In [55] the average consumption of previously connected customers (by municipality) is used to estimate consumption for unelectrified households. This would not capture the variations amongst households. [56] use support vector regression to study the relevance of 48 household survey variables (demographics, appliance ownership, household personality traits) in predicting household consumption. [57] use an energy end-use model to estimate demand for off-grid communities in Myanmar, Indonesia and Laos

through household surveys that measure appliance ownership and usage. [58] take a similar approach to estimating residential electricity consumption in Nigeria by collecting survey responses on appliance ownership and usage. [59] use machine learning to predict daily electricity consumption tiers upon connecting to a microgrid, using features obtained from customer application surveys pre-electrification. All of these studies use data that would be difficult and/or expensive to obtain at scale for a country. Our approach provides a scalable and faster approach to estimating consumption from proxy household features which are available in satellite imagery.

Satellite Imagery and Machine Learning: Recently, there has been a surge of studies applying CNNs to satellite imagery to assess building damage [60], measure road quality [61], detect solar farms [62], segment roads and buildings[63], estimate rooftop density by type [64] and measure poverty. One approach [65] predicts wealth for multiple African countries by combining overhead daytime images with CNNs. The authors use high resolution daytime images in training a CNN to predict nighttime lights; features extracted from the trained model were then used to estimate household expenditure and wealth at a 10 x 10 km resolution. Their results suggest that predictions about economic development can be made from satellite image derived features; this insight provides additional motivation for developing methods that extract information from imagery for electricity consumption prediction. Building on [65], multiple works [66, 67, 68, 69, 70, 71, 72] have assessed wealth, poverty and development using satellite images. [73] use VIIRS nighttime lights, gridded population data and land cover to estimate binary electricity access rates and electricity consumption tiers at 1 x 1 km grids. These studies demonstrate the value of satellite imagery in serving as a proxy measure for varying features such as poverty, electricity access and consumption. However, all these studies are carried out at a larger spatial scale to preclude evaluations of poverty levels or electricity consumption tiers of individual households[1]. [47] is the only study to the best of our knowledge that predicts individual building energy consumption using overhead imagery in Gainesville, Florida, and San Diego, California. While performance improves at a spatially aggregated level, at the individual building level, they report low correlation (r^2=0)

[1]Household consumption tiers are relevant for electrification planning

between predictions and the training data in Gainesville. The context of this study is quite different given the consumption levels, the size and formal construction practices in the U.S.. Also, the problem they address is that of estimation for electrified households rather than consumption prediction for unelectrified households.

To the best of our knowledge, ours is the first study of its kind that predicts electricity consumption at an individual household level using overhead imagery. We formulate our task as a classification rather than a regression problem, and provide model interpretation around learned features.

2.3 Models

2.3.1 Problem definition

Given a set of households found in buildings $\mathbf{B} = \{b_1, b_2, b_3 ... b_n\}$, where each building has a corresponding satellite image prior to the household being electrified $\mathbf{X} = \{x_1, x_2, x_3 ... x_n\}$, the objective of our proposed model $\mathcal{F}(x_i)$ is to use each building's corresponding satellite image pre-electrification, to predict its consumption class ($\hat{y}$) after the building has been electrified (i.e. $\hat{y_i}$= $\mathcal{F}(x_i)$). Data from electric meters after electrification serve as ground truth values. Binary labels y_i are obtained by applying a threshold *thres* (e.g. <= *thres* kWh/month) to the average monthly consumption values of each individual household given its electric meter readings.

We propose a two-phase supervised method to predict binary consumption classes ($\hat{y_i}$). First, we prioritize learning about building features through the help of a building segmentation task. Next, the building segmentation model is used to initialize a supervised model to classify consumption levels. Figure 2.1 illustrates the steps used for training, and their corresponding losses. Our method is compared to other models that match commonly used approaches to predict electricity consumption levels after electrification.

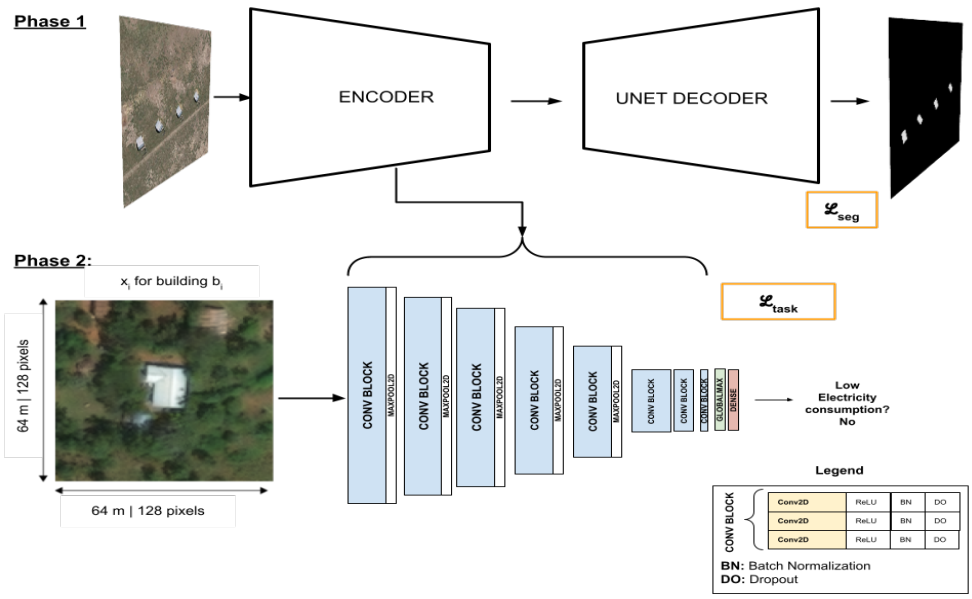


Figure 2.1: Phase 1: A building segmentation model is trained using the encoder in Phase 2 and a UNET decoder. The segmentation model is trained with a dissimilarity loss ($\mathcal{L}_{seg}$). Skip connections are omitted to maximize information funneling. Phase 2: The pretrained encoder is used in phase 2 to learn the electricity prediction task. An image (x_i) containing a household's building is input into the pretrained encoder. This encoder is trained with the negative log-likelihood ($\mathcal{L}_{task}$) loss to predict electricity consumption levels upon electrification.

2.3.2 Electricity Prediction Models

Here we present 4 different approaches to predicting electricity consumption. The first two are widely used approaches, the third is our proposed method, and the fourth is an approach to support the interpretation of our method.

Model A: Average Historical Consumption

Model A represents the simplest model and serves as a baseline. This method mimics approaches commonly used by energy planners to estimate consumption for the unelectrified. Energy providers have historical consumption data for already-connected households. Thus, this model assumes that the average historical consumption of households in the same administrative unit is sufficient to approximate the consumption of customers that will be electrified in the future. In the case of Kenya, energy planning is done at administrative levels (following Kenya's policy of decen-

tralized energy planning). For each connection year t_k, households electrified prior to t_k are used to calculate the average monthly consumption ($\bar{y}(c_j, t_k)$) of each constituency c_j. A constituency is an administrative unit and there are 290 constituencies across the country. Households who share the same constituency and electrification year are assigned the same $\bar{y}(c_j, t_k)$ value. A threshold value *thres* (e.g. <= *thres* kWh/month) is applied to the assigned consumption of every household to determine the expected consumption class ($\hat{y}_i$). $\hat{y}_i$ is compared to the true consumption class y_i.

Model B: MLP with Non-Visual Data

Varying lower-resolution datasets are widely available and can serve as proxies for electricity access. In fact [73, 72] use non-visual data to evaluate electricity access and economic development. We present *model B*, based on publicly available non-visual datasets to evaluate their performance in predicting consumption levels upon electrification. *Model B* offers more complexity than an average historical consumption strategy. Non-visual features are inputted into a Multi-Layer Perceptron (MLP) containing 3 dense layers with 64, 32, and 16 filters respectively (Appendix 2.7.1). The model is trained by minimizing the Negative Log-Likelihood loss as shown in Equation 2.1.

$$\mathcal{L}_{task} = \sum_{y} -\log(p(y_i; \theta)), \tag{2.1}$$

θ and y_i are the model weights and consumption labels.

Model C: Building Characteristics and Context

Model C combines both information about building characteristics with information about the surrounding context. In this model both the building of interest (b_i) and its surrounding context (in the form of a 128 x 128 image patch pre-electrification) are used to predict consumption levels post-electrification. Electricity consumption levels post-electrification are learnt in two phases. First we train an encoder-decoder building segmentation model to learn relevant building features. Next, we extract the trained encoder, add a classifier head and use the learnt building weights to initialize the consumption prediction task. Below we present a description of each phase as shown

in Figure 2.1.

Phase 1: Learning about buildings: Deep learning has been shown to thrive in the presence of large amounts of labels. Although our electricity billing dataset is the largest of its kind (i.e., in a similar context) ever studied, its size remains small relative to the amounts frequently used to train data-hungry CNNs. We hypothesize that learning a proxy task (such as building segmentation) could provide relevant image encodings for predicting levels of electricity consumption, especially when small numbers of labels are available. We employ a much larger dataset of 6,928,078 building footprint polygon geometries in Uganda released by Microsoft[74] for building segmentation. Building polygons from Uganda are used because there is no large high quality building footprint data in Kenya, and Uganda is the closest geographic country to Kenya with building polygons.[2] Noisy (misaligned or missing) building polygons were observed within the Microsoft data in some parts of Uganda. Nevertheless, RGB patches of 128 x 128 pixels were used to train the building segmentation model in Uganda. We combine a custom encoder with a UNET-decoder to perform building segmentation (Figure 2.1). This encoder architecture is used both as an encoder for building segmentation and as an encoder for the classifier in phase 2. This architecture was inspired by the DeepSense architecture[76] and has been shown to be helpful in remote sensing applications such as building segmentation. Skip connections between the encoder-decoder are excluded to maximize information funnelling through the encoder during phase 2. 64 filters were used in each layer of the encoder. The building segmentation model was trained with a dissimilarity loss ($\mathcal{L}_{seg}$) as shown in Equation 2.2, which builds on the Jaccard index $\mathcal{J}$(U,$\hat{U}$).

$$\mathcal{L}_{seg} = 1 - \mathcal{J}(U, \hat{U}) = 1 - \frac{(U \cdot \hat{U}) + \epsilon}{(U + \hat{U} - U \cdot \hat{U}) + \epsilon} \tag{2.2}$$

where U represents the true footprints, $\hat{U}$ represents the predicted footprints and ϵ is used for numerical stability. The learnt encodings are later used in the downstream consumption level prediction task to bootstrap the classifier.

Phase 2: Predicting electricity consumption levels: After training the building segmentation

[2]This work was done prior to the release of Google Footprints [75]

model using Uganda data, the encoder-decoder network is initialized with the best building segmentation weights. The encoder is extracted and merged with a classification head (consisting of a global max-pooling and a dense layer) to predict consumption levels. The image patch is fed into the encoder with the classifier head, which outputs the predicted class ($\hat{y}_i$), and is trained with $\mathcal{L}_{task}$. Data augmentations (e.g vertical and horizontal image flips, 90 degree random rotations and 15% zooms) are performed during training.

Model D: Building Characteristics Only

The goal of *Model D* is to provide additional interpretation around the black box CNN in *Model C*. Rather than evaluate the whole image, this model aims to evaluate the importance of **only** roof characteristics of the building of interest, while ignoring the surrounding context of the household. To achieve this, *Model D* utilizes only building roof characteristics (area and type) as predictors of consumption levels. Specifically, building roof area and the RGB 3-channel intensities are extracted and used as features for prediction. Building roof area and color are inputted into the previously defined MLP to predict consumption levels (Appendix 2.7.1). The MLP is also trained with $\mathcal{L}_{task}$.

Roof-top area extraction: The point indicator approach proposed by [77] is chosen over conventional segmentation because the building polygons available for segmentation (Microsoft Building Footprints in Uganda [74]) suffer from misaligned and omitted labels when compared to our satellite images. First we select only polygons that are well aligned with structures in satellite imagery. The well-aligned polygons together with the Pointer Segmentation Network[77] are used to train a segmentation model that learns when some of the instances within the images are omitted. This segmentation model was also trained with the dissimilarity loss ($\mathcal{L}_{seg}$). After training on Uganda, we also generate 1000 hand-labelled footprints in Kenya and use the small sample from Kenya to tune the Pointer Segmentation Network. Once the model is tuned to Kenya, the GPS locations of the buildings (b_i) in our dataset, are used to obtain a point within each image (x_i). This point when combined with the tuned Pointer Segmentation Network is used to extract the footprint

for building b_i. The extracted footprint is then used to crop out the pixel intensities of the roof. We assume building roofs have a uniform color, thus the roof pixel mean for each channel is used in addition to the roof area (obtained from images pre-electrification) as input features to predict consumption levels after electrification.

2.4 Data

The dataset used in this work has 3 components: 1) Monthly post-paid electricity bills, 2) Overhead daytime satellite imagery, and 3) Public data sources. We unify these 3 data sources by matching the billing dataset to images or public data sources using customer locations within the billing dataset. Following are some details about each.

2.4.1 Ground truth electricity data

Previously[40], we conducted a longitudinal study of 100k+ randomly sampled electrified households, observing that median customers in Kenya typically reach a consistent level of electricity consumption roughly 12 months after receiving an electricity connection. Given this observation, we define the average monthly consumption of a household after 12 months of a connection as the *expected stable electricity consumption*. For each household, all bills after one year of connection are averaged to obtained a single stable estimate of electricity consumption. The World Bank's Multi-Tier Framework (MTF) divides electricity consumption into a series of Tiers, based on levels of electricity services. We consider low levels of consumption as corresponding to Tiers 0 - 2 of the framework while high consumption levels correspond to >= Tiers 3. Our levels of consumption are obtained by placing a threshold (*thres*) at 30 kWh/month. Figure 2.2 illustrates our levels of consumption relative to the MTF tiers. We select a 30 kWh/month boundary because it aligns with MTF break points and energy access practitioners rely on the MTF tiers to support spatial electrification planning. Rather than defining the binary class with low being <= 30 kWh and high being > 30 kWh, we select a discontinuous boundary where stable monthly consumptions (kWh)

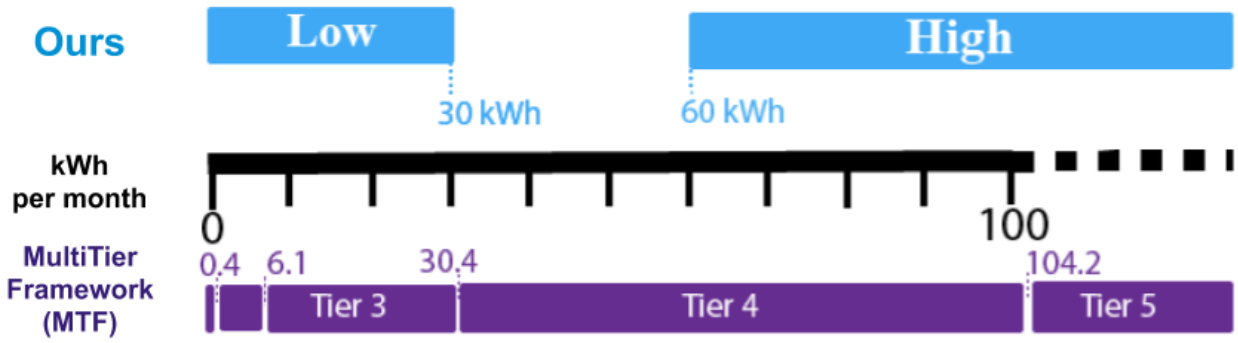


Figure 2.2: Illustration of World Bank Multi-Tier Framework Consumption Tiers relative to our levels of consumption .

<=30 kWh are considered low while >= 60 kWh are considered high [3]. To develop a matched dataset of bills and images, customers are grouped by location to obtain electrified buildings. We select residential buildings with only one customer account and these buildings are matched to contemporaneous daytime satellite imagery. From 135,702 Kenya Power customers, 52,083 single customer buildings are used to calculate the monthly stable consumption of each customer. Single customer buildings are chosen over multi-customer buildings because our billing dataset does not contain all the customers in each multi-customer building. Keeping in mind our goal of predicting expected levels of consumption upon electrification ($\hat{y}_i$) using images pre-electrification (x_i), satellite image acquisition dates are used to select buildings with satellite imagery acquired prior to the stable consumption phase. Our selection assumes that the socioeconomic benefits of electrification do not become apparent within a daytime satellite image immediately (< 1 year) after the household is electrified. Labels are obtained by applying the discontinuous threshold, to obtain binary consumption levels.

[3]Customers between 30-60 kWh represent a harder set to study given that we use proxy measures (building characteristics) from satellite imagery. The disjoint boundaries enable electricity planners to still identify customers with low electricity consumption (<=30 kWh) to target lower-cost electricity technologies and customers, while also enabling planners to target customers likely to have high electricity consumption (>=60 kWh) for more traditional grid-based connections. This design choice, made with significant input from electricity system planning practitioners, supports the twin goals of enhancing the financial sustainability of electricity providers and preserving model performance for the relevant task.

Table 2.1: Non-visual data used for electricity prediction.

Census (% of ward)
Water Source (Surface | Improved | Unimproved)
Sanitation (Improved | Unimproved)
Lightfuel (Finished | Rudimentary)
Floor material (Finished | Rudimentary)
Cook fuel (Finished | Rudimentary)
Wall material (Finished | Rudimentary | Natural)
Rooftop material (Finished | Rudimentary | Natural)
Intensity
VIIRS Nighttime lights

2.4.2 Satellite Imagery

Satellite imagery used in this work consists of 3-band (RGB) 50 cm daytime DigitalGlobe Satellite Imagery obtained between 2002 and 2020. The DigitalGlobe imagery while providing country-wide coverage only contains a single image per location (there are no temporal images for the same location). To train the building segmentation task, images with corresponding building polygons were used irrespective of the image acquisition date. To predict electricity consumption levels, buildings whose images (x_i) occurred pre-electrification are selected as part of the training, validation and test datasets.

2.4.3 Public data sources

Census Information: The 2009 Kenya census [23] provides low-resolution demographic information on households at the ward administrative level, for which there are 1450 wards in Kenya. The 2009 census is selected over the more recent 2019 census because the recent census data are not yet publicly available and also occur significantly after our electricity consumption data. In addition, the 2009 census better aligns with our formulation for latent electricity prediction using data the occurs prior to when the household was electrified. Table 2.1 shows a summary of parameters obtained from the 2009 Kenya census, grouped by semantic meaning. The census reports the % of households in a ward for every category. Seventeen census indicators were used as additional data. Customers in the same ward are assigned the ward census value.

Intensity: 15 arcseconds/pixel (450m at the equator) VIIRS Satellite Nighttime Light data [78] is often used to study economic development and electricity. Average monthly nighttime light intensities for every year (2012 - 2015) were calculated using monthly VIIRS composites. The nighttime light intensity for the year prior to when the building was electrified is retrieved, for the grid cell in which the building is located. If the building was electrified before 2012, the 2012 intensity is used, as VIIRS composites are only available after 2011.

2.5 Experiments & Results

2.5.1 Experimental Setup

After matching satellite images pre-electrification to the mean electricity consumption of households in the stable phase, the datasets consist of 20,000 individual households. 75 % was used for training, 15 % for validation and 10 % were held-out as the test set. The distribution of the overall electricity data is preserved within each sub-group of train, val, and test. Results are reported for the 10 % in the held-out test set. All models were trained with an Adam optimizer with a learning rate of $1e^{-5}$. This learning rate was chosen over others ($1e^{-3}$, $1e^{-4}$ and $1e^{-6}$) as it offered the best overall performance and training convergence. The building segmentation model in Phase 1 was trained for 30 epochs (as both the train and validation curve had converged). The MLP models are trained for 20 epochs and the CNN model is trained for 100 epochs. A batch size of 64 was used and 25% dropout was applied on all models to prevent overfitting. Feature standardization and normalization was performed. We used an input patch size of 128 x 128 pixels to provide enough field of view that captures the building in the centre and some context around it.

2.5.2 Performance Evaluation

Table 2.2 shows the performance of each of the four models presented in Section 2.3. Our evaluation metrics include: 1) Class Accuracies shown as True Negative (TN - low consumers) and True Positive (TP - high consumers) 2) Equally weighted F1-score, and 3) Area-Under-Curve (AUC).

(A) Average Historical Consumption

Using average historical consumption levels as predictors for yet-to-be connected customers results in a highly-skewed prediction (0.35 F1-Score), with 99% of high consumers correctly predicted while only 2% of low consumers are correctly predicted. The strong performance skew is because of the electrification bias, where high consumers (who are often wealthier) are electrified first while lower consumers are added over time. The average historical consumption will always over estimate the consumption levels of the newly connected (often lower consuming) customers. An energy planner using administrative level averages will spend large investments to connect low consumers via grid when cost-effective alternatives might be more suitable.

(B) Non-Visual Data

Census indicators offer a range in F1-scores (0.57 - 0.65), with the highest obtained from rooftop materials. This suggests that building characteristics are important proxy features for predicting consumption levels upon electrification. Census parameters while performing better than *Model A*, also show a performance skew towards the lower consumption class. Nighttime lights only achieved a 0.51 F1-score in predicting individual consumption levels of future electricity connections. Overall, when all census and nighttime light features are combines F1-scores and AUCs are still below that obtained with images.

(C) Building Characteristics and Context

Using only daytime satellite images as the basis for prediction, our approach (*Model C*) achieves an balanced F1-score of 0.68 with an AUC of 0.75. This image-based model ensures good performance in both classes (70 % and 66 % correctly predicted as low and high respectively). Good performance in both classes is crucial for energy planners (especially in highly heterogeneous regions) and suggests that images better support local class differentiation, compared to the other lower resolution data sources. Our CNN architecture performs comparable to well-known architectures such as VGG16[79] and ResNet50[80], even though our custom encoder only has 728k

Table 2.2: Comparison of electricity prediction models in Kenya. Area-Under-Curve (AUC) & Balanced F1-score metrics are presented. True Negative (TN) shows the fraction of low consumers that were correctly predicted while True Positive (TP) shows the fraction of high consumers that were correctly predicted.

Model	**Method**	**Data Input**	**AUC**	**F1-score**	**TN**	**TP**
A	Historical Consumption	Average kWh	NA	0.35	0.02	0.99
B	Census	i) Water Src.	0.69	0.63	0.82	0.47
		ii) Sanitation	0.62	0.57	0.63	0.51
		iii) Lighting Fuel	0.68	0.63	0.82	0.47
		iv) Floor Mat.	0.67	0.61	0.84	0.41
		v) Cooking Fuel	0.67	0.60	0.86	0.39
		vi) Wall Mat.	0.66	0.63	0.69	0.57
		vii) Rooftop Mat.	0.69	0.65	0.66	0.64
B	Nightime Lights	VIIRS	0.52	0.51	0.77	0.30
B	Census & Nighttime Lights	i)- vii) and VIIRS	0.65	0.65	0.75	0.55
C (Ours) Building Seg. Weights without Contrastive loss	**Building Characteristics & Context**	**Images**	**0.75**	**0.68**	**0.70**	**0.66**
C (Ours) Building Seg. Weights with Contrastive loss	Building Characteristics & Context	Images	0.73	0.67	0.66	0.67
D	Building Characteristics	Roof Area	0.65	0.61	0.66	0.56
		Roof Color	0.66	0.62	0.56	0.68
		Roof Area & Color	0.69	0.64	0.65	0.64
B & C	Building Characteristics & Context, Census Nighttime Lights	Images, i-vii, VIIRS	0.77	0.70	0.76	0.65

trainable parameters compared to millions in VGG-16 and ResNet-50 (Table 2.3).

The classifier encoder was pretrained on a building segmentation task. The value of this pre-training step is validated by 2 approaches. First, we compare classification performance with and without pretraining and noticed that pretraining the encoder on building segmentation improves the electrification classifier accuracy from 0.63 to 0.68 when all the electricity training data is used. When the training data is reduced, performance is preserved for the building segmentation-

Table 2.3: Performance comparison of well-known architectures compared to our encoder

	Weights	F1-score	# Parameters
VGG16	Random	0.62	14,714,688
Resnet-50	Random	0.62	23,587,712
Our Encoder	Random	0.63	728,0065

pretrained model when compared to a model trained from scratch (Appendix 2.7.3). Second, we also compare the performance with and without a supervised constrastive loss[81] when initialized with building segmentation weights. Here, we hypothesize that if relevant embeddings are obtained through building segmentation, further optimizations of the embeddings (through a contrastive loss) would provide no additional performance gains. The classifier (initialized with building segmentation weights) was trained with a supervised contrastive loss (temperatures: 0.08 and 0.1) prior to finetuning the final layer for classification. Adding a contrastive loss did not further improve performance (0.67 F1-score), suggesting that the building segmentation task learnt relevant embeddings needed for the classification task.

Combining visual and non-visual features (*Model B & C*) through a multi-modal architecture (Appendix 2.7.2), increased the F1-score to 0.70. Using multi-modal data can be helpful to improve electricity predictions.

The image model was evaluated on households with monthly consumption between 31-59 kWh. We observed a 4% decline in F1-score when a threshold of <=30 kWh and > 30 kWh is used for low and high, respectively. [4]

2.5.3 Model Explainability

In this section we explore quantitative and qualitative approaches to uncovering the relevant features learnt by the CNN when predicting consumption levels from satellite image. We present three model explanability approaches (1 quantitative and 2 qualitative), with the goal of shedding some light into the black box CNN models. The first approach evaluates the performance of a machine learning model that takes in only building characteristics and outputs consumption levels. This quantitative approach measures the amount of relevant information held in building roof characteristics only. The second approach applies a Gradient-based Class Activation Map (Grad-CAM) that highlight portions within the input image responsible for the predictions. The third approach uses Generative Adversarial Networks (GANs) to tease out human-interpretable features that lead

[4]This set make up 28 % of single household customers within our data.

Figure 2.3: Sample segmentation outputs using an indicator point to specify which building(s) to segment[77]. White dots show input points given to the model to specify which buildings to segment. Green shows predictions and blue ground-truth.

to class differentiation. We present an in-depth discussion of each of these approaches.

(D): Building Characteristics Only (Quantitative Interpretation)

Complementary to the CNN, *Model D* isolates and quantifies the relevance of building characteristics (**only roof-top area and type**) when learning to predict consumption. Prior to discussing *Model D*'s performance, we first discuss the performance of the pointer segmentation model used to obtain building footprints. **Performance of rooftop segmentation:** Hand-labelled polygons in Kenya showed a validation Intersection-Over-Union (IOU) of **0.54**. Figure 2.3 shows sample segmented footprints in Kenya given indicator points (white dots) specifying buildings. This segmentation model was applied on the electricity training, validation and test set to extract building footprints (roof area) and average roof pixel intensities for each channel (rooftop type).

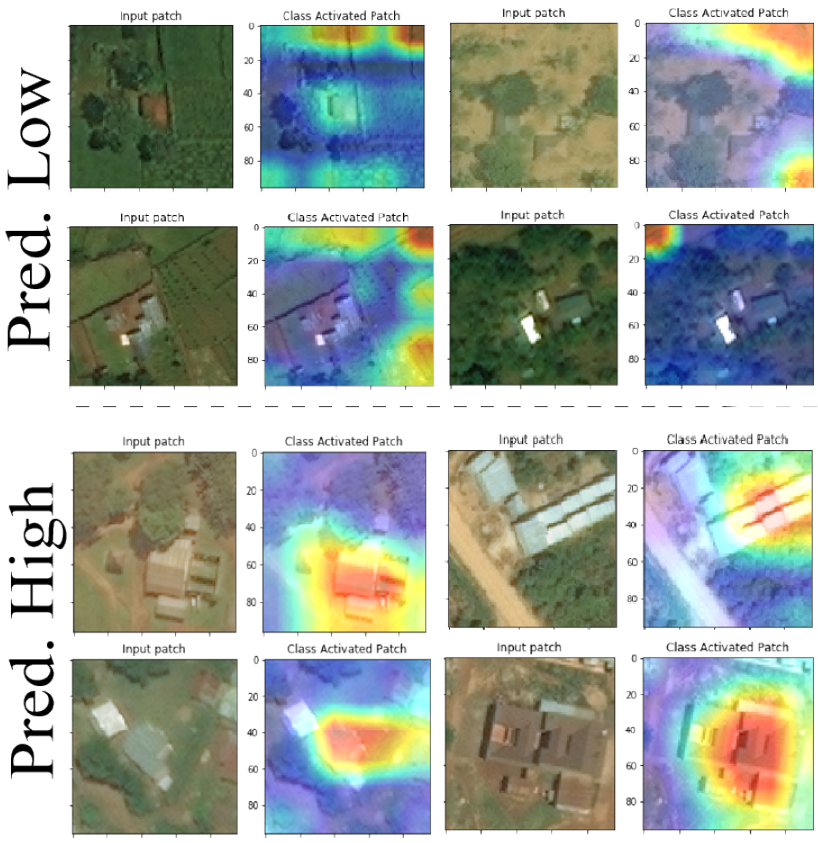


Figure 2.4: Gradient-based class activation maps for sample in test set. Stronger neural activations are in Red while weaker neural activations are in Blue. Buildings are strongly activated when predicting high levels of consumption while the activation is more distributed between the building and its surrounding context when predicting low levels of consumption.

Model D Performance: Roof area and roof color features respectively predicted 66% and 56% of the low class correctly while respectively predicting 56% and 68% of the high class correctly. However, combining both roof area and color reduced the skew in performance while encouraging better predictions for both low and high levels of consumption. This suggests that individual roof sizes may be more indicative of low consumers while roof materials (from mean pixel intensities) are helpful for better identifying high consumers. Direct use of images, which includes both the building characteristics and the surrounding context of the building improves the F1-score (relative to using only building characteristics) by 4% . This added benefit is likely a combined effect of bypassing segmentation error and the additional information found within the surrounding context of the building.

GRAD-CAM (Qualitative Interpretation)

Class activation maps are used to qualitatively explain the relevant features that the image-based classifier in *Model C* uses to predict levels of electricity consumption. GRAD-CAM [82] is used to visualize portions of the image that have high neural activations when making predictions. Some GRAD-CAM visualizations are shown in Figure 2.4. Strong activations (red) on buildings are observed when predicting high-consuming buildings while the activation is more distributed between the building and its surrounding context (blue) when predicting low-consuming buildings. The image-based model utilizes both building size and surrounding land as indicators of consumption levels.

Visualizing Explanatory Features for Class Differentiation (Qualitative Interpretation)

One might ask the question, why does a model classify one image as low rather than as high? More formally, the question asked might be: what high-level features is the model using to discriminate between low and high electricity users. To answer this question, we utilize decision boundary crossing transforms specifically in the form of an unpaired image-to-image Generative Adversarial Networks (GAN). GANs can be used to identify and visualize features that impact classification decisions, as they allow a user to inspect how the addition of certain feature vectors cause an image to fall on the other side of a decision boundary [83].

Consider an image x_i belonging to an area with low electricity consumers and an image y_j belonging to an area with high electricity consumers. In this context low and high are defined by user thresholds, less than t_l and greater than t_h, respectively.

Given these unpaired image examples [5], the goal is to learn distinct, localized features that shift x_i from being classified as a low consumption image, across the classifier decision boundary to being classified as a high consumption. Figure 2.5 illustrates the generative decision boundary mapping functions G and F that will create the features of interest which lead to a change in the

[5] A place cannot be both low and high consuming at the same time, thus there are no two class labels defining the same place.

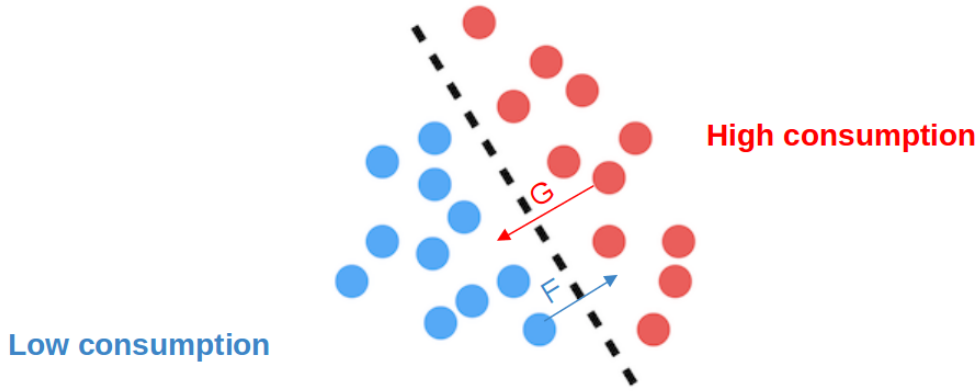


Figure 2.5: Illustration of decision boundary transforms G and F that transform an image from a given class across the decision boundary to a new class. G transforms images of high consumption areas to that of low consumption areas, while F does the reverse.

classification decision. In this work, we use two generators to create decision boundary crossing transforms: G which generates $\hat{y}_i$ given x_i ($G : X- > \hat{Y}$) and F which generates $\hat{x}_j$ given y_j ($F : Y- > \hat{X}$). The quality of the generated images are evaluated by two discriminators. The goal of the generators are to fool the discriminators into thinking the generated images are real. This approach can be generalized for multiple classes, showing that a tune-able GAN can be created to understand relevant features for multiclass classification (Refer to [84]).

Similar to [85], this work uses an adversarial loss L_{gan} and a cycle consistency loss L_{cycle}. These two losses enable both the discriminators and generators to learn from each other while prioritizing image quality. To ensure that distinct, localized features are generated an illumination control loss (L_{illum}) is added. This loss prevents the generators from merely increasing or decreasing greenness to generate images in the alternate domain, instead incentivizing distinct image changes while any global change in the pixel values is kept minimal. Each loss is weighted by some λ, thereby controlling the contribution of each loss to the total model loss. The losses are presented below:

$$L_{gan}(G, D_y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log(D_Y(y))] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_y(G(x))] \tag{2.3}$$

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[|F(G(x)) - x|_1] + \mathbb{E}_{y \sim p_{data}(y)}[|G(F(y)) - y|_1]) \tag{2.4}$$

$$L_{illum}(G, F) = (\sum \mathbb{E}_{x \sim p_{data}(x)}[|G(x) - x|_1] + \sum \mathbb{E}_{y \sim p_{data}(y)}[|F(y) - y|_1]) \tag{2.5}$$

$$L_{total} = \lambda_1 * L_{gan}(G, D_y, X, Y) + \lambda_2 * L_{gan}(F, D_x, X, Y) + \lambda_3 * L_{cyc}(G, F) + \lambda_4 * L_{illum}(G, F) \tag{2.6}$$

Figure 2.6 presents the results for the model implementation, when only the low and high class are used to train the model. In transitioning from a low-consumption image to a high-consumption one (top set of images), the method performs two primary localized and distinct changes: road and building footprints are both enlarged and brightened. These changes have the effect of making roads and buildings stand in sharper contrast to background features. These results make intuitive sense, as the lighter roads in the transformed images look to have a higher quality than roads in the original one, indicating more development in the generated image, which usually corresponds to higher electricity consumption. Similarly, tin roofs are typically seen as higher-status home improvements, and making the upgrade from a thatched roof to a reflective one likely parallels an increase in electricity consumption for a particular household. Transforming from the high-consumption class to the low-consumption class (bottom set of images) largely makes the inverse changes to the input imagery: road and buildings footprints are dimmed and made to blend in with their surroundings. These generated images on average look more rural than their high-consumption counterparts. Results from this approach agree with our previous findings and intuition that while buildings contribute to driving predictions, other contextual features such as

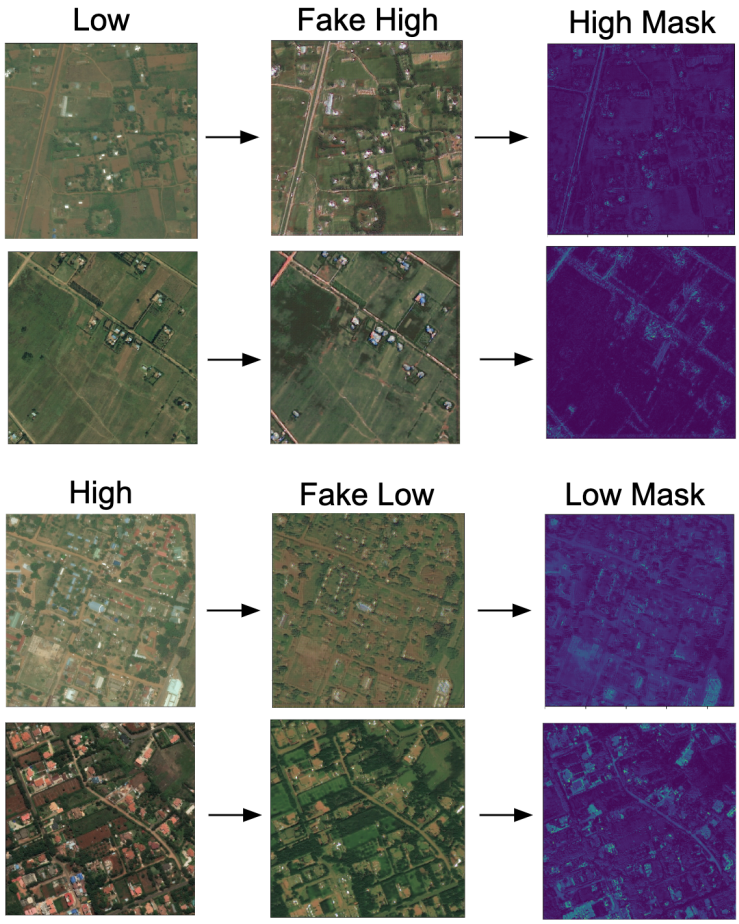


Figure 2.6: Results from the binary CycleGAN. The first set of transformations take a low-consumption image and transform it into a high-consumption image; the second set of image does the inverse. Left to right, the columns indicate the original image, the transformed image, and the absolute value of the transformed image minus the original image.

road presence and quality, and surrounding vegetation also provide guiding signals to the CNN.

2.5.4 Validation with independent survey data

We present an extra validation of our approach against an independently collected and nationally representative baseline household survey (4473 households) of both electrified and unelectrified households[86]. The Kenya Multi-Tier Framework (MTF) Survey conducted between 2016 - 2018, asks grid-connected households how much electricity they consumed in the most recent month. The reported and binned onetime consumptions are compared with our predictions for Kenya Power residential grid-connected customers. Images alone are used to predict consumption

Table 2.4: County-level consistency between independently collected Multi-Tier Framework (MTF) Survey and predictions for 5.3 million Kenya Power residential customers. Results (p-value <0.0005) for counties with at least 15 MTF samples.

	29 Counties	28 Counties (Excluding Nairobi)
Pearson correlation	0.64	0.82

levels of 5.3 million Kenya Power residential customers connected by the start of 2016. MTF samples are binned at <=30 kWh as low and >=60 as high. Sample weighted Pearson correlations (Table 2.4) between MTF and predicted consumption levels are reported for 29 counties with at least 15 MTF survey samples of grid-connected customers. Using all 29 counties, a correlation of 0.64 is observed. Excluding Nairobi county increases, the correlation to 0.82[6]. These correlations show strong agreement (p <0.0005) given an independent source of national data.

2.5.5 Country-wide predictions

After training, we inferred consumption levels for 11.9 million buildings in Kenya using building GPS locations collected as part of the Kenya National Electrification Strategy - Structures Survey. GPS locations and corresponding image patches are used to predict consumption levels for all buildings. Statistics for each predicted level of consumption are reported in a 6-band TIF for Kenya at resolutions of 250m, 500m, 1000m and 10,000m. Band 1 shows the predicted number of buildings with low levels of consumption, band 2, the mean predicted probabilities for band 1 and band 3, the standard deviation of prediction probabilities. Bands 4 through 6 capture similar information as the first three but are for high levels of consumption. The Kenya map in Figure 2.7 shows predictions (aggregated at 250m) for the 11.9 million buildings. We show the fraction of buildings that have low levels of consumption. This is obtained by dividing band 1 in our generated TIF by the sum of band 1 and 4. Blue shows regions where more buildings have low levels of expected stable consumption, while red shows regions where more buildings have high levels of expected stable consumption. Because our training data is a sample of consumption data – there

[6]Nairobi (the largest city in Kenya), is excluded because the survey over-samples recently-electrified informal settlements.

are no areas where we have exhaustive coverage – we are unable to obtain performance metrics for our aggregations. We observe that wealthier areas such as Nairobi have larger number of structures with high levels of consumption; this aligns with our intuition.

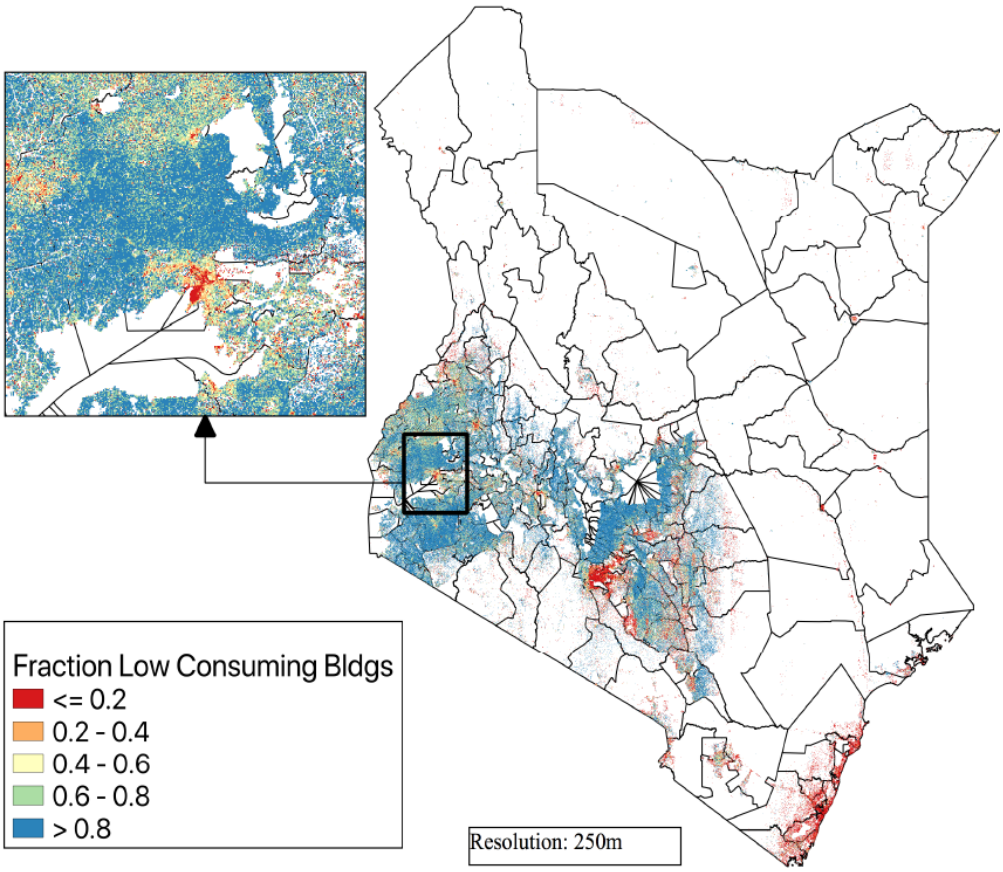


Figure 2.7: Novel predictions of electricity consumption levels for Kenya, aggregated at 250m. Blue shows regions with a large fraction of low-consuming buildings while Red shows regions with a large fraction of high-consuming buildings.

2.5.6 API and Users

An API was developed to share building consumption prediction estimates freely to the general public. The building consumption estimates are aggregated at 250m, 500m, 1000m and 10,000m cell resolutions for privacy concerns. Users are able to access consumption predictions of a single cell or collection of cells using the available cell resolutions. Users can make point or polygon coordinate queries. Figure 2.8 shows a sample JSON response given an input request polygon. The

Figure 2.8: Sample consumption level API JSON response given an input polygon request. Structure counts for each class and prediction confidence levels are returned.

query response contains a summary count of all buildings in the queried area of interest that lie in both the low and high consumption classes. For polygon queries, probability figures are included that convey the model's confidence in the structure classifications made. Summary statistics for cells 1-15 are returned if the summary feature is selected, else individual statistics for each cell is returned. The API has seen significant engagement from a variety of NGOs, institutions and individuals since its launch. The API has so far registered 7 active users who combined have made nearly 15,000 requests. We fully expect the engagement to grow as we make more countries and consumption categories available to users through the API. For more information on how to access and use the API, please use: `https://eguide.io/#api`

2.5.7 A note on proper applications of our work

This chapter proposes a methodology to estimate anticipated levels of electricity consumption. Such an exercise given the premise itself of estimating how much electricity a specific household will consume- is fraught with many uncertainties in the prediction itself, the embedded assumptions and (im)proper applications of the results. First, this paper focuses on residential customers only with an estimate of whether they are expected to be in a low or high level of consumption if they are grid connected. Secondly, an electrification bias may be present, as the analysis cannot and does not evaluate customers that are currently electrified with off-grid systems. Thirdly given that our validation results show that the odds of correct predictions are roughly two out of three, electrification planners risk classifying individual households or groups of households (perhaps in some geographies/landscapes or perhaps based on roof materials/footprints) with otherwise high consumption as low, potentially leading to biased outcomes. Hence we believe that there is no substitute for individual and community agency and representation; and no substitute for utility/planner surveys. On the flip side, utilities could uniformly end up simply estimating that all new consumers are low-consuming, extrapolating from their recent observations. Analysis such as that presented here could be one additional input in decision making. Utilities could improve their own predictions with the much larger and comprehensive data (e.g. bills and locations of all existing customers) that they have. Our novel results are aimed at providing a new methodology and a high-level guidance, making them suitable for site prioritization across larger landscapes, where a *human-in-the-loop* approach such as surveys can be taken, to validate the true consumption (through appliance ownership etc) after initial sites have been determined.

2.6 Summary

This chapter proposes a method to estimate levels of electricity consumption for unconnected households using pre-electrification images. Our results show that our novel methodology of using satellite images for electricity prediction outperforms existing approaches currently used by

energy planners. We also present a multi-modal approach that combines satellite images with other data sources to further improve the overall prediction performance. The predictions of our model (currently deployed in Kenya) provide a birds-eye view of relative levels of consumption upon electrification throughout the country and equip decision-makers with a direct measure of expected energy usage as well as a novel proxy for economic activity. This can enable better system planning and stretch investments in electrification to connect more people to modern energy sources.

Predicting electricity usage from satellite images remains a difficult task, mainly because elements in satellite images (rooftops, roads, fields) are only proxy measures for electricity. Utilities could improve their own predictions with additional much larger and comprehensive data (e.g. bills and locations of all existing customers) that they possess. We are keen to co-develop such methodologies with partners. We also plan to evaluate the transferability of our method by extending our approach to other countries and sectors (e.g., commercial and industrial).

2.7 Appendix

2.7.1 Multi-Layer Perception (MLP) Architecture

Figure 2.9 shows the MLP architecture used to train *Model B* (Non-visual Data) and *Model D* (Building Characteristics Only). The MLP consists of 3 dense layers with 64, 32, and 16 filters respectively, all with ReLU activations. The last dense layer consists of a softmax activation. 25 % dropout was applied to minimize overfitting.

2.7.2 Performance of building segmentation

Additional evaluation of the building segmentation task is done by observing how the classifier performs at varying training data sample sizes. Figure 2.10 shows the F1-score at different training data sample sizes when random subsets of the data are selected and either random weights or building segmentation weights are used to initialize model training. At each sample size increment, samples from the previous sample size are included. E.g. the 20 % dataset contains all the samples

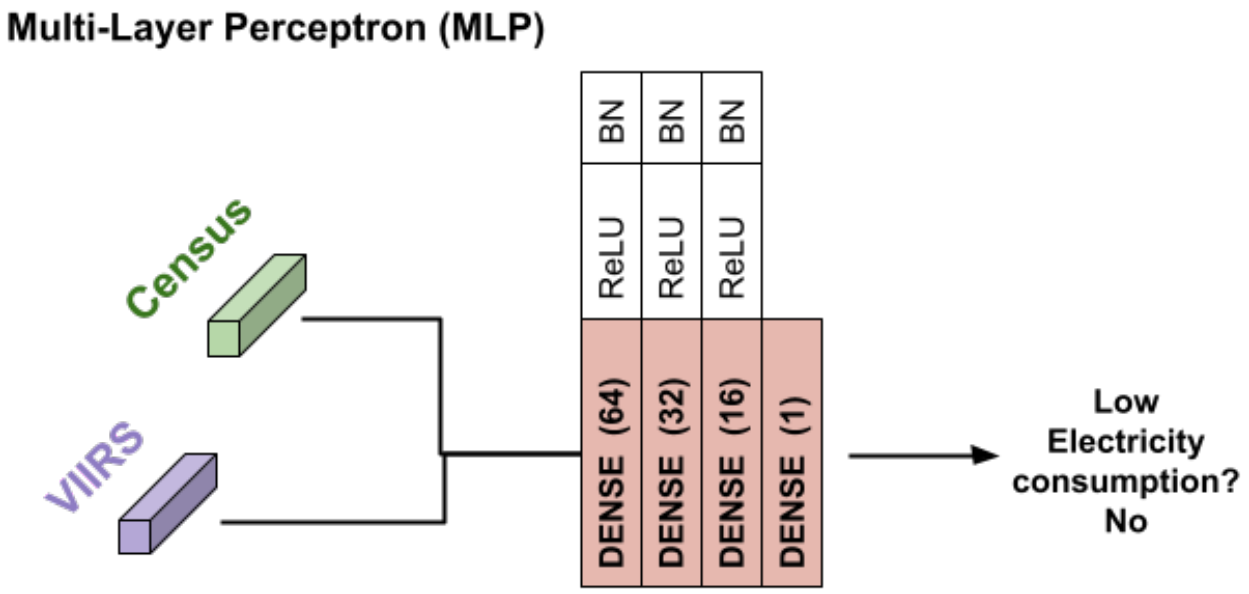


Figure 2.9: MLP architecture used to train *Model B* and *Model D*

from the 5 % dataset. Initializing with building segmentation weights offers performance gains especially at smaller sample sizes. The improved performances with building segmentation weights suggests that underlying characteristics about buildings (rooftop type, color, size) provides relevant features for consumption prediction. This is inline with our initial findings that building characteristics are relevant in predicting consumption levels. In addition to improved model performance, building segmentation weights make the classifier less susceptible to label quality. Specifically when random weights are used for initialization, it is observed that the randomly selected sub-sample at 60 % of the full dataset, performed the best and performance dropped as more samples were added. This suggests that the ease | difficulty of the sub-sample significantly affects performance. Building segmentation weights initializes the model in a suitable learning space and has a regularizing effect even as harder labels may be introduced, allowing only additional useful information to be extracted.

Obtaining large amounts of useful samples to appropriately predict consumption of yet to be connected customers can be challenging. For energy practitioner looking to apply our approach, we show that learning about buildings from using a segmentation task, provides useful weight

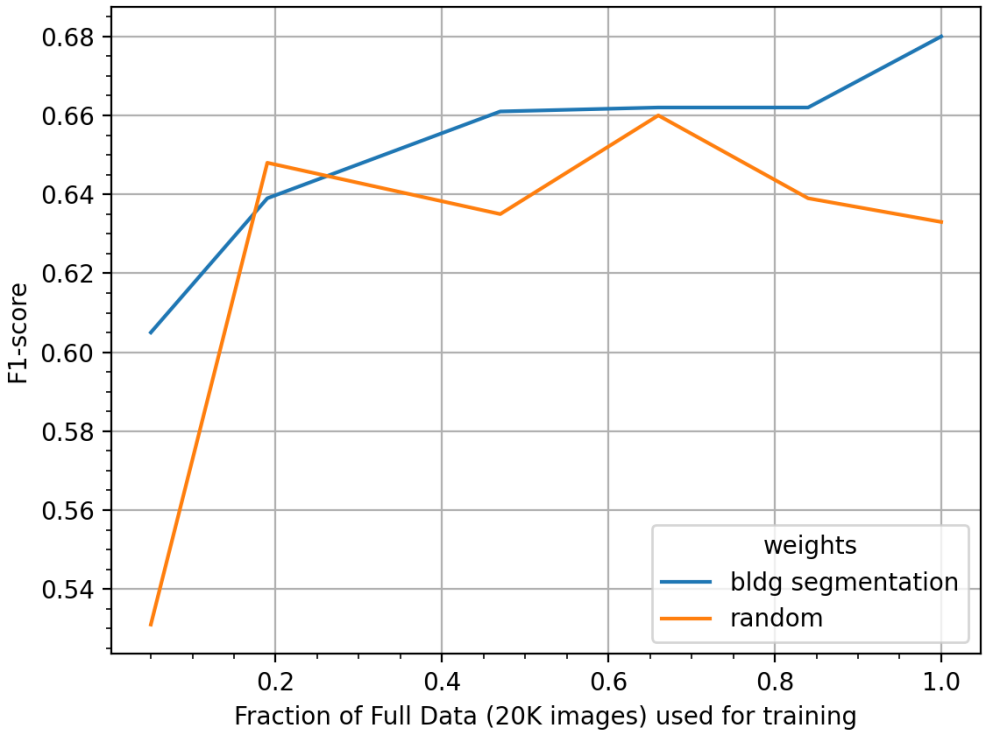


Figure 2.10: Comparison of prediction performance when the classifier is initialized with random weights versus building segmentation weights. Learning about building segmentation improves performance in low-data regimes and makes performance less susceptible to harder labels thereby offering a regularizing effect.

tuning needed for appropriate prediction of consumption tiers.

2.7.3 Multi-modal architecture: Encoder and MLP

Figure 2.11 shows the multimodal architecture used to combine satellite images with public data sources.

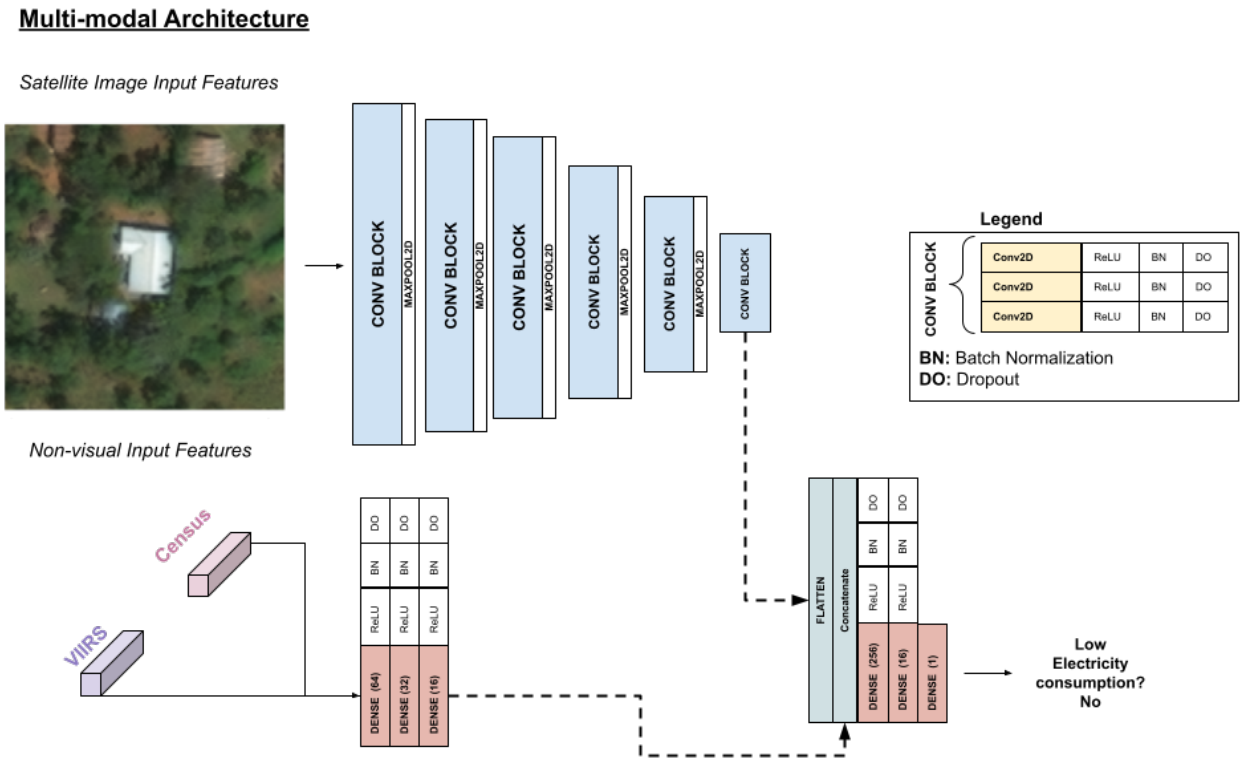


Figure 2.11: Multi-modal architecture combining the CNN image-based encoder with an MLP to predict consumption levels using visual images and non-visual public data sources.

Chapter 3: A scalable framework to measure the impact of spatial heterogeneity on electrification

3.1 Introduction

Sustainable Development Goal 7 (SDG7) was adopted in 2015 by the United Nations member states to provide access to "affordable, reliable, sustainable, and modern energy" to all by 2030 [87]. Although there has been significant progress towards reaching SDG7 in recent years, 840 million people still live without electricity as of 2019 [4]. The lack of access to electricity in developing regions necessitates rapid and informed decision making on electrification options. Among the options available today, isolated or individual customer-scale solar-battery systems, frequently referred to as solar home systems (SHS), do not require any network at all. Networked options, such as a grid connection, rely on one or more large power plants located at multiple points on a network, where transmission lines carry the power over long distances (generally hundreds to thousands of kilometers) on a high-voltage backbone. This backbone in turn feeds a medium-voltage (MV) network, which distributes electricity directly to large consumers and transformers. The transformers drop down the voltage and allow a low-voltage (LV) wire to connect smaller customers in roughly a kilometer radius. The MV and LV network combined with transformers is called the distribution system. In the context of investments for access to grid electricity, this system generally represents the largest fraction of the total system cost and therefore, understanding the requirements of the distribution systems is quite important for proper rural electrification planning.

Determining the best electrification option for a region is particularly challenging especially when a mixture of solutions is possible. In fact, Carvallo et al. show that in places with low electrification rates, hybrid solutions that pair networked systems with standalone decentralized

options typically offer an attractive approach to electrification [88]. To aid utilities in identifying electrification options, a number of electrification planning tools that are capable of choosing between decentralized and networked options have been developed [89]. These tools apply least-cost methods to determine the demand points, which may be better served by grid extensions and those whom would gain more benefits from off-grid systems. Depending on the techno-economic model used and the availability of the data, granularity level of these tools varies. Literature suggests that using all consumer locations for large-scale planning imposes strong computational constraints on many models. Thus, the studies aiming for large-scale electrification such as at the country level tend to make simplifications by grouping individual structures into villages or large cells of 1km [89].

When consumer points are aggregated over large areas for planning purposes, it is not possible to understand the impact of the settlement patterns on the components of the distribution systems and this may lead to misleading results when determining the electrification option at the local level. In order to address this problem, we first propose a data processing strategy for Kenya to convert structure locations, identified from satellite imagery, to estimated household locations using census data. Then, we present a computational framework that involves a two-level network design algorithm to find an abstract representation of the power distribution system involving low-voltage wires, medium voltage wires, and the transformers between the two levels of the system. Given the system components, we introduce three simple metrics for per-household connectivity requirements of LV wire, MV wire, and transformers to interpret our results at the administrative unit level and the sub-administrative unit level. With our administrative level analysis provided for 9.2 million structures in Kenya, we show that traditional rural/urban classification based on population density is often deceiving in estimating the cost of electrification and a new categorization based on our metrics (combination of MV and LV wire requirements and the number of structures per transformer) provides more relevant estimates on the total cost. Moreover, in the sub-administrative analysis, our metrics can help determine the least-cost electrification option (e.g.,grid, mini-grid, or stand-alone systems) for expanding access and create a platform to perform

sensitivity analysis based on different cost components. To the best of our knowledge, there is no focused study that evaluates the value of different connectivity metrics, highlighting their roles and strengths in facilitating the electrification planning process in a scalable manner. In addition, our work shows how these connectivity metrics complement and clarify the composite cost metric, which is usually the only metric reported in many planning studies.

This chapter adds to the existing knowledge in three ways. First, the work demonstrates a data processing strategy to estimate the residential connection locations at the country level. Second, the chapter proposes a framework for applying large-scale computationally-intensive network optimizations on millions of consumers. Third, the chapter introduces three complementary connectivity metrics for evaluating electrification choices agnostic to the network planning approach. The methodology that we put forward can assist the decision-making process in electrification planning and serve as a decision support tool for identifying suitable electrification options. While we present results for Kenya, we believe that this tool can be applied to places with little to no access to electricity. Meeting the targets set in SDG7 requires consideration of multiple consumers across large landscapes with varying settlement patterns; our work outlines a feasible approach to perform planning at scale to support electrification objectives.

The remainder of the chapter is organized as follows: In Section 3.2 we present relevant contributions from literature, in Section 3.3, we discuss the data used for this work and present a method to estimate residential connection locations from building structures identified by satellite images. In Section 3.4 we describe the two-level network optimization algorithm used in our framework and our computational improvements. In Section 3.5 and 3.6, we show the metrics computed using the two-level network algorithm and their applications at varying resolutions. In Section 3.7 we also show the sensitivity of our metrics to cost. Finally in Section 3.8, we propose feasible extensions to our work and conclude.

3.2 Related Work

In a comprehensive review paper by Ciller et al. [89], planning tools used for rural electrification are classified into three groups: pre-feasibility studies, intermediate analysis tools and detailed generation and network design tools. Although not all efforts towards rural electrification are presented or used as a software tool in the literature, we review the studies related to our work using the same classification.

Pre-feasibility studies as in [90, 91, 92, 93] estimate the least cost approach for different technology choices using simplifying assumptions, allowing for a first pass at the planning problem. These studies do not typically include network design and are likely to group consumers into villages or cells (e.g., 1 km x 1 km). Grouping of consumers reduces the computational granularity, and therefore, pre-feasibility studies have lower model complexity, high computational speed, and are valuable for quickly evaluating technology choices over large-scale areas at low resolution given varying generation options. Cost remains the key reported metric of evaluation used with pre-feasibility methodologies.

The studies that are used for intermediate analysis have various complexity levels depending on the network design and the technical details considered. Similar to the pre-feasibility studies, the resolution of the data used in the intermediate analysis studies is low. An intermediate planning approach presented in [94] proposes a spatial cost minimization electricity planning model for Kenya to decide between grid-based electrification and off-grid solutions. The model provides the basis for Network Planner (NP), an online decision-support tool that has been developed to explore grid, mini-grid, and off-grid options for rural communities [95] and has been used in national electrification studies of countries such as Senegal [96], Ghana [97] and Nigeria [98]. In [99], Abdul-Salam and Phimister propose an approach based on hierarchical lexicographic programming that considers both cost efficiency and political economy to give large populations a priority for grid connectivity. Bolukbasi and Kocaman propose a prize collecting Steiner tree approach to choose between grid and off-grid options and to determine the network design for the

grid-compatible nodes in a least cost manner [100]. Although these studies offer great value by folding in more modeling complexity, they reduce the computational difficulties by aggregating individual consumers and therefore neglect the effect of settlement distribution. Similar to many electrification planning models, intermediate studies report cost as the key metric of evaluation.

In Ciller et al. [89], Reference Electrification Model (REM) [101] is described as the only planning tool that falls under the detailed generation and network design class. REM aims to design a power system configuration evaluating the demand profiles for the individual customers. To overcome the computational burden of a detailed plan using local level data, REM uses a sequential approach to plan the sub-systems in a hierarchical manner. Although it provides a very detailed network configuration, it is acknowledged in [89] that, the network design approach used in REM is not designed for rural electrification planning and may perform poorly when designing small networks.

There are also some studies in the rural electrification literature that use customer or household level data as in REM [101], however, aim for obtaining quick estimates for the network structure and associated costs, rather than being used for detailed implementation. The main objective of these studies is to show that rural settlement patterns – especially in Sub-Saharan Africa – can be diverse and the effect of settlement patterns on the electrification options might be overlooked in the pre-feasibility and intermediate analysis studies due to the aggregated data considered. Using several datasets of structure locations developed from satellite imagery, Zvoleff et al. propose a metric, called the homogeneity index, that serves as a proxy for the degree of dispersion of the structures. They provide solid evidence about the impact of geographic patterns on the cost of energy infrastructure. However, they assume that all identified structures within the images are households and these households can be connected via single level LV network [102]. Kocaman et al. [103] use the same structure locations as [102] to propose a computationally-intensive two-level (MV and LV) network optimization approach and evaluate the cost of grid extension for the distribution systems in limited-size rural regions. In [104], Adkins et al. use inter-community and inter-household distances as proxies to estimate MV and LV wire lengths.

In this paper, we build upon the approach presented by Kocaman et al. [103] and present a computational framework to incorporate a large number of connection points into electrification planning, thereby improving modeling capacity at reasonable computational speed. In this direction, our study is the first to propose a detailed data processing strategy to estimate the residential connection locations from hand-labelled structure points. Moreover, we propose a set of per-household connectivity metrics - low-voltage wire, medium-voltage wire and transformers - that can be used to rapidly evaluate electrification choices agnostic to the network planning approach. We show how network outputs from detailed models such as REM [101] can be used to compute our metrics and how these metrics facilitate rapid analyses of the electrification landscape within a country. We discuss all our results for Kenya, for which, to the best of our knowledge, no similar findings are available in the literature.

3.3 A Data Processing Framework

In this section, we first discuss the source of our structure locations data and propose a data processing framework to estimate the household locations.

3.3.1 Structure locations

Our study is principally built upon 11.9 million human-labelled building structures in Kenya from satellite imagery data obtained in 2017. This data was obtained from the Kenya National Electrification Plan - Structures Survey and includes latitude and longitude pairs for each identified structure within the images. No additional information is provided on the structure type or its pertaining attributes such as rooftop type and area.

3.3.2 Estimating household locations

It is quite common for rural households to own multiple structures (shed or outhouse in addition to living quarters), while in more urban locations, multiple households may dwell within the same structure [105]. We propose a method to obtain an estimation of households from human-labelled

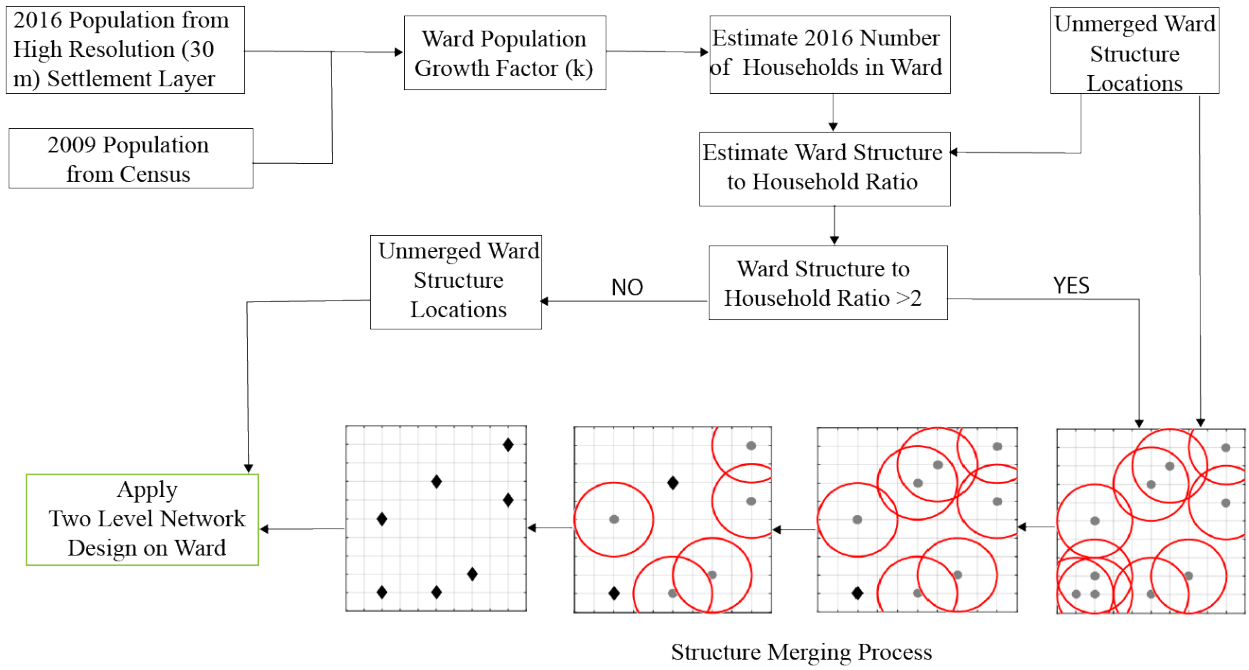


Figure 3.1: Data processing framework: 2016 population from the High Resolution Settlement Layer and 2009 population census are used to estimate a population growth factor (k), which is used to estimate 2016 household counts. Wards with structure to household ratios > 2 are further processed, where structures are merged using a set-covering merging algorithm. The two level network design is ran on resultant structures.

satellite imagery data. Census data provides the number of households at varying administrative levels. For the case of Kenya, the census provides household counts of each ward. Wards in Kenya (about 1400 in number) represent the smallest administrative unit in Kenya. The household counts from census data, provide only aggregates with no information on household locations. Because the Kenyan census is decennial and there is readily available 2009 Kenya census data, we apply a correction strategy to estimate household counts in 2016. Facebook's 2016 High Resolution Settlement Layer (HRSL), provides population data at a 30m resolution [106]. Given the 2009 population data at the ward level, and using HRSL population data, we estimate a population growth factor *k* for each ward, which represents the growth a ward has experienced between 2009 and 2016. We assume household counts scale linearly with population, thus we use a 1:1 relation between population growth and the growth in the number of households [1]. Applying this growth

[1]From recently released 2019 Kenya census data, we observe roughly 10 % difference between population growth and the household count growth from 2009 to 2019.

factor k to the 2009 ward-level household data, we can estimate the expected number of households in 2016 for each ward. Upon obtaining the 2016 household estimates, a direct comparison can be applied to the 11.9 million structures obtained from satellite images.

Next, we compute a per-ward Structure To Household ratio (STH) that is the ratio of 2017 identified structures to estimated households (obtained from the census data adjusted to 2016). This ratio is frequently greater than 1, as observed by Kenya 2014 DHS results [105]. In this paper, we assume every household in a ward to have the same number of structures; we allow this ratio to vary from ward to ward. Where the STH ratios are higher than 2, we apply a merging algorithm described below. We present our full data processing framework, including estimating household locations and our merging algorithm in Figure 3.1.

3.3.3 A merging algorithm

A set-covering algorithm was applied at different radii and the resulting structure counts were compared to each ward's household count. The set-covering problem is an NP-complete problem and aims to find the minimum number of sites and their corresponding location to cover all demand nodes [107]. Here, we adopt a well-known heuristic approach proposed by [108] to find the reduced set of structures that cover all building structure locations within a radius r of interest. Figure 3.1 highlights the merging process when STH are greater than 2. The steps of this approach are as follows:

1) Draw a circle around each building structure location with a specific radius r.
2) Count the number of points in each circle.
3) Take the circle with the maximum amount of points (Ties are broken arbitrarily).
4) Eliminate the building structure points 'covered' with the circle in Step 3.
5) Repeat 1-4 with the remaining points until each building structure point is 'covered'.

A merging radius of 20 meters was found to be most suitable to match household counts with

the adjusted census data, a distance which reduces the 11.9 million human-labelled structures to a merged structure count of approximately 9.2 million. The average STH ratio for all wards is 1.3 with a maximum of 2.6. See Appendix 3.9.1 for a more detailed discussion on merging radius. The merged structures and their corresponding locations are subsequently used in the rest of the paper. The paper treats each merged structure as requiring a separate electric connection.

3.4 A Computational Framework for Distribution Systems Planning

We propose a computational framework to estimate the i) per-structure LV wire requirement; ii) per-structure MV wire requirement needed for each structure to be connected to the network; iii) the number of structures per transformer, and; iv) a per-structure connection cost. In this section, we detail how we compute these four metrics. Motivated by the need to evaluate cost estimates and additional metrics which highlight spatial diversity, this paper adopts a two-level network design (TLND) approach proposed by [103] and proposes a decomposition approach to obtain results over a large spatial extent.

3.4.1 A two-level network design approach

The TLND combines the transformer location problem and the LV and MV network design problems into a single optimization framework by modeling a two-level radial power distribution system. The two-level network connects demand points (in this case post-merged structure locations) via intermediate transformers, which reside on a primary MV network. The merged structure points are connected to the transformers with a secondary multi-point LV network. As in [103], transformers are assumed to be uncapacitated, i.e. they can handle unlimited demand. However, there is a limitation on the distance between a merged structure point and its serving transformer. The TLND does not consider the presence of the legacy grid, high voltage (HV) network[2], load balancing requirements, or power flow.

[2]High voltage transmission networks are strongly dependent upon the specific location of central power generation systems

Determining the layouts of both LV and MV networks while locating distance-limited transformers that connect them in a continuous space is an NP-hard problem, since the continuous space location-allocation problem is NP-hard [109]. The algorithm proposed by [103] to solve this NP-hard problem leverages an agglomerative hierarchical clustering approach. This bottom-up approach starts with locating a transformer on each demand point (where each demand point represents a singleton cluster) and iteratively decreases the number of transformers as a pair of clusters is agglomerated (merged) in a greedy manner based on a dissimilarity measure. In this paper, the centroid method is used as the dissimilarity measure: the closest pair of transformers which can be replaced by a single transformer located at the centroid of the demand points without violating the distance constraint is merged at each step. The minimum spanning tree problem aims to find a tree (a network containing no cycles) that spans all the points minimizing the total cost of the connection. At any iteration of the clustering algorithm, once the transformer locations are updated, the MV network between them and the source point is found using a minimum spanning tree algorithm with the guarantee of an optimal solution [110]. Once the clusters are formed at each iteration of the agglomerative hierarchical clustering approach, the multi-point LV network between the transformers and the demand points is obtained by solving the capacitated minimum spanning tree problem. This problem aims to find a spanning tree rooted at the transformer considering a distance or a number of nodes on each sub-tree emanating from the root point. In the TLND, a distance limit is used on the length of a sub-tree and the problem is solved using Essau and Williams's heuristic approach [111]. The maximum distance between demand points and the transformer is assumed to be 500m, which is a widely accepted limit for open-wire LV lines. For each step of the agglomerative clustering, the algorithm calculates the minimum spanning tree as the MV network and the capacitated minimum spanning trees within each cluster as the multi-point LV network. The overall cost is computed at each step and the least cost design is outputted.

In order to run the TLND, we also assume that a transformer cost USD 2000, a meter of MV wire cost USD 25, while a meter of LV wire cost USD 10. While we use costs obtained from [103], our TLND can be run with costs that are reflective of any region of interest. Given the cost

parameters and the constraints, the objective of the algorithm is to find the number and locations of the transformers and the least-cost layouts of MV and LV networks. In the next section 3.4.2, we demonstrate how we integrate the TLND into the computational framework for estimating the metrics at the country level.

3.4.2 A decomposition approach for large-scale planning

Planning at a national scale with individual structures result in millions of demand points: in the case of Kenya, 9.2 million merged structure locations need to be considered for planning. Even at the resolution of the smallest Kenyan administrative unit, the median and maximum per-ward merged structure count is 6,872 and 32,321, respectively. In response to the significant computational requirements of large-scale optimizations, [112] proposes micro-optimizations for small zones as an approach to applying network algorithms for large-scale distribution planning. Inspired by this micro-optimization strategy, we devise a framework to run the two level network design algorithm on millions of demand points, without sacrificing spatial heterogeneity.

We develop our computational framework to minimize run-time without sacrificing performance. Our approach considers the smallest administrative unit as the entry point to apply the framework. For Kenya we apply the framework in parallel on each ward. Given a ward, the framework consists of three steps: 1) recursively decompose the ward into cells, 2) parallelize the TLND for all cells, and 3) reconstruct the ward. Figure 3.2 shows our computational framework for a synthetic ward and its corresponding structures. In Figure 3.2(a) we take a ward as shown in i) and check the ward against three predefined parameters M, N and R. We compute the number of structures in a ward (m) and compare it to a predefined threshold (M) which represents the maximum number of structures that can be present. Next our approach computes the number of structures for the largest cluster in that ward. Clustering is performed by the two level network design algorithm to assign structures to a given transformer: by limiting the maximum number of structures in a cluster to a predefined threshold N we are able to reduce the time it takes to design a low voltage network for the structures in the cluster. Similarly, the ward radius (meters) is computed and

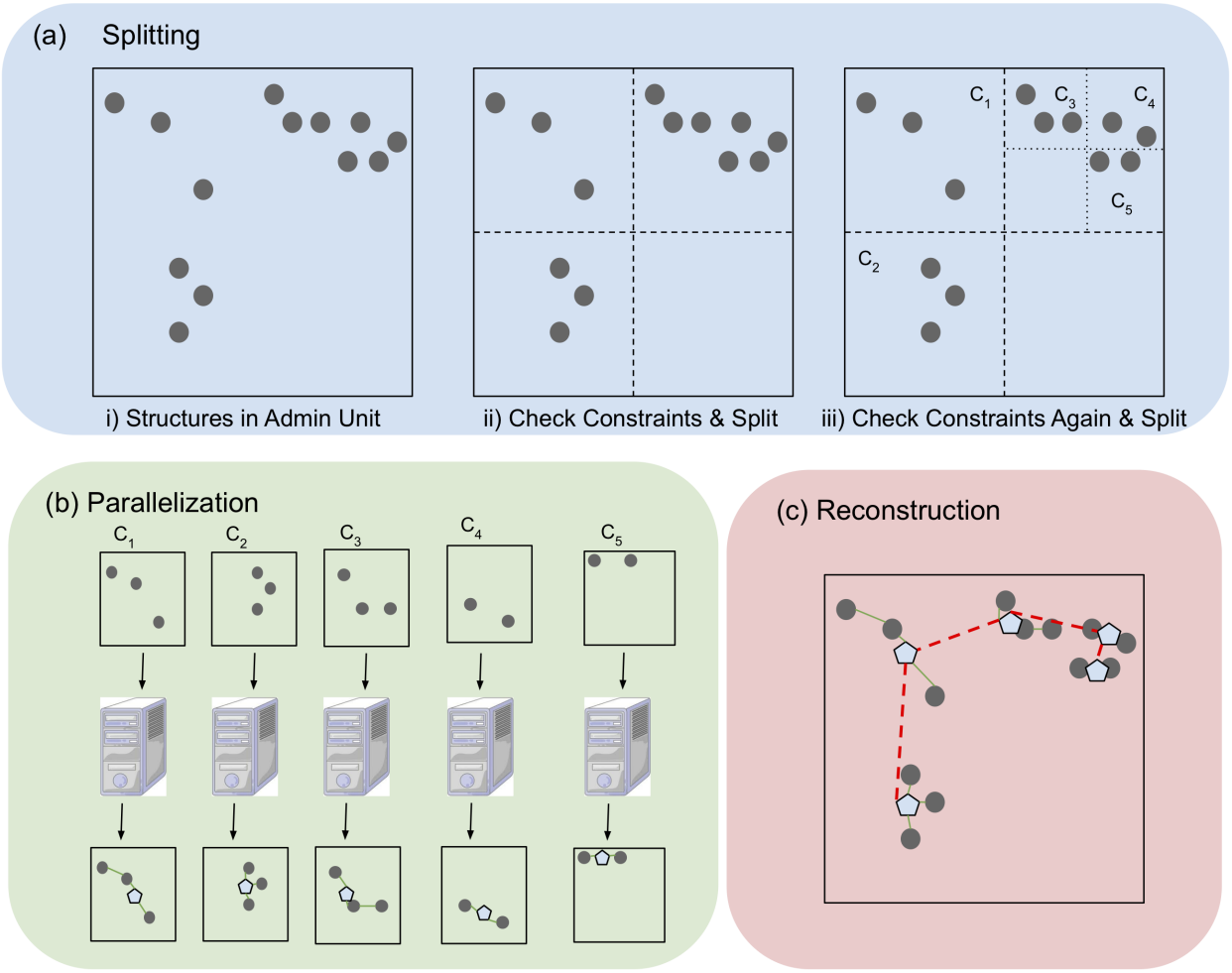


Figure 3.2: Computational framework for planning using multiple demand points. (a)*Splitting*: A recursive split is used to obtain valid cells for the network planning algorithm. Splitting continues until all three constraints are met (Number of structures in cell < M; Number of structures in largest cell cluster < N; cell radius > R) (b)*Parallelization*: The network planning algorithm is run in parallel on all valid cells to obtain transformer locations, the low voltage network and a local medium voltage network (c)*Reconstruction*: Transformer locations from all cells in a ward are used to compute the medium voltage network for the ward.

compared to a predefined minimum radius R, which ensures that the connecting radius of a utility is preserved and the number of structures connected to a transformer is maximized. The radius parameter counterbalances the splitting and prevents the wards from being excessively split. If r is less than R, the ward is accepted as a valid cell for the network planning algorithm; if r is greater than R, then m and n are compared to M and N, respectively.

Taking the example presented in Figure 3.2(a)(i), in which the per-cell maximum number of

Algorithm Recursive Split of Structures in Administrative Unit

```
Inputs:
M, maximum number of merged structures in a cell,
N, maximum number of structures in the largest cluster,
R, minimum cell radius.
validcells = empty list
for each adminUnit in allAdminUnits do
    allcells = [adminUnit]
    while the number of cells in allcells is greater than zero do
        for cell in allcells do
            Remove cell from allcells
            m = compute number of merged structures in cell
            n = compute number of structures in largest cluster in cell
            r = compute cell radius
            if r <= R do
                Append cell to validcells
            else do
                if (m < M) and (n <N) do
                    Append cell to validcells
                else do
                    newcells = split cell
                    Append newcells to allcells
```

Figure 3.3: Recursive Split Algorithm

structures M is assumed to be 3, Figure 3.2(a)(ii) shows the results of the initial splitting. The split cell that does not meet the constraints is further split until the constraints are met, as shown in Figure 3.2(a)(iii). Formally, our recursive split algorithm splits the ward into cells C_i such that they obey the following constraints: 1) the number of merged structures in C_i must be less than a predefined threshold M; 2) the number of structures for the largest cluster in C_i must be less than a predefined threshold N; and 3) the radius of C_i must be greater than a predefined radius R in meters to allow any further splitting. The predefined parameters of M, N, R, are all user-defined parameters which can be determined a priori by running tests on a small number of wards in order to understand the effect of number of structures, settlement patterns and the search radius on the runtime of the network planning algorithm. We discuss the effect of runtime and our choice of parameters in Appendix 3.9.2.

Figure 3.3 presents pseudo-code for our splitting algorithm. Once valid cells are obtained, the TLND is applied in parallel. Transformer locations, the low voltage network and a localized medium voltage network is obtained for each C_i cell as shown in Figure 3.2(b). The localized medium voltage network does not consider transformers in other cells belonging to the same ward; we address this in a final step by putting cells back together and rerunning the medium voltage computation (minimum spanning tree algorithm) with transformer locations across all cells in the ward. We show in Appendix 3.9.2 that splitting the ward does not have adverse effects on the obtained results.

Detailed computing specifications are as follows: Running 9.2 million structure locations was done on a computer cluster with two Intel Xeon E5-2680 v4 processors with 14 cores each, 128GB RAM and 200 GB local SSD. 17,330 cells were generated for Kenya and the TLND was ran on each cell. With the longest allowable runtime being 21 days, this resulted in 98.8% of cells completing the TLND. Given our framework, 90 % of the cells ran in under 12 hours with more than 50 % of the cells taking less than 1 hour to run the TLND. 98 % of the cells ran the TLND in under 4 days.

3.5 An Analysis on the Administrative Boundary Level.

In this section, we first discuss the value of our proposed metrics to measure the impact of spatial heterogeneity on the electrification cost using the smallest administrative unit resolution (i.e. ward). Next, we show the performance of our metrics compared to population density at this resolution. Finally, we discuss the effect of real settlement patterns on our computed metrics.

3.5.1 Proposed metrics calculated for Kenya

Results for each ward are averages across all merged structures within the ward. Here, we do not include the existing grid in Kenya but rather focus on evaluating the impacts of networking given the structures internal to the ward. Figure 3.4 shows the average ward level metrics by decile: per-structure low-voltage wire (meters), per-structure medium voltage wire (meters),

per-transformer number of structures, and per-structure cost (USD).[3] Given a desired proximity of structures to each other and to the transformer, our method allows for the quick and easy identification of suitable wards for different types of electrification. For example, an energy provider may be interested in determining which wards have an average distance between merged structures of less than 30m and correspondingly can be networked through LV connections. As shown in Figure 3.4(a), the 30m threshold corresponds to approximately 25% of the wards – primarily those in Eastern Kenya. Similarly, an energy provider might be interested in wards where transformers are

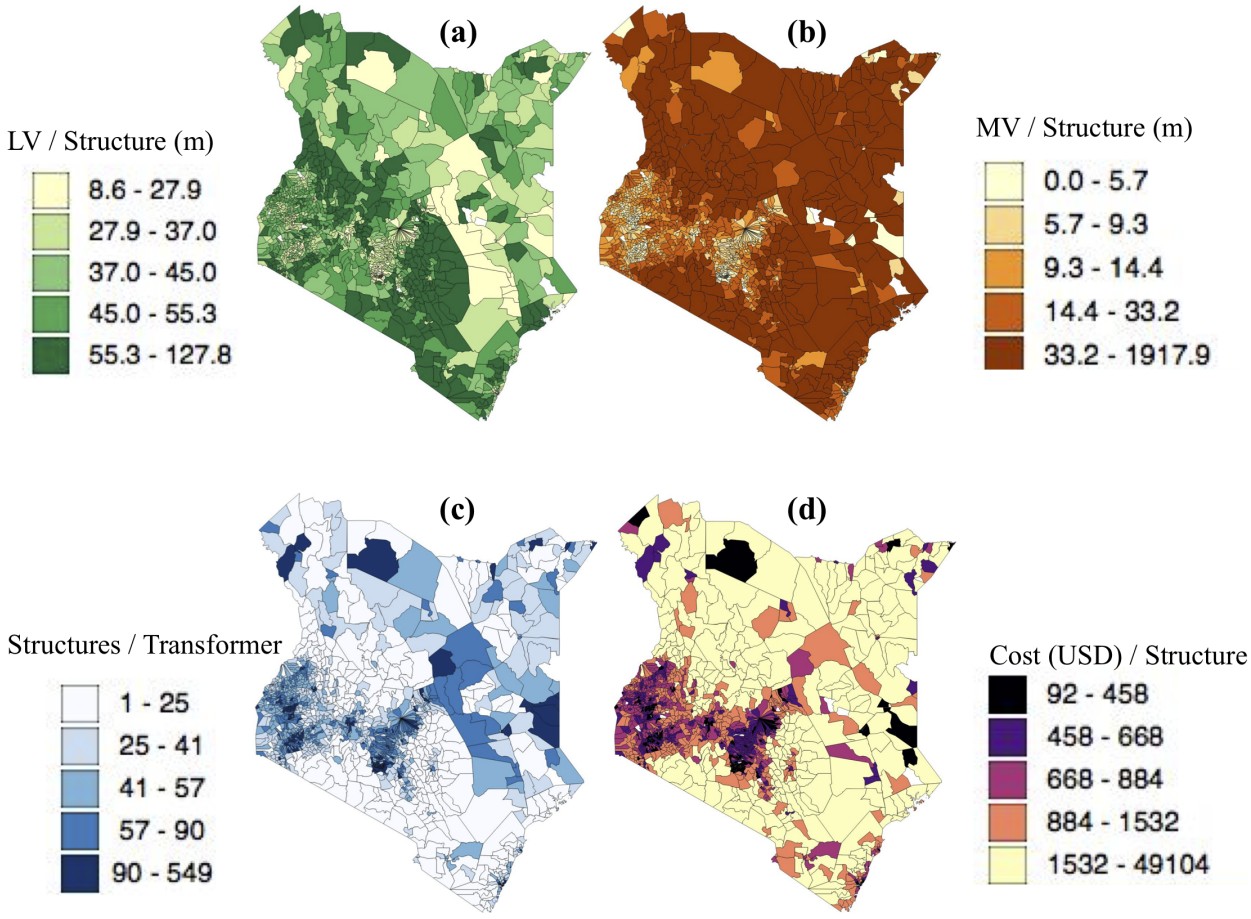


Figure 3.4: Average ward connectivity metrics for Kenya by decile.

in close proximity to each other and consequently are suited for MV networks. In Figure 3.4(b) we show that almost 50% of wards require less than 10m of MV wire per structure. The ability to

[3]It is important to note that the two-level network design enforces a limitation of 500 m for connecting structures on the same LV wire (due to voltage drop considerations).

specify both LV and MV requirements outside of costs allows planners to quantify the effects of regional geography on network design.[4] Figure 3.4(c) shows the average number of structures per transformer. Wards with the highest number of structures per transformer are found in more urban regions in Central Kenya. Generally, number of structures per transformer decreases in more rural regions even though there are a few otherwise rural wards in Eastern Kenya with higher transformer capacity.

Figure 3.4(d) shows the average ward per structure connection cost of electricity access: this cost reflects the average combined wire and transformer costs needed to connect a structure in the ward. The connection cost metric shows which wards are suitable candidates for networked grids and which wards are more suited for alternative electrification modes like mini-grids or solar home systems (SHS). Differentiating between wards suited for mini-grids versus those for SHS requires leveraging the 3 other metrics in Figure 3.4; the exact cost cutoffs for each technology choice would depend on the price of these alternatives and the utility's cost-sensitivity. The four metrics presented in Figure 3 capture the complexities of geography-dependent network design, the benefits of which are explored in the next section.

3.5.2 Why do we need new metrics?: A comparison with population density

Population density is a metric that is often used for estimating the location and type (rural or urban) of demand centers. For energy access problems, we observe that rural/urban classification based on population density may not be enough and is often deceiving in estimating the cost of electrification. A new categorization based on a combination of MV and LV wire requirements and the number of structures per transformer provides more relevant metrics to anticipate the total cost and create a platform to perform sensitivity analysis based on different cost components. For this purpose, we compare our metrics against population density to quantify the additional gains which our metrics may offer.

[4]It is important to note that computed wiring requirements are distances as a crow flies, and practical routing considerations might lead to distances which are larger than those presented here. This concern could be addressed by incorporating topology into the methodology.

Table 3.1: A new categorization based on a combination of our metrics to anticipate the cost of electrification

Category	Proposed Metrics			Population Density
	MV / structure	LV / structure	structures / transformer	
Urban & Suburban	Low	Low	High	High
Nucleated Rural	High	Low	High	Low
Non-nucleated Rural	Low	High	Low	Low
Extreme Sparse Rural	High	High	Low	Low

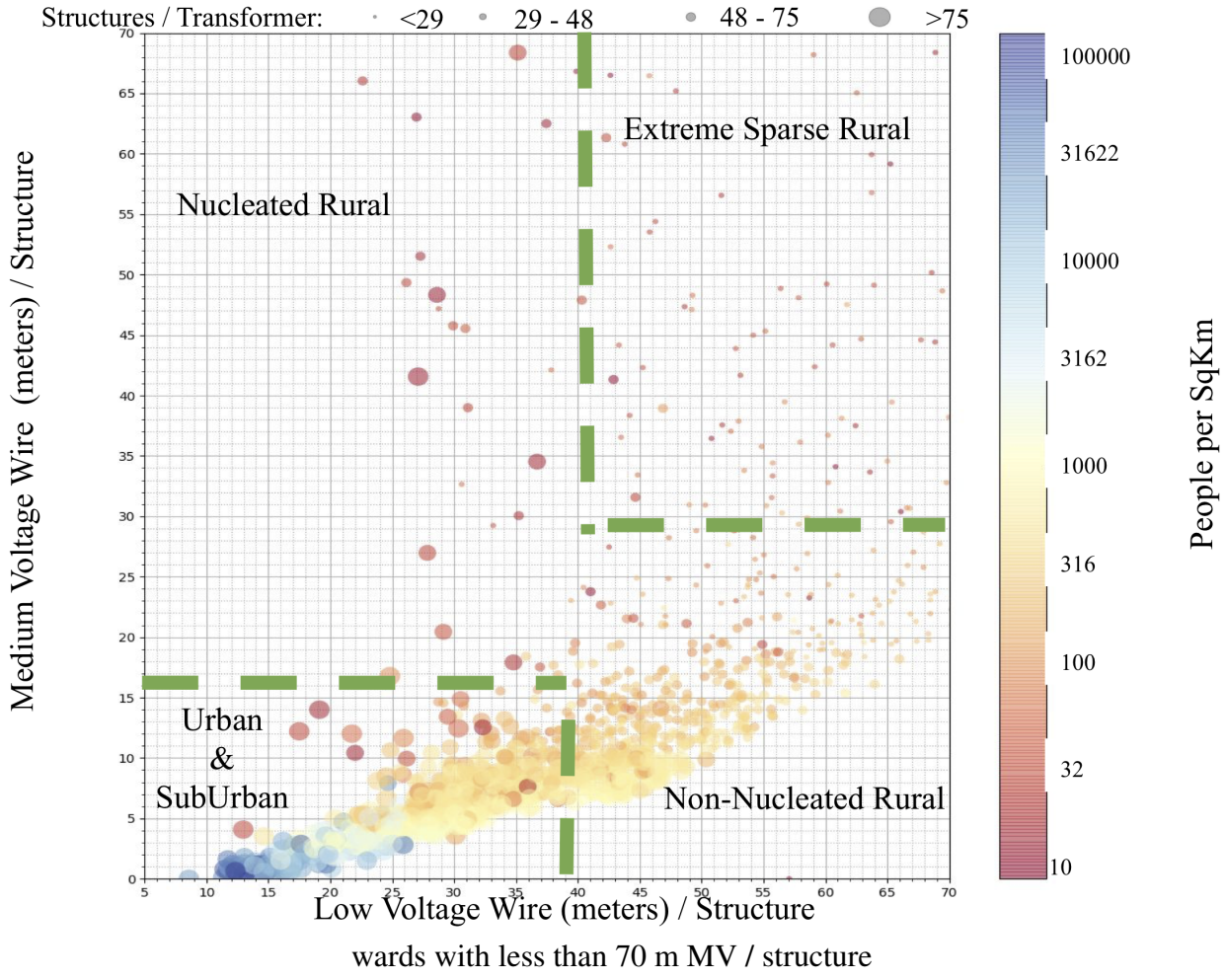
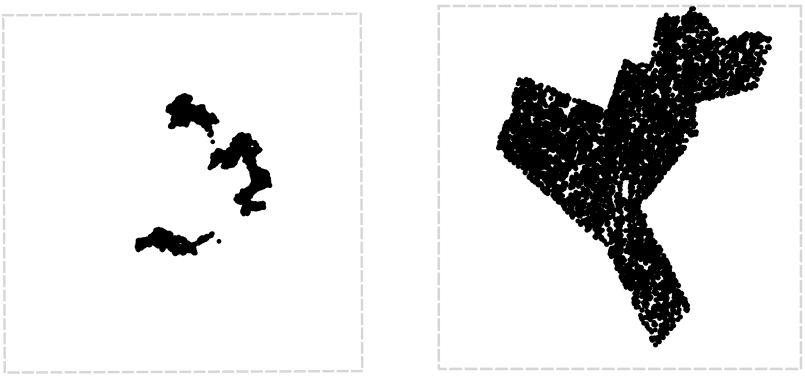


Figure 3.5: A scatter-plot showing per structure LV wire requirement against per structure MV wire requirements. Each bubble in the figure represents a ward in Kenya and the bubble size indicates the average number of structures per transformer by quartiles. People per sqkm are captured by the coloring of the bubbles. There are multiple wards with similar population densities that have varying MV and LV requirements. Thus our connectivity metrics capture more spatial diversity than population density alone.



(a) E, Siaya (32.5, 6.9, 7047) (b) F, Makueni (88.4, 34.5, 8610)

Figure 3.6: Two wards with around 120 people per sqkm are shown. The per-structure LV requirement, per-structure MV requirement, and the structure count of the ward are shown respectively in brackets. The grey boxes surrounding each ward represent 30 km^2 area for scale and do no show the administrative boundaries. Figure (a) and (b) show that wards can have similar population densities but varying settlement patterns which can influence the computed metrics.

Figure 3.5 shows a scatter plot of the per-structure MV requirement as a function of LV requirement. In this figure, each bubble represents a ward, and the bubble sizes show average number of structures per transformer of the ward. The average number of structures per transformer are grouped by quartiles and the quartile ranges are shown in the figure. The coloring in Figure 3.5 shows the people per square kilometer (sqkm). As expected, wards with higher population density (i.e. those in blue), tend to be grouped at the lower left hand corner of the figure, with low MV and low LV wire requirements and with higher number of structures per transformer. These wards tend to be more urban, likely with established grids. The upper right hand corner of Figure 3.5 contains sparse rural wards with high LV and high MV requirements and low number of structures per transformer. However, it is important to note that not all wards that can be considered rural (based on population density) reside in this quadrant. Given our proposed metrics, these rural wards should be further categorized as nucleated and non-nucleated (or dispersed) rural settlements, given their LV and MV combination. The details of this classification are summarized in Table 3.1.

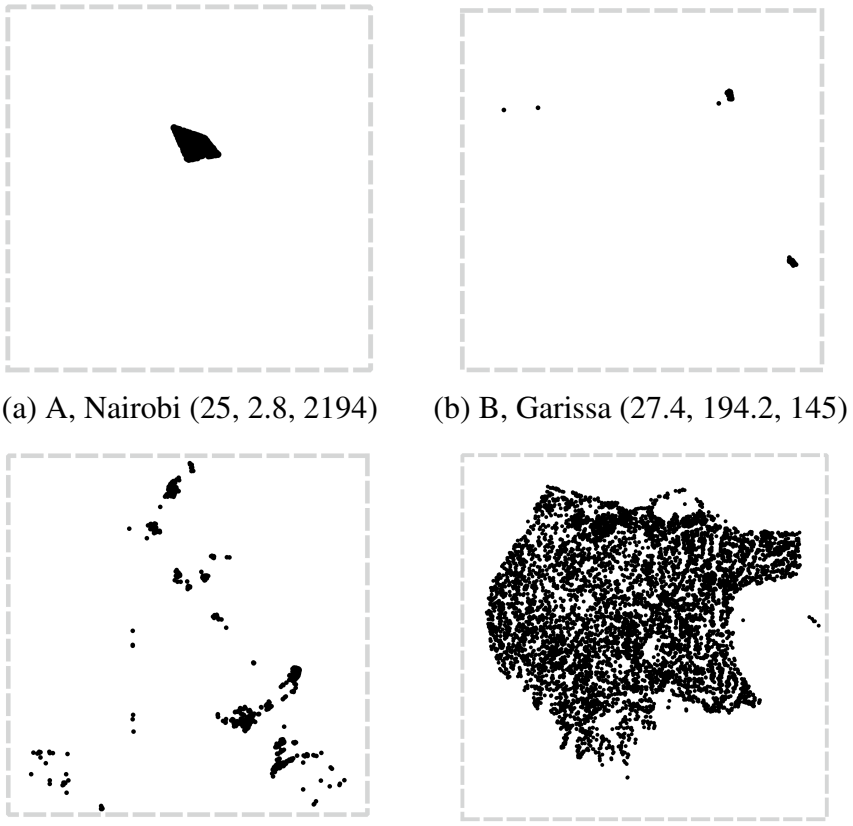
A strong observation from Figure 3.5 is that there are a number of wards with varying connectivity metrics at similar population densities. To explore this observation, we analyzed two such wards with similar population densities of 120 people per sqkm. Figure 3.6 shows both wards in a

30 km^2 box for scale but does not show the administrative boundary of the ward. The figure shows the per-structure LV length, per-structure MV length and the number of structures in brackets, respectively. Upon comparing both wards, we see that ward E in Siaya has very different LV and MV requirements to ward F in Makueni, although they have similar population densities and a similar number of merged structures. LV and MV requirements in ward E are significantly lower because of high structure nucleation, while the LV and MV requirements in ward F are much higher because structures are further away from each other on average. The varied infrastructure requirements of both wards results in an average difference in connection cost of $1341. By using our proposed metrics, we capture more insights on the diversity of wire requirements and by consequence connection costs needed to provide electricity access. We further quantify the dissimilarity in wire requirements for wards with similar population densities in Kenya. For every ward, we identify wards of similar population density (within 10 %). We compute the average LV and MV difference between wards with similar population density and the ward of interest. On average, 47 % of the wards with similar population density have LV or MV differences greater than 20%. This indicates that using population density as a metric for connectivity would be misleading approximately half of the time. This distribution of system requirements is lost when population or structure density alone is used as the metric of evaluation, or when residential consumption nodes are aggregated to form population centers.

3.5.3 Effect of settlement patterns

Zvolef et al. [102] show that geography and by consequence settlement behavior affect network lengths. Similarly, Kocaman et al. [103] discuss that settlement patterns play a role in the results obtained from the two-level network design. In this section, we aim to understand the effect of real settlement patterns on our computed metrics.

Figure 3.7 shows four wards with varying settlement patterns, where each point represents a merged structure (points in close proximity might appear as a single point in the figure). The grey dashed boxes surrounding the structures represent a 25 km^2 box. The figure also shows the ward



(a) A, Nairobi (25, 2.8, 2194) (b) B, Garissa (27.4, 194.2, 145)

(c) C, TanaRiver (31.1, 39, 1991) (d) D, Embu (96.5, 40.8, 8071)

Figure 3.7: Four wards with varying settlement patterns are shown. In brackets are the per-structure LV requirement, per-structure MV requirement and the structure count of the ward, respectively. The grey boxes surrounding each ward represent a 25 km^2 area. Figure (a) and (b) show similar LV requirements with significantly different MV requirements. Figure (c) and (d) show varying LV requirements at similar MV requirements.

labels and their county name. In brackets we report the per-structure LV requirement (m), the per-structure MV requirement (m), and the structure count, respectively for the ward. Figures 3.7(a) and 3.7(b) show wards with similar per-structure LV requirements and varying per-structure MV requirements, while Figures 3.7(c) and 3.7(d) show wards with similar per-structure MV requirements and varying per-structure LV requirements. At similar per-structure LV requirements as seen in 3.7(a) and 3.7(b), the per-structure MV needed in ward A is 70 times lower than that needed in ward B due to the proximity of clusters. In Figure 3.7(b), significant MV is required to connect clusters of structures. These clusters may be villages or communities. However in Figure 3.7(a), all structures and their clusters are in tight proximity. The per-structure MV requirement in Figure

3.7(b) is even higher due to the smaller number of structures present in ward B when compared to ward A. At similar MV, Figure 3.7(c) has one-third the LV requirement of Figure 3.7(d). There is an even spread of structures throughout the 25 km^2 grid in Figure 3.7(d), which influences the per-structure LV requirement. With a higher structure count in Figure 3.7(d), it is expected that the per-structure LV requirement would be low as the total LV wire length and cost is spread out among a higher number of structures, however this is not the case. Because structures are more evenly spread out in ward D, the LV wire requirement is high. We observe that nucleation of structures drops the per-structure LV requirement while nucleation of clusters (villages, communities) reduces the per-structure MV requirement. We are able to show that our proposed connectivity metrics capture the effects of settlement patterns.

3.6 An Analysis on the Sub-administrative Boundary Level

We recognize that decision making about electrification technologies occurs at a granular level and that a single technology choice cannot be assigned to an administrative unit. As a result, we leverage the data and methodology for analysis at sub-administrative boundaries. To explore this in depth, we present the complete network for a sample ward of 7047 structures. Figure 3.8(a) shows transformer locations and the MV network for all the structures within the ward. The blue pentagons represent transformer locations, red solid line shows the MV network, and the grey points represent the structures. In Figure 3.8(b), we include the LV network (as green dashed lines) for a subset of the ward, showing connections between individual structures and transformers. Given our proposed methodology, the MV and LV network with individual connections can be visualized as demonstrated by the figure. Energy planners can inspect connections across transformers and structures and subsequently aggregate the metrics to a level that is most useful to support their decision making.

With our methodology we can identify which transformer locations and connecting structures can be networked with minimal LV wire. For the same ward, Figure 3.10(a) shows the number of structures per transformer at each transformer by quintile. Blue transformers are connected to

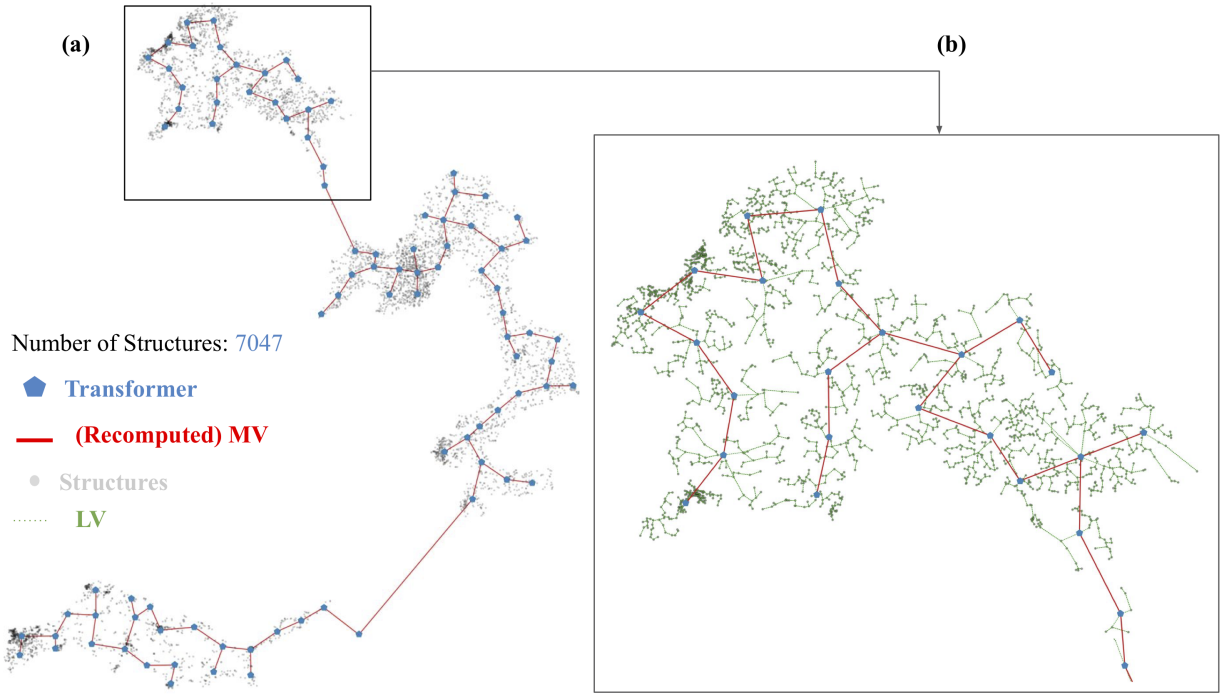


Figure 3.8: Complete network for a sample ward with 7047 structures. Figure (a) shows transformer placement and the MV network connecting the transformers. Figure (b) includes the LV network for a small section of the ward, showing connections between structures and transformers.

many structures while red transformers are connected to few structures. In the figure, we observe that transformers with few surrounding grey dots have a lower number of connecting structures, while transformers with many surrounding grey dots have a higher number of connected structures. Figure 3.10(b) shows the distribution of structures per transformer for all transformers in the ward. With a ward average of 77.5 structures per transformer, 10% of wards have more than 160 structures per transformer (twice the ward average). The distribution within the ward can be missed when only considering averages of our metrics along administrative boundaries or at lower resolutions. The flexibility to evaluate the proposed metrics at multiple scales allows for deeper evaluation of varying electricity technologies.

Using the same ward, we show that our methodology and metrics can be used to identify opportunities for varying electrification technologies. Table 3.2 presents four scenarios that align with the numbers presented in Figure 3.9(a) and Figure 3.10(a). Each scenario shows the combination of two of our metrics which may lead to a different electrification strategy. We refer the reader to

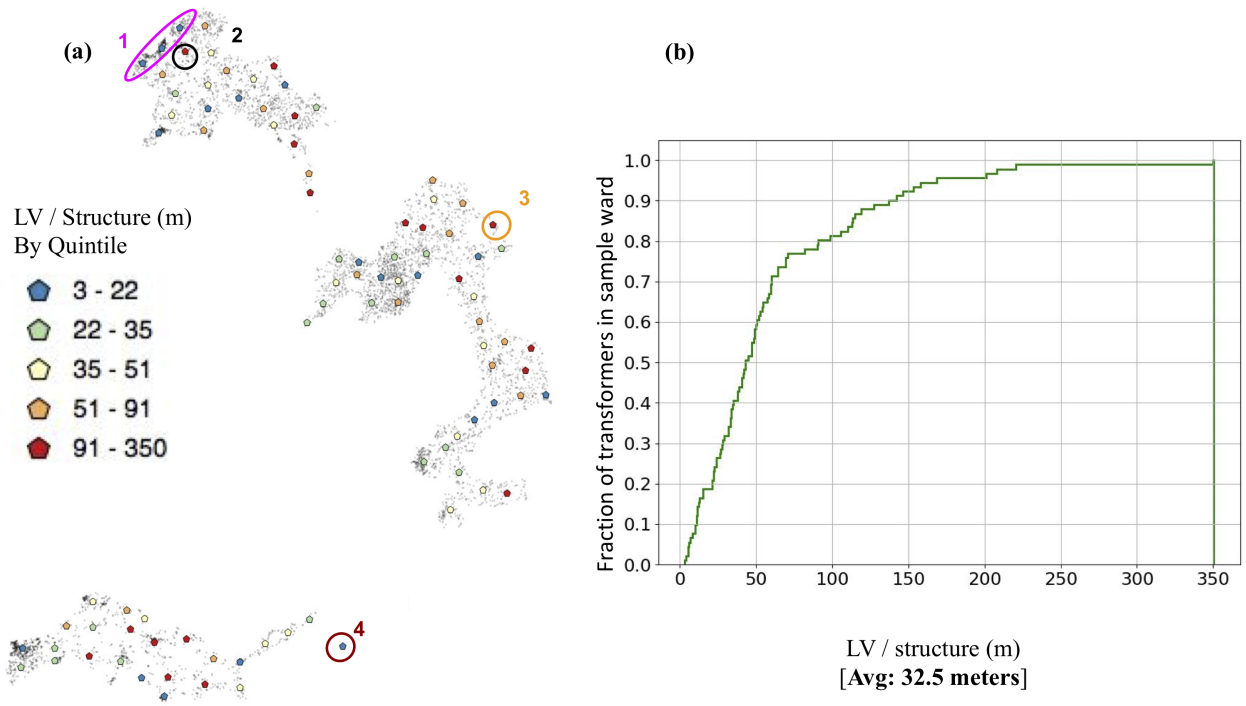


Figure 3.9: Low Voltage (LV) per structure, for each transformer in sample ward. a) Spatial distribution of LV per structure, binning transformers by quintile. b) CDF of LV per structure for all transformers in ward. The ward average is 32.5 meters. Four scenarios are presented, each with different implications for networking. See Table 3.2 for details

both Figures 3.9(a) and 3.10(a) for spatial visualization. In Table 3.2, the transformer colors are given in brackets for each scenario. Scenario **1** occurs when there are many structures connected to a given transformer and there is a small LV wire requirement for structures connected to the transformer. With a large number of structures connected to the transformer, the cost of the transformer is spread across multiple structures, thereby reducing the cost to any individual structure. Coupled with a low LV wire requirement, the choice of electrification is heavily dependent on the per-structure MV wire requirement. A low MV wire requirement suggests a centralized system like grid extension is a viable option for structures connected to these transformers. Scenario **2** shows there are many structures connected to a transformer but the structures are not clustered around the transformer.[5] Although the per-structure transformer cost is low due to high number of connecting structures, the high LV wire requirement becomes a major bottleneck to networking

[5]Note that we show 7047 structures which may appear as though they are in close proximity but represent multiple kilometers of coverage.

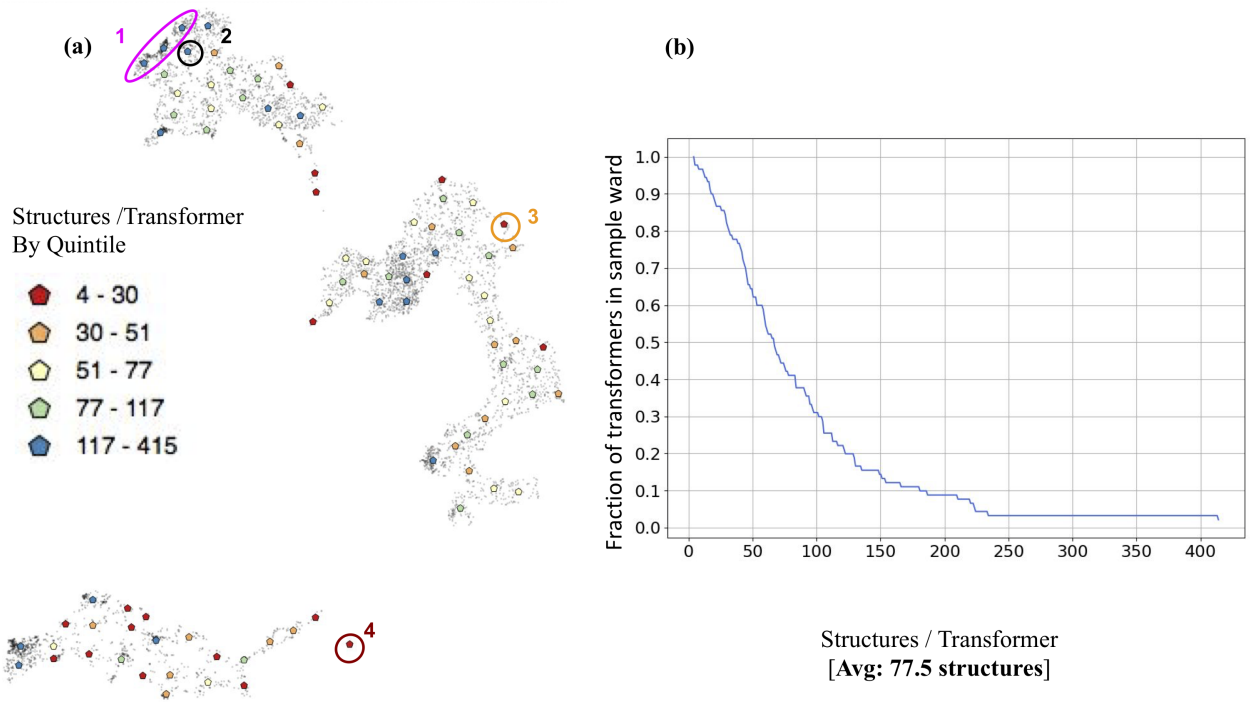


Figure 3.10: Number of structures per transformer, for each transformer in the sample ward. a) Spatial distribution of structures per transformer, binning transformers by quintile. b) CDF of structures per transformer for all transformers in ward. The ward average is 77.5 structures per transformer. Four scenarios are presented, each with different implications for networking. See Table 3.2 for details

this transformer and the structures associated to it. Solar home systems might prove to be suitable alternatives in this scenario. Scenario **3** presents a worst case scenario from a networking standpoint. Here there are few structures connected to the transformer and the structures are not in close proximity to each other. Similar to scenario 2, solar home systems might be worth considering as the cost to connect structures is high. Scenario **4** represents a case where there are few structures connected to the transformer, but the structures are in close proximity to each other and the associated transformer. In this scenario local generation and distribution through the low cost LV network would seem the most suitable approach. Because our approach uses individual structures, energy providers can explore the implications of networking at multiple resolutions, right down to the individual transformers. We do not show the MV wire metric at sub-administrative boundaries, as the existing grid network is needed in order to assign an MV wire requirement to a given transformer.

Table 3.2: Scenarios highlighting different electrification strategies which can be identified with our method.

Scenario	Structures / Transformer	LV / Structure (m)	Possible System(s)
1(Purple)	High (blue)	Low (blue)	Grid Extension or Minigrid
2(Black)	High (blue)	High (red)	Solar Home System (SHS)
3(Orange)	Low (red)	High (red)	Solar Home System (SHS)
4(Dark Red)	Low (red)	Low (blue)	Local Generation or Minigrid

3.7 Sensitivity Analysis

We evaluate the robustness of our proposed metrics by performing a cost sensitivity analysis. Table 3.3 presents our proposed metrics under 3 cost scenarios: i) baseline cost previously discussed, ii) double MV and LV wire cost iii) double transformer cost. The sensitivity analysis is performed on four previously presented wards A through D, first introduced in Section 3.5.3. From this sensitivity analysis we show that our proposed per structure MV, LV and transformer metrics are stable (less than 3 % change) under the three cost scenarios. We also observe that the wire cost is the primary driver of cost. This observation is apparent when doubling transformer cost results in less than 6.5 % change in the cost per structure across all four wards, while doubling wire costs, doubles the cost per structure across all wards.

Through this cost sensitivity analysis, we show that our proposed metrics can support infrastructure planning, where the actual unit wire and transformer installation costs (best known by the planner) can be directly multiplied by our metrics to obtain realistic cost estimates to support electricity infrastructure decision making.

3.8 Conclusion

In this paper we assess the effects of regional geography and settlements patterns on electrification strategies. By estimating the locations of residential structures through our proposed merging process, we are able to capture settlement behaviors of structures over a whole country. Through our novel computational framework that involves a network design algorithm, we develop a two-level distribution network between the structures. We present a a set of connectivity metrics on the

Table 3.3: Cost Sensitivity Analysis under three scenarios i) baseline cost (MV =$25/m , LV = $10/m, Transformer=$2000) ii) 2X MV and 2X LV wire cost, iii) 2X transformer cost. Sensitivity analysis is presented for 4 wards (A,B,C,D) previously in Section 3.5.3.

		Baseline Cost	2X Wire Cost	2X Transformer Cost
LV Per Structure	Ward A	25	25	25
	Ward B	27.4	27.4	27.4
	Ward C	31.1	31.1	31.1
	Ward D	96.5	96.3	96.5
MV Per Structure	Ward A	2.87	2.89	2.89
	Ward B	194.2	194.2	194.2
	Ward C	38.99	38.99	38.99
	Ward D	40.85	40.92	40.85
Structures Per Transformer	Ward A	137.12	137.13	137.13
	Ward B	29	29	29
	Ward C	32.1	32.1	32.1
	Ward D	14.7	14.6	14.7
Cost per Structure	Ward A	336	660	352
	Ward B	5198	10326	5266
	Ward C	1348	2634	1411
	Ward D	2123	4109	2259

wire requirements, number of structures on a transformer, and connection cost on a country level without sacrificing spatial resolution. We discuss that easily accessible metrics such as population density ignore the interplay between structure locations, and accordingly the true connection cost of a structure.

We demonstrate that metrics which capture settlement behavior are crucial when planning efficient electrification on a large scale. Meeting the targets set in SDG7 requires considerations of multiple consumers across large landscapes with varying settlement patterns and our proposed metrics can easily be folded into existing planning approaches to support these objectives. In addition, thanks to its scalability, our framework can support decision making at a granular level by recommending electrification strategies such as solar home systems, mini-grids and grid.

Our future efforts will involve relaxing some of the assumptions made in this work. Relaxing the assumption on uniform consumption would potentially lead to different network outcomes and would allow for variable transformer sizing. We also intend to capture existing grid infrastructures in our planning approach, for all settings have some initial network backbone that influences optimal electrification strategies. Finally, in our current implementation, the two-level network does not account for environmental and topological constraints such as protected areas, rights-of-ways, and elevation. As we believe these constraints would influence the medium voltage computation, we aim to incorporate them in future work.

3.9 Appendix

3.9.1 Merging Approach

We considered various merging radii to merge the 11.9 million identified building structures.

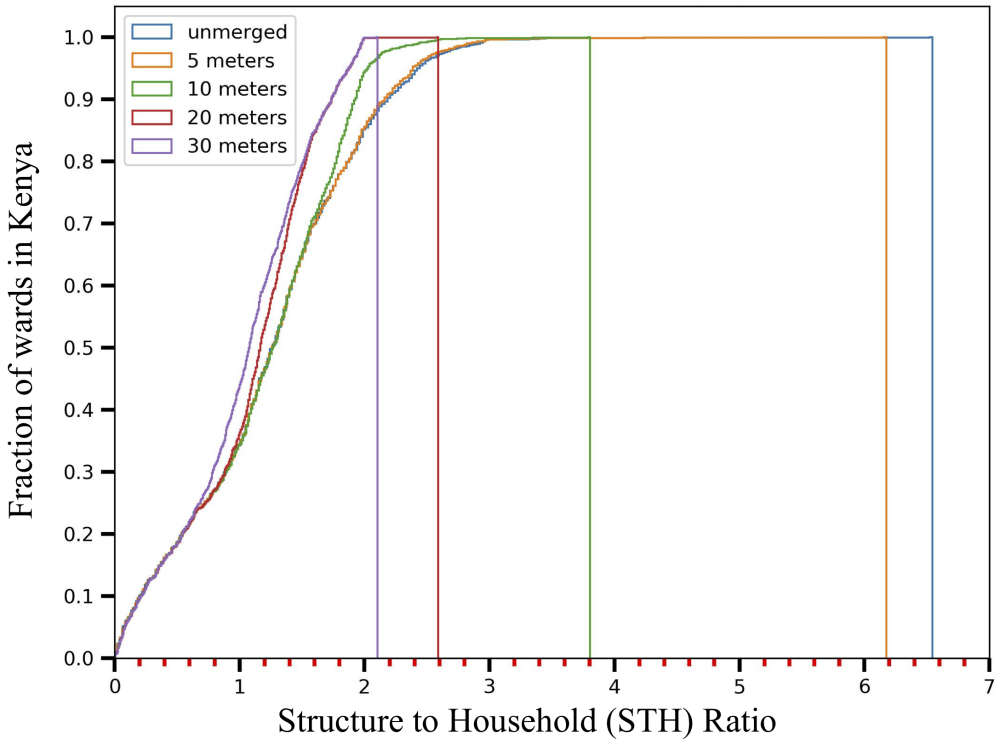


Figure 3.11: CDF of STH ratios for all wards in Kenya under varying merging radii.

Figure 3.11 shows the effect of merging on STH ratio under varying merging radii. We see that

the maximum STH ratio is 6.5 for unmerged structures, with multiple wards well above 2 structures per household. This implies that at the worst case, for a specific ward, every household has about 6 structures. We believe this estimate to be wrong as it does not account for other building types (commercial, industrial, etc). For merging radii from 5 to 30 meters, we observe a drop in the STH ratio, where at 20 m and 30 m, the maximum STH ratios are 2.6 and 2.1 respectively. We decided on the 20 m merging radius because it reduced the STH ratio for wards with exceedingly high STH ratios, without compromising those wards with STH less than 1. In the case of a merging radius of 30 m (as seen by the purple line), the STH ratios of less than 1 were further depressed.

3.9.2 Sensitivity to Scaling Strategy

We evaluated our framework by looking at some wards under vary split configurations. The selected wards were split into 4, 9, 16 and 25 cells and the TLND was applied to each cell. The runtime, per-structure low voltage, per-structure medium voltage and transformer capacity for the split configurations were evaluated against the unsplit ward. In this experiment, we only control the number of cells generated and do not apply limits on the number of structures in the cell or the cell radius. Figure 3.12 shows the worst case completion time in hours for five wards split into the aforementioned number of cells. The worse case completion time represents the completion time for the cell that took the longest to run. The computational time is cut by more than half for 4 of the 5 wards when the ward is split into 4 cells. Subsequent splitting further improves the completion time for the 4 wards.

The computational time for Kendu Bay in Figure 3.12 oscillates as the number of cells increases, although the worst case always takes less time when the ward is split than when it is left unsplit. To better understand this oscillation, we looked at the number of structures for the cell with the longest runtime in each of the wards. Figure 3.13 shows the number of structures under varying splits for the cell with the longest completion time. Capping the number of structures in a cell (M) at 3000 structures, significantly decreases the completion time. In our computational framework our choice for the hyper-parameter M was 3000 and thus ensured that large wards were split to

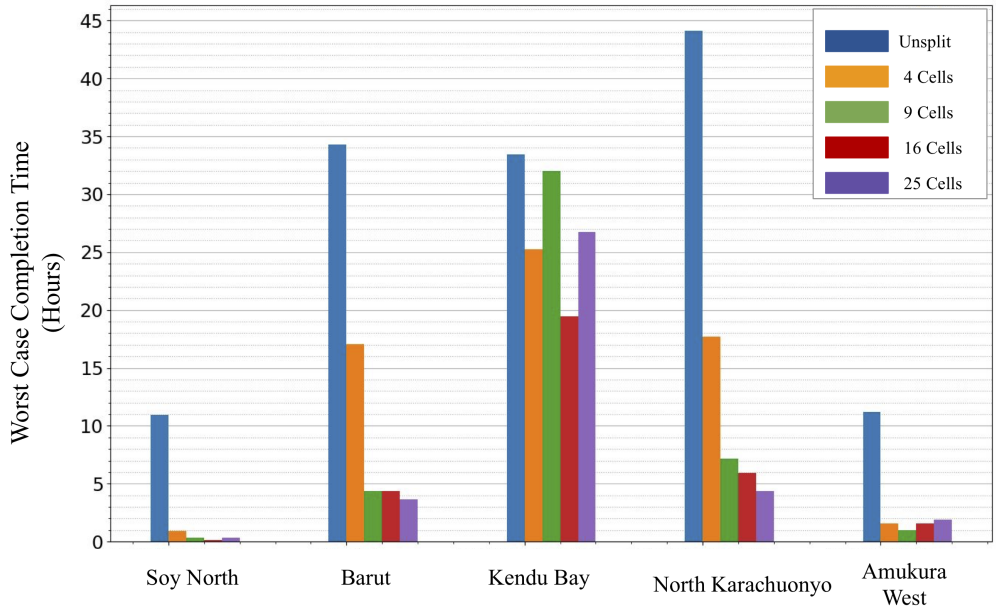


Figure 3.12: Completion time of the TLND in hours for the cell that took the longest time. Four out of five times, splitting a ward into 4 dropped the completion time by half.

cells with manageable number of structures. Revisiting Kendu Bay ward, where completion time oscillated, we observed from Figure 3.13 that dropping the number of structures in the cell is not the only contributing factor to completion time. Figure 3.13 suggests that the settlement pattern or spatial layout of structures within the cell influences the completion time. It also suggests that without enforcing minimum limits on the cell radius R, over-splitting a ward can have negative effects thereby increasing the computational time. Thus we used a minimum cell radius of 500m to stop over-splitting and capped the maximum number of structures in the largest cluster (N) at 300. This ensured computational gains while minimizing degradation in performance of our metrics.

Figure 3.14 shows our average connectivity metrics for 5 wards under varying split approaches. The figure also shows the results when the algorithm is run on the whole ward using the **Unsplit** label. These average connectivity metrics are obtained by first summing the metrics across all cells in a ward, then normalizing the sums by the number of structures in the wards. Figure 3.14 (a) and (b) show that our LV and MV connectivity metrics are not heavily influenced by splitting the ward into cells and applying our reconstruction strategy. However, we notice that the number of structures per transformer varies under different split strategies and tends to drop as we increase

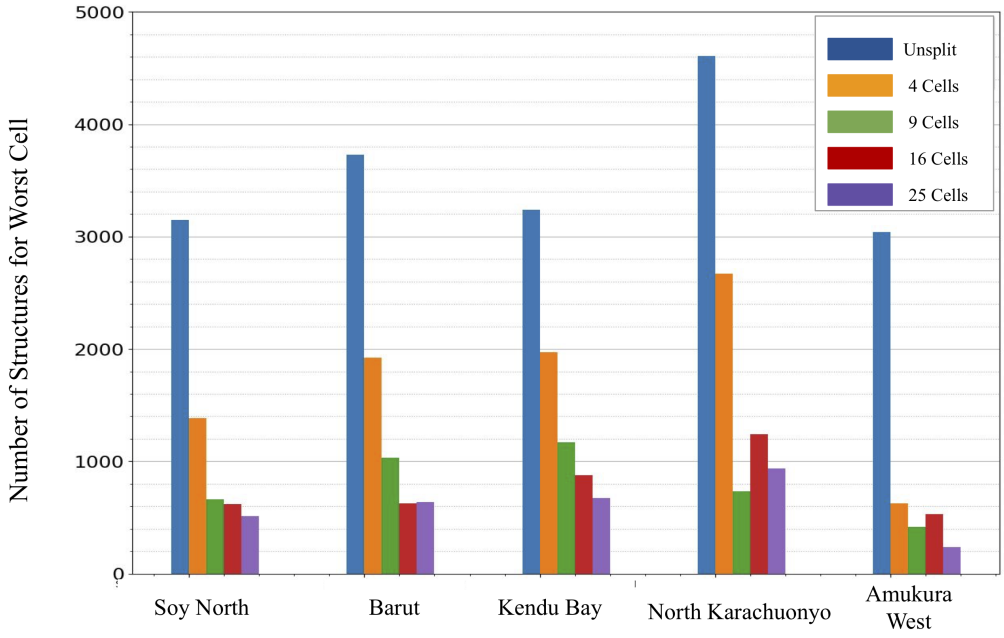
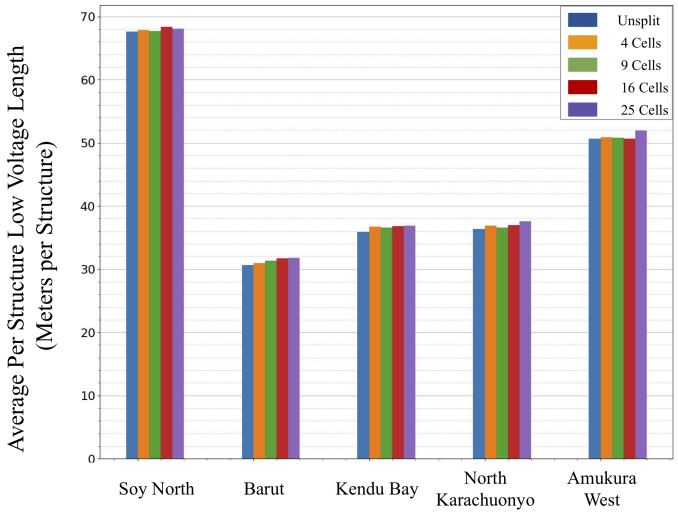


Figure 3.13: Number of structures for the cell will the longest run time. Splitting decreased the number of structures. However, number of structures is not the only driver of completion time. As in the case of Kendu Bay, spatial layout of structures also influences the computational time.

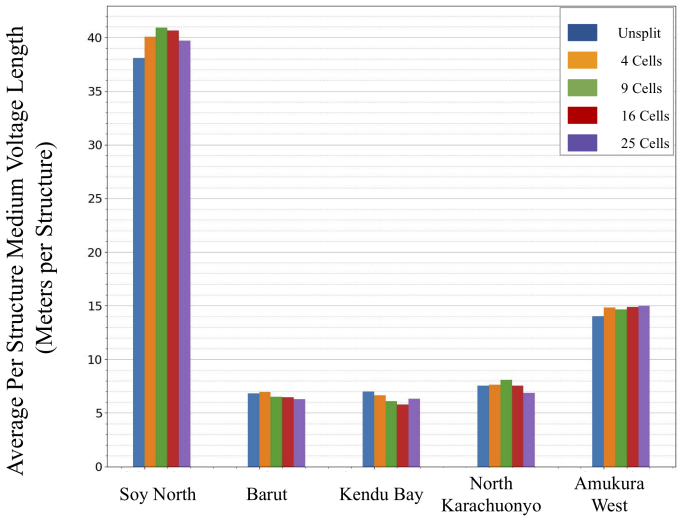
the number of cells a ward is split into. From these wards, we observe that transformers tend to be more under-loaded as the number of cells increase. We apply a minimum radius R in our splitting algorithm to prevent excessive splitting, thereby ensuring that the number of structures per transformer is maximized.

3.9.3 Code

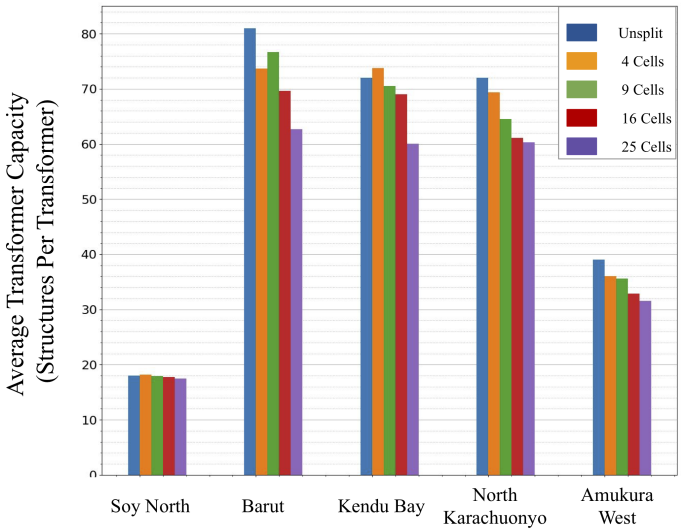
The repository and code needed to replicate this work can be found here: `https://github.com/SEL-Columbia/two_level_grid_network_planner`.



(a) LV after Reconstruction



(b) MV after post-processed Recomputation



(c) Number of structures per transformer after Reconstruction

Figure 3.14: Effect of splitting and MV reconstruction on our proposed connectivity metrics. The two-level network design is applied to each cell. Averages for the ward are reported here.

Chapter 4: High resolution estimates of household electricity usage as a proxy for household overall expenditure

The first three chapters of this thesis have focused on analyzing electricity consumption growth overtime, predicting future electricity consumption levels for unelectrified households and measuring the impact of settlement patterns on grid connection costs. This chapter departs from the electricity access question and rather focuses on how monitoring and evaluation of socio-economic indicators can be performed given electricity data.

We currently find ourselves in a data revolution, where the volumes of global data are in the zettabytes. These large volumes of data bring with them budding opportunities to extract new insights about human development across multiple indicators. The data, stemming from both private and public sector can be coupled with new analytical approaches to better measure impacts of investments and progress towards sustainable development goals.

Despite this explosion in global data products, many emerging economies still lag behind in the acquisition, storage and usage of valuable datasets needed to support rigorous and recurrent monitoring and evaluation of multiple socio-economic indicators. In contrast, governments in an attempt to meet sustainable development goals are providing millions of new customers with access to electricity amongst other services. This chapter evaluates the effectiveness of re-purposing electricity usage data collected by utilities to provide new insights to other domains such as economic well-being.

4.1 Introduction

Indicators of human well-being and access to services are critical for measuring the impact of investments and guiding development policies. Countries invest billions of dollars annually to

improve access to electricity and water services, modernize agriculture, provide relief to vulnerable groups , all with the objective of meeting the Sustainable Development Goals. Data-driven approaches to guide and monitor the impact of such investments over long temporal horizons can support better evidence-based decision-making. [113]

A lack of sufficient high resolution data to monitor and inform investment remains a key challenge[114]. Household survey deployment has traditionally been the approach to collect detailed ground data about household access to services, wealth and expenditures. However, because surveys are an *active*[1] data collection approach, they only capture a one-time snapshot of a household every few years. While surveys may be nationally representative samples of the population, their implementation is costly, the data collection process is time consuming and the survey collection can sometimes be too slow to be useful. Moreover, the sample sizes and spatial sampling of these surveys do not support decision making at high resolutions (sub-administrative level). Census data which captures every household, is collected once in a decade, thus can not serve as a good alternative for frequent evaluation.[115]

What if high frequency, *passively* collected data could be leveraged for improved policy making and service delivery? *Passively* collected data is data already being collected by governments and various institutions, without the need to deploy further resources or surveys for collection. Examples of *passively* collected data include electricity, water and mobile phone usage. *Passively* collected data presents an opportunity to re-purpose already collected data to answer new questions. [116] show how mobile phone data can be used to improve targeting of humanitarian aid to vulnerable households. While mobile phone usage was not design for poverty measurement, this work demonstrates how passively collected data can be repurposed to answer pertinent questions in new ways.

In this work, we leverage already collected electricity usage data, which electric utilities have access to, and analyze how well it lends itself to estimating other indicators. Specifically, we first estimate household overall consumption expenditure using electricity usage data from Rwanda,

[1]Surveys require the deployment of additional human resources to collect one data point compared to approaches that continually collect data

showing that higher electricity consumption is correlated with wealthier households. Income, consumption expenditures, and wealth are the three indicators typically used to ascertain the economic status of a household. Household income can be difficult to measure as it is self-reported thus may not reflect the broad ranges of non-salary based earnings [117]. This work relies on survey reported household overall consumption expenditure using the Fifth Integrated Household Living Conditions Survey (EICV5) from Rwanda. The survey is used to establish an approach to estimating household overall consumption expenditure from electricity usage data. We then present and evaluate our machine learning based approach to predicting electricity usage for individual households given high resolution daytime satellite imagery. Predictions from satellite imagery provide a pathway for non-governmental stakeholders (e.g. businesses, market analytics providers, investors, researchers, national bodies etc) to repeatedly and independently measure electricity usage and by consequence household wealth, given small amounts of label data for model training.

While decision-making can be improved by higher resolution data, predictions at the highest resolution (individual building or household) may be fraught with more error and privacy concerns. However, at the lowest resolution (country, province or district level), the variance within the administrative level is not captured, thereby leaving out a diversified set of solutions which may be relevant for different groups within the administrative level. Thus, we discuss our predictive performances at multiple resolutions, shedding relevant insights on how to preserve both performance, privacy and resolution to support decision making.

This work is situated within the context of two bodies of literature: one that establishes the relationship between electricity usage and economic development and the other that evaluates the use of non-conventional data sources and machine learning to predict socio-economic indicators. Electricity usage is often positively correlated with economic growth, where countries with higher electricity usage also tend to have higher per capita GDPs. There is however conflicting literature on whether electricity consumption leads to economic development or vice versa. Some studies have found that increased grid-access and electricity usage improves well-being indicators such as income, consumption, respiratory health, education and overall expenditure [118, 119, 120, 121].

Other have found little or no short to medium-term impact of electrification on economic indicators [8]. While the nature of causality between electricity usage and economic benefits remains harder to tease out, there is overall agreement about there exists a correlation between electricity usage and well-being indicators [122]. In this work, we seek to exploit this correlation, to estimate household overall expenditure as a proxy for wealth, using already collected electricity usage data.

With advances in machine learning and the growth in remote sensed products, there has been an explosion of methods estimating different household indicators. Multiple works [66, 67, 68, 69, 70, 71, 72, 123] estimate wealth and poverty from satellite imagery. Survey collected wealth data is used as a supervisory signal to the proposed models to guide the extraction of relevant features from satellite imagery. Other studies [124, 125, 126, 127] use Call Detail Records (CDR) from mobile phones to demonstrate that at the individual level, CDR data is predictive of household wealth. Social media data, in combination with satellite imagery has also been used to map socioecnonomic indicators [70, 128]. As an alternative to survey measured household characteristics or CDR, we evaluate the value of electricity consumption data to estimating household wealth. Within the context of predicting electricity usage, [129, 47, 130] use daytime satellite imagery to predict electricity usage of individual buildings. Only one of these studies is performed in a low-access region, where large scale electrification is still being carried out.

Our work shows that electricity consumption data can be valuable in estimating other socio-economic indicators such as household overall consumption expenditure. This work also shows that actual electricity consumption can be predicted from daytime satellite imagery with good model performance at the individual building level. Finally, we show that aggregation can preserve prediction performance and privacy of the households.

4.2 Data Overview and Processing

This work is centered on the link between electricity usage and household overall expenditure, first establishing the relationship between both and then predicting electricity usage from daytime satellite imagery. To understand the relationship between electricity usage and household expen-

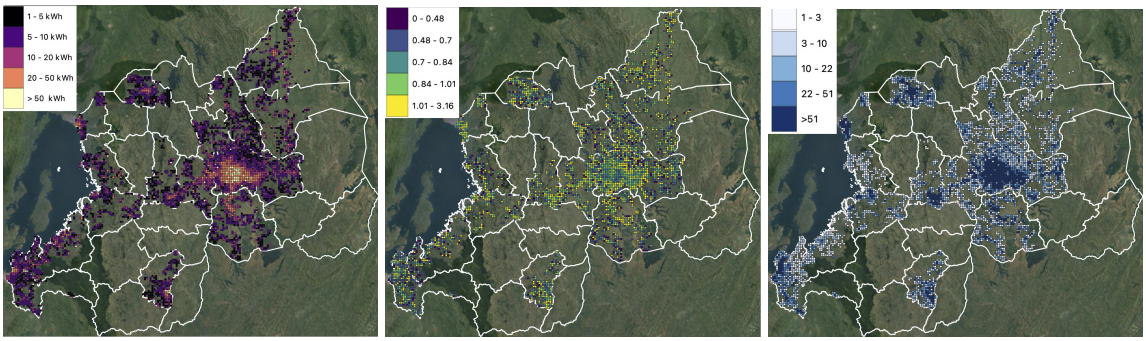
diture, we analyze the Fifth Integrated Household Living Conditions Survey (EICV5). To predict electricity consumption, we unify monthly household electricity consumption data from electric meters with daytime satellite imagery using household locations. In this section, we describe the relevant datasets used.

4.2.1 Electricity Data

The monthly electricity data used in this work was obtained from the Rwanda Energy Group (REG) who is the primary national electricity grid provider. Prepaid electricity purchases between 2012 - 2020 for 811K customers were provided by the utility of which 687K are residential customers. Each residential consumer id and meter id is matched to a separate dataset of meter locations also provided by the utility. However, only about half of the customers could be paired to GPS registered electric meters. For customers with corresponding GPS locations, prepaid transactions were converted to monthly electricity consumption (kWh) by spreading the purchased units over the days between two consecutive purchases. The daily consumptions were then aggregated over each month to obtain monthly electricity consumption. For periods where the duration between 2 consecutive purchases was greater than the median purchase frequency (days) for a given customer, the consumption was spread over the median purchase frequency. The conversion to monthly consumption values ensured that the customer's aggregate consumption was preserved.

Having obtained monthly electricity consumption for every customer, customers with the same electric meter GPS location (given REG's meter location dataset) were grouped to obtain the average monthly building consumption in each year. The building GPS locations were matched to Digital globe satellite imagery, and only buildings with imagery obtained between 2017 and 2020 were selected. Given the average monthly building electricity consumption in a year, the electricity consumption data was clipped at the 2nd and 98th percentile to remove outliers. A final filtering step was applied to obtain residential single customer buildings, resulting in 176,081 single household buildings.

To support good spatial sampling and model evaluation at multiple resolutions, Rwanda was



(a) Average monthly electricity consumption (b) Coefficient of variation (c) Number of households

Figure 4.1: 1km X 1km grid cell statistics, showing the spatial variation in average monthly electricity consumption, the variation in electricity consumption within the cell and the number of households in each cell

split into 1km x 1km grid cells. Figure 4.1 shows for every grid cell i) the average monthly electricity consumption ii) coefficient of variation of monthly electricity consumption iii) the number of households within the grid cell. Figure 4.1(a) shows that the highest monthly electricity consumption is experienced in Kigali with smaller pockets in other districts while average monthly electricity consumption outside of the city is mostly lower than 20 kWh/month. Despite low average electricity consumption, about 80% of the cells have a coefficient of variation greater than 0.5, suggesting that there exist a distribution of household consumption within the 1km x 1km grid cells, despite low averages. This distribution is at least 50 % of the average monthly cell electricity consumption. Finally, the household counts figure show that about 20 % of the cells have very few households (1 - 3). The objective of the predictions, would be to capture both the correct cell means while adequately preserving the variation within each cell.

4.2.2 Remote Sense Data

Satellite Imagery: High resolution 50 cm daytime satellite imagery obtained from Digital Globe is paired with GPS locations from electric meters to make predictions. The high resolution 50 cm daytime imagery obtained between 2017 and 2020, contains four image band (NIR, Red, Green and Blue). DigitalGlobe imagery provides country-wide coverage containing only a single image per tile (there are no temporal images for the same tile). To train a model to predict

electricity consumption from satellite imagery, satellite imagery acquisition dates were matched to average monthly electricity consumption occuring in the same image year. The value of the NIR band is discussed when all 4-bands (NRGB) are used compared to using 3-bands (RGB).

High Resolution Electricity Access (HREA) data: The HREA dataset provides annual composites of statistically estimated brightness levels at 15 arcseconds resolution. These brightness estimates are derived from temporal analysis of nighttime light imagery from the VIIRS sensor, dating back to 2012. The statistically estimated brightness levels as suggested by the authors is an indicator of outdoor lighting usage, which can be correlated with overall energy consumption. From the HREA dataset, only cells with recorded brightness levels are selected. This dataset is used to predict household overall expenditure and is compared to the performance of utility data when estimating household overall expenditure.

4.2.3 Rwanda Fifth Integrated Household Living Conditions Survey (EICV5)

The EICV5 survey is a nationally representative survey of Rwandan households. The fifth iteration of the survey was collected between October 2016 and October 2017, with responses from 14,580 households. The survey provides information on household demographics and well-being such as poverty, inequality, living conditions, education, housing conditions, household electricity consumption, overall household expenditure amongst others. While the survey reports household information, the GPS locations of the households are not shared. The survey however reports the corresponding districts of the households. Note that there are 30 districts in 26,338 square kM of Rwanda. This work looks at two indicators i) overall consumption expenditure of households and ii) expenditure on electricity. Overall consumption expenditure represents the value of good and services purchased by a household in a given year. This expenditure includes rent, expenditure on food, water, electricity and more, and is given in Rwandan Francs (RWF). Secondly, this work also looks at the expenditure on electricity by households connected to the national electricity grid (3,589 households). This set of households does not include households using electricity primarily from solar systems. Only grid connected households are considered to ensure a fair comparison

with our utility dataset. The survey's electricity expenditure is reported in RWF. Tariffs (fixed tariff and the 2017 block tariff implemented in Rwanda) account for the differences between the survey electricity expenditure (RWF) and the utility reported electricity consumption (kWh). The survey data is used to establish the relationship between electricity usage and household overall expenditure. This relationship provides justification for leveraging electricity usage data for varying purposes such as household wealth estimation.

4.2.4 Model Data Split

Here we discuss our data splitting strategy which is used to test the transferability and robustness of our predictions. 1 square km grid cells are created to support spatial sampling across the whole country. In order to train our models, 4577 (90 %) grid cells were selected for in-sample train, validation and test, while leaving 518 (10 %) grid cells are withheld for out-of-sample testing. The buildings found in the withheld out-of-sample cells are never seen during model training and validation, and are only used to validate the transferability of our models to unseen regions. For the in-sample grids, buildings within the grid cells with electricity data were split into 70 % train, 20 % validation and 10 % held-out testing. The distribution of electricity consumption from meter readings in the overall metering data was preserved in each set. Figure 4.2 shows the spatial sampling used to obtain the in-sample and out-of-sample set.

4.3 Methods for Estimating Electricity Consumption from Satellite Imagery

4.3.1 Model Architecture and Training:

Predicting average monthly electricity consumption from daytime satellite imagery was done using a Convolutional Neural Network (CNN). Varying CNN architectures have been proposed to support computer vision analysis. For this work, MobileNets were used due to their attractive lighter weight architecture. The encoder portion of the MobileNet architecture was combined with a light-weight predictive head consisting of 5 dense layers. Four of the 5 layers were activated with Rectified Linear Units (ReLus) while the last layer had a linear activation.

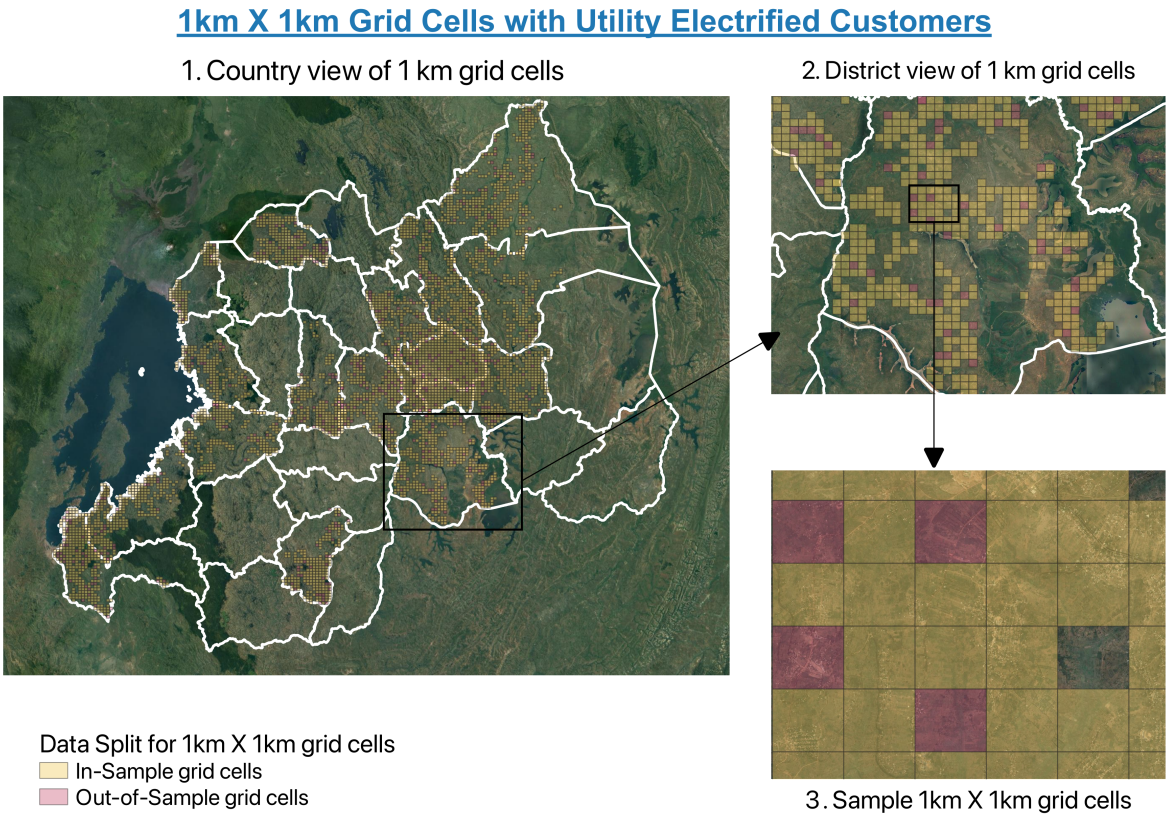


Figure 4.2: Spatial sampling of in-sample and out-of-sample sets using 1km X 1km grid cells with utility electrified customers.

For model training, a 64 x 64 m image patch is fed into the network where the network predicts the average monthly kiloWattHour consumption of the building found at the center of the image patch. An L1 norm loss (Mean Absolute Error) was used to train the model

$$MAE = \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{4.1}$$

, where y_i is the true monthly electricity consumption and $\hat{y}_i$ is the predicted electricity consumption. An Adam optimizer at a learning rate of 1e-5 proved to be the best in minimizing the L1 norm. A minibatch size of 16 images was used to train the model at each iteration for 20 training epochs. The model loss converged after 20 epochs with no additional gains observed.

4.3.2 Metrics:

Model performance is evaluated under both a regression and classification lens. Three regression based metrics are used for evaluation: i) Mean Absolute Error (MAE) which measures the average absolute error between the true monthly consumption and the predicted monthly consumption and ii) Mean Absolute Percentage Error (MAPE) which measures the average absolute percentage error between the true monthly consumption and the predicted monthly consumption iii) the R^2 which measures how well the model captures the variability within the electricity consumption data.

Classification metrics are obtained by binning the predicted consumption values and computing the class accuracies and F1-scores. The class boundaries are set at the 50 % percentile for the binary classification and at the 33rd and 66th percentile for the 3 class classification.

4.3.3 Results aggregation:

While predictions are performed at the individual building level, we report our model performance at three aggregation levels: 250 x 250 m, 500 x 500 m and 1 x 1 km grid cells. To perform aggregations, the average monthly consumption of single customer residential buildings within each grid cell (at a given level of aggregation) is computed using the utility reported electricity consumption. Average predicted monthly consumptions for the same single customer residential buildings are also computed. The average predicted consumptions are compared with the utility predicted consumptions at the given aggregation level. The regression metrics are then applied to the grid averages to obtain the reported performances.

4.4 Results

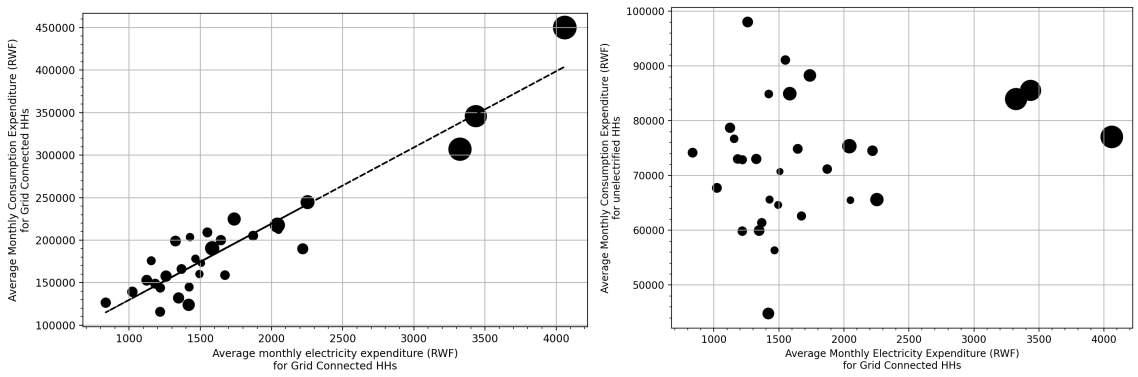
We present our results in two parts: First we present how well our method lends itself to measuring overall household consumption expenditure. Household consumption expenditure is defined as the total money a household spends on goods and services within a given period of

time. Household overall consumption expenditure is an indicator of the household's purchasing power and wealth. Thus the first part of our results discusses how well our method can estimate household wealth given electricity usage data. In the second part of the results we present how well our method predicts average monthly household electricity usage (kWh) in a given year. Here we discuss model performances at varying resolutions and under varying metrics.

4.4.1 Measuring household consumption expenditure

A study of 22 Sub-Saharan African countries reveals that about 3% of household total expenditure goes towards electricity expenditure [39]. We validate this relationship in Rwanda by analyzing a nationally representative survey of Rwandan households taken between October 2016 - October 2017. The Integrated Household Living Conditions Survey 5 (EICV5) obtained from the National Institute of Statistics of Rwanda (NISR) sampled 14,580 households and reports both household annual expenditure on electricity and overall expenditure. While household locations are not provided, the survey indicates the corresponding district of each household. Thus we measure district-level correlations between expenditure on electricity and household overall expenditure for all 30 districts in Rwanda. Figure 4.3 shows district-level average monthly electricity expenditure relative to household overall consumption expenditure in 30 districts for both grid connected households and unelectrified households in the survey. Note that while unelectrified households do not spend on electricity, Figure 4.3(b) uses the electricity expenditure of grid connected households to see how is agrees with the overall expenditure of unelectrified households.

Figure 4.3(a) supports the well-studied relationship that electricity expenditure correlates well (adjusted R-squared 0.98) with overall expenditure at the district-level and thus electricity consumption data can serve as a good proxy for overall household expenditure for grid connected households. Due to the lack of household GPS locations, this relationship could only be validated at the district level. Nevertheless, the relationship between electricity usage and overall consumption expenditure is well captured by the linear relationship, where districts with higher average electricity usage also have higher overall consumption expenditure and vice versa. A linear model



(a) Grid connected (3589 households) (b) Unelectrified (9778 households)

Figure 4.3: District-level correlation between average monthly electricity expenditure and average monthly overall consumption expenditure for grid connected and unelectrified households.

(as shown by the black line) was fitted to the district-level data. A Mean Absolute Percentage Error (MAPE) of 10.6 % is observed between the survey reported overall consumption expenditure and the approximated overall consumption expenditure from the linear model. Given the self-reported nature of the survey, we anticipate that part of the discrepancy between the linear model and the survey values may be due to self-reporting errors. Nonetheless, this shows that a simple linear model correctly approximates the relationship between electricity usage and household overall expenditure.

Assuming that district-level electricity expenditure (for grid connected households) might be useful in measuring overall expenditure in unelectrified households, we show in Figure 4.3(b) the average district-level household electricity expenditure for grid connected households against household overall consumption expenditure for unelectrified households. The figure shows no observable district-level correlation between average monthly electricity expenditure (in grid connected households) and overall household expenditure for the unelectrified. This behavior comes as no surprise, where electricity expenditure of electrified households can not be used as a proxy to measure household overall expenditure in unelectrified households.

Currently, about 65 % of Rwanda is electrified, we expect the relationship between electricity usage and overall consumption expenditure to be more relevant as nations push for universal grid connections. It would be critical to validate the strength of the correlation between electricity and

consumption expenditure at higher resolution (though beyond the scope of this work, as household locations are not provided).

Estimating District-level Overall Expenditure with Utility Data

Having shown (through the independently collected survey), that there exist a strong correlation (at the district-level) between electricity usage and overall consumption expenditure, this section discusses the performance of utility collected electricity usage data in estimating overall consumption expenditure.

Utility Data The utility data, is a collection of electricity consumption for 176,081 households between the periods of 2017 - 2020 obtained from Rwanda Energy Group (See Section 4.2.1 for in-depth data description and processing). The electricity usage data as reported by the utility is passively collected through electric meters deployed at homes. We have shown that there exist a linear relationship between survey reported electricity expenditure and overall consumption expenditure. Thus we use a linear model to estimate survey reported overall consumption expenditure from the utility reported electricity usage data [2]. Given the linear model, the MAPE between true district overall consumption expenditure and the model estimated overall consumption expenditure is reported. We observe a MAPE of 11.9 % when utility reported electricity usage is used as inputs to the model to estimate overall consumption expenditure. This compares to a MAPE 11.5 % when survey electricity expenditure is used, for the 18 districts with utility data. The 0.4% performance difference between utility reported electricity usage and survey reported electricity expenditure, shows that large amounts of passively collected electricity consumption data can be used to estimate overall consumption expenditure and by extension household wealth, thereby bypassing the need for repeated annual surveys.

Open Source Datasets We compare the performance of utility based electricity consumption with that of the High Resolution Electricity Access (HREA) dataset, a widely available proxy of

[2] In the appendix we show strong agreement between survey electricity usage and utility electricity usage

energy usage. Specifically, HREA reports statistically significant brightness levels derived from nightly VIIRS satellite imagery at 15 arcseconds resolution. This dataset is often correlated with overall energy consumption. Also using a linear model, we evaluate the ability of HREA to estimate district-level reported overall consumption expenditure. We observe a MAPE of 14.7 % and 14.9 % for the linear model that takes HREA as its inputs and predicts consumption expenditure in all districts and the 18 REG districts, respectively. This suggests that while lower resolution datasets such as brightness levels might be indicators of consumption expenditure, better performance can be obtained by using the utility data. An added advantage of using utility reported electricity consumption is that overall consumption expenditure can be estimated at resolutions higher than 15 arcseconds. Moreover errors with non-utility data may increase with increasing resolution.

Table 4.1 summarizes these results, reporting the MAPE when different datasets are used to estimate overall consumption expenditure.

Table 4.1: Mean Absolute Percentage Error (MAPE) between model-based approximations of overall consumption expenditure and EICV5 survey reported consumption expenditure, when different datasets are correlated with survey consumption expenditure.

	All Districts	18 Districts
EICV5	10.6	11.2
Utility Data	NA	11.9
HREA	14.7	14.9

4.4.2 Measuring household electricity usage

Thus far, we have shown that electricity usage data can be relevant for measuring other indicators such as household overall consumption expenditure. In this section, we discuss our results from predicting residential electricity consumption using remote sensed daytime imagery. Beyond measuring wealth indicators, electricity consumption estimates are also relevant for multiple stakeholders (e.g. national bodies, marketing insights, investors etc) looking to understand current electricity usage at scale. Here we present performances of our regression-based predictive Convolutional Neural Network (CNN) that takes in image patches (64 x 64 m) and outputs the

kilowattHour consumption of single customer residential buildings. We report performances at varying resolutions and discuss the implications.

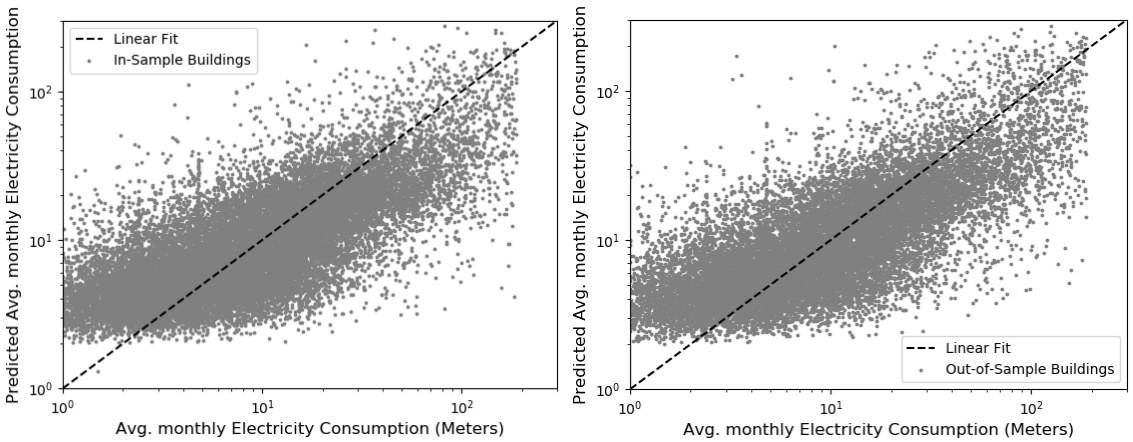
Predictive performance at varying resolutions

We present electricity consumption prediction performance at 4 resolutions (individual buildings, 1 x 1 km grids, 500 x 500 m grids and 250 x 250 m grids). Decision-making can occur at varying resolutions where trade-offs between error and electricity consumption heterogeneity (at a given resolution), exists. In the case of distribution transformer connection policies, utilities may choose to electrify household within a certain radius of the distribution transformer. That radius can range from 500 m to a few kilometers. Understanding predictive performance of electricity consumption models can provide useful guidance for such connection policies.

Predictive performance at the building level First we discuss performance at the individual level as the predictions are made on individual buildings. Figure 4.4 shows scatters of true average monthly electricity consumption for buildings against the predictions obtained with satellite imagery for the in-sample and out-of-sample test sets.

The regression-based CNN model is better able to differentiate the average monthly consumption for residential buildings consuming more than 10 kWh/ month. While the model is able to identify that a building is in the low range (< 10 kWh/month), the model cannot correctly estimate the actual consumption of the household, given the narrow range. This behavior is likely because of homogeneity in the satellite imagery of low consuming buildings in rural areas.

Regression Results: Table 4.2 shows the prediction performances for individual buildings under both regression and classification metrics. First we discuss the observations of model performance given regression metrics. For all residential buildings in our dataset, the MAE is close to 9 kWh/month while the MAPE is around 75 %, when making predictions for individual buildings. This suggests that at the individual level, the regression-based CNN records large relative prediction errors as shown through the MAPE but small absolute errors as shown through the MAE. Not-



(a) In-sample test (b) Out-of-sample test

Figure 4.4: Compares avg. monthly electricity consumption of buildings to those predicted using satellite imagery. The model is more sensitive to variability in consumption for buildings that consume on average more than 10 kWh/month. While it correctly places low consuming buildings (<10 kWh/month) in the below 10 kWh category, it is not as sensitive in differentiating household below that cutoff.

ing an average monthly consumption of 18 kWh/month, and a standard deviation of 25 kWh/month for this dataset, the MAE of 9 kWh/month indicates average prediction errors are within a third of the standard deviation. The next observation is that when households with consumptions < 3kWh/month are removed, the MAPE improves from 77 % to 52 % for the out-of-sample test set and from 75 % to 51 % for the insample test set. This indicates that the recorded high MAPEs are mainly driven by lower consumers where error represent a larger proportion of their consumption. A third observation is that including a 4th image band (NIR) does not significantly improve the model performance. The MAPE remains comparable when using three versus four image bands. Finally, given the scatter in Figure 4.4, R^2s range from 0.6 - 0.64 and show that the individual predictions capture around 60 % of the variability in the data. All R^2s showed high statistical significance ($p_{values} < 0.001$).

These regression metrics suggests that predicting average monthly electricity consumption for individual buildings from satellite is more feasible for higher consuming buildings. For low consuming buildings, it is hard to know whether the consumers are truly low or whether other contemporaneous variables may be at play (e.g. the building is a vacation home with occasional tenants).

Table 4.2: Prediction performance for individual buildings, reported for the In-Sample and Out-Of-Sample Test Sets. Regression metrics are reported for the CNN under 3 image bands (RGB) versus 4 image bands (NRGB)

		In-Sample Test	**Out-of-Sample Test**
Regression *All Buildings* (3 Band: RGB)	R2	0.60	0.61
	MAE	8.72	9.68
	MAPE	75	77
Regression *Buildings >3 kWh/month* (3 Band: RGB)	R2	0.61	0.61
	MAE	10.76	11.81
	MAPE	50.7	51.7
Regression *All Buildings* (4 Band: NRGB)	R2	0.64	0.64
	MAE	8.63	9.49
	MAPE	76	78
Binary Classification *All Buildings*	Low (<= 10.1 kWh)	0.83	0.83
	High (>10.1 kWh)	0.78	0.74
	F1-score	0.80	0.78
	Accuracy	0.80	0.78
3 Class Classification *All Buildings*	Low (<= 5.8 kWh)	0.65	0.64
	Medium (5.8 - 17 kWh)	0.57	0.59
	High (>17 kWh)	0.61	0.65
	F1-score	0.62	0.63
	Accuracy	0.61	0.62

Nonetheless, the regression model was able to extract useful features from satellite imagery to differentiate electricity usage.

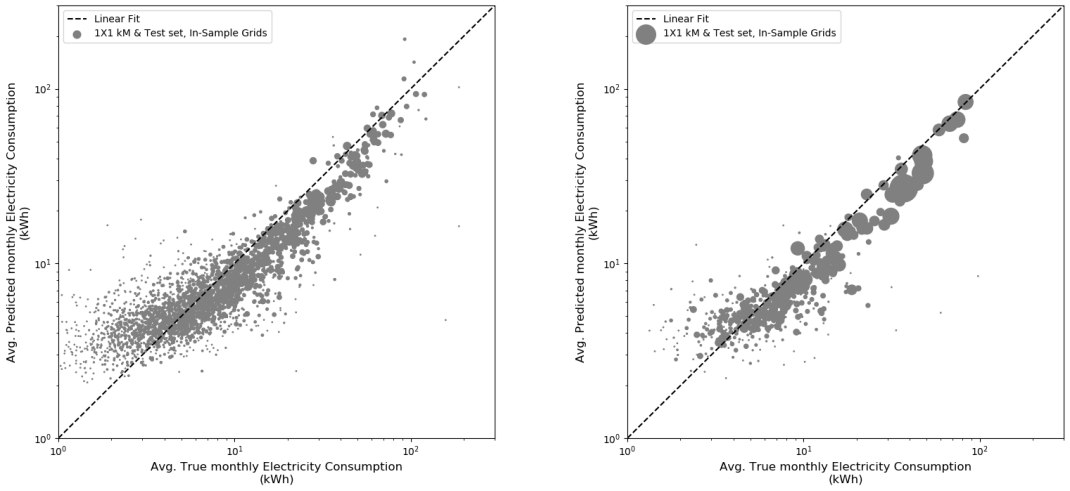
Classification Results: The regression results suggest that high predictive performance is harder for lower consuming households. Thus, we evaluate how classification metrics would perform. Table 4.2 also shows the classification performance when the predictions are binned into 2 and 3 classes. For the binary classification task, we observe a prediction class accuracy of 0.83, when identifying buildings consuming less than 10 kWh/month and a class accuracy of 0.78 when

identifying buildings consuming more than 10 kWh/month, for the in-sample test set. The 10 kWh/month threshold is selected because this represent the 50 % percentile of the data, thus allowing the identification of the top half versus bottom half of consumers. Compared to the regression results, this suggests that while images may not predict the actual building consumption with high fidelity, identifying consumption classes for individual buildings is feasible at high performances (balanced accuracy and F1-score of 0.8 and 0.8, respectively for the in-sample test set). The 3-class problem presents a harder task, where electricity consumption is thresholded at the 33rd and 66th percentile, to obtain low, medium and high groups. It is observed that the predictions are better at identifying the lower consumers (<5.8 kWh/month) while the middle group remains the hardest group to identify. For the 3 class problem, a balanced accuracy of 0.61 and 0.62 is observed for the in-sample and out-of-sample test sets.

While making predictions for individual buildings provides the highest resolution into electricity usage, there exist higher errors when predicting low consuming buildings. These errors drop for higher consuming buildings. An alternative to reporting actual kilowatt-hour values would be to estimation consumption classes, as this is an easier task with better performance.

Performance Aggregated at 1 kM x 1 kM grid cells Individual building predictions provide the highest spatial resolution though the mean absolute percentage errors are high due to larger errors for low consumers. Here, we analyze the value of aggregated results on performance. Results are first aggregated to 1 sqkm, where the utility reported averages of the grid cells are compared with the averages of predictions within the grid cells. Note that predictions are made at the building level and residential buildings within the cell are averaged for comparison. At this resolution, we make a few key observations. Aggregating the results to 1 sqkm significantly improves the performance. The R^2 improved from 0.6 to 0.81 for the in-sample test set and from 0.61 to 0.83 for the out-of-sample test set. Figure 4.5 shows this agreement between utility electricity measurements and our predictions, for the in-sample and out-of-sample test sets, aggregated at 1km resolution. Each bubbles represents a 1 sqkm grid cell and the bubble sizes represent the number of residential

buildings within our dataset found in the grid. Larger bubbles are grids with more residential buildings. (See Section 4.2.4 for further description on the 1 sqkm grid cells).



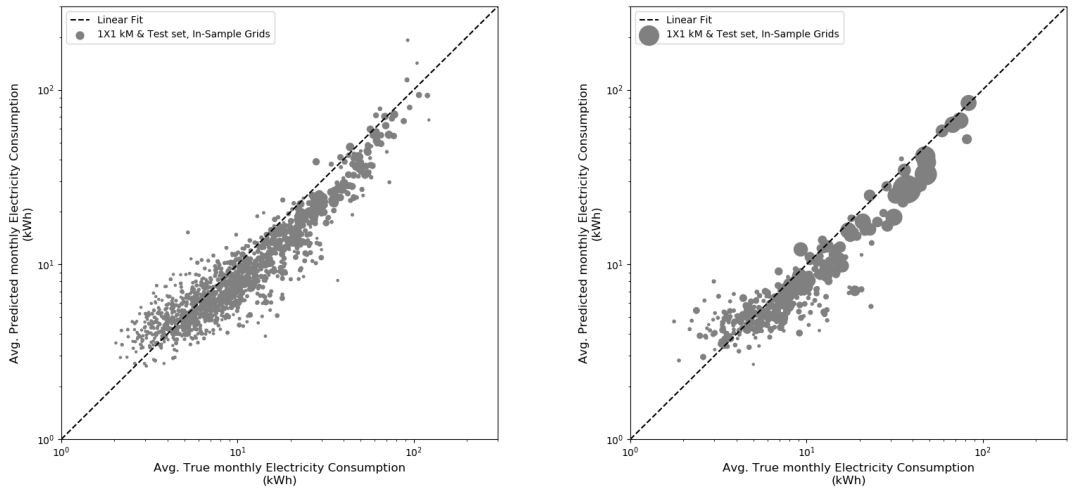
(a) In-Sample Test (2528)

(b) Out-of-Sample Test (495)

Figure 4.5: Agreement (of 1kM grid cells) between average monthly electricity consumption from electric meters and the predicted averaged monthly electricity consumption using imagery. Each point represents a grid cell and the size the number of households in a cell.

From the Figure, we also observe that grid cells with few number of customers (smaller bubbles) occur further away from the linear fit line, suggesting that aggregation is not helpful for grid cells with fewer customers. This observation is intuitive as there are fewer examples within the grid cell that the predictive model could learn from and also fewer examples to smooth the noisy predictions. By removing grid cells with fewer than five samples the observed R^2 increased to 0.9 and 0.95 for the in-sample and out-of-sample test, respectively. Figure 4.6 shows the scatter when grid cells with fewer than 5 samples are removed. Removing grid cells with few households, reduces the number of grid cells by 47 % and 29 % for the in-sample and out-of-sample test, respectively. However, the model performance at this resolution is better when there are more samples within the grid. As more households get electrified, we can expect that fewer grid cells will have a low number of households or buildings for the model to learn.

To further understand performances by electricity consumption levels, we evaluate perfor-



(a) In-Sample Test (1341 grids)

(b) Out-of-Sample Test (350 grids)

Figure 4.6: Agreement (of 1kM grid cells) between average monthly electricity consumption from electric meters and the predicted averaged monthly electricity consumption using imagery. Each point represents a grid cell and the size the number of households in a cell.

mances for different consumption ranges. Here we try to unpack which consumption ranges are hardest to estimate via satellite imagery.

Table 4.3 shows detailed prediction performance for different electricity consumption ranges. The table shows the MAE and MAPE for each consumption range. In addition, the table shows the number of grids within that range, the true average consumption for the grids (not weighted by sample size) and the total number of buildings in each grid, for grids with at least 5 buildings. By breaking down the performance by consumption groups we observe that the MAPEs are highest for grid cells with average consumptions less than 4 kWh/month. However, very small MAEs around 1 kWh/month are also observed for this group. The higher MAPEs are observed because the consumptions are themselves very small for this group.

In contrast, while the MAEs for the 4-10 kWh/month group is also around 1 kWh/month, the MAPEs for this group are about half those of the lowest tier. This suggests that while aggregation improves performance overall, the gains are smallest where the model itself struggles to predict correctly. Because model predictions at the building level are worse for very small consuming

Table 4.3: Prediction performance aggregated at 1 km for grid cells with at least 5 buildings. Results show both the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE) for in-sample and out-of-sample test sets.

All	**MAE (kWh/month)**	**MAPE**	**# of grids**	**Avg. electricity consumption (kWh/month)**	**# Individual Buildings**
In-Sample Test	3.1	24.3	1341	12.3	19053
Out-of-Sample Test	2.5	25.4	350	10.3	17590
1 - 4 kWh/month					
In-Sample Test	1.1	35.3	159	3.3	992
Out-of-Sample Test	1.3	45.2	53	3.2	950
4 - 10 kWh/month					
In-Sample Test	1.2	18.5	667	6.6	6528
Out-of-Sample Test	1.2	17.4	193	6.5	6914
>=10 kWh/month					
In-Sample Test	6.9	28.3	515	24.4	11347
Out-of-Sample Test	5.5	27.9	104	21.5	9682

households, this pattern is maintained even when buildings are grouped. Nevertheless, grids with consumptions between 4 -10 kWh/month have much lower prediction errors. For grids with consumptions >= 10 kWh/month, the aggregation also results in much lower MAPEs compared to grids with <4 kWh/month.

While we observe that aggregation allows for good estimation of average monthly electricity consumption, we evaluate the prediction variability within the 1 km grid cell, which gives further justification to the value of aggregation. Given the building level consumption values, the Coefficient of Variation (CoV) at the 1 sqkm grid cells is computed. The CoV captures the ratio between the standard deviation of the electricity consumption within the grid and mean electricity consumption of the grid. The true CoV reflects the variation using the utility reported electricity values for the households within the dataset and this is compared to the CoV given the predicted electricity consumptions of individual buildings. Comparing the CoVs gives an indication of how well the prediction model performs in capturing the variability in electricity consumption between the buildings in the same grid cell.

Figure 4.7 shows the CoV of the true consumption and that of predicted consumption for the

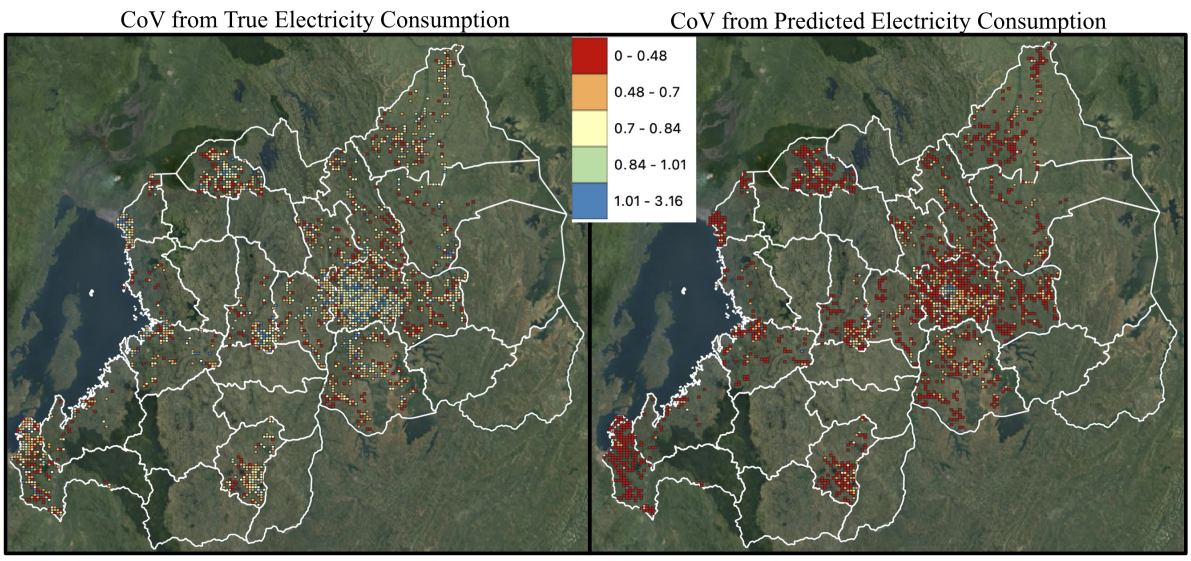


Figure 4.7: Coefficient of Variation (CoV) for each 1 kM grid cells from the in-sample test set. **Left:** CoV using true utility consumption data. **Right:** CoV using predicted consumption given satellite imagery. While average grid monthly consumption are accurately predicted, beyond the urban center, the model does not capture the variability within the grid cell.

in-sample test set. Outside the urban center of Kigali, the model struggles to capture with high fidelity, the variations in consumption within each grid cell. 76 % of grid cells have a true CoV greater than 0.48, while only 28 % of grid cells have a CoV greater than 0.48 when predictions are used. An L1-norm loss was used in training the model and this loss function optimizes for the average. The CoV shows that given this loss function, the model is better at estimated the average consumption of buildings within a 1 sqkm grid and does not perform as well in capturing the variability in consumption within the grid cell. This effect might be amplified due to the obvious differences between urbanized Kigali and more rural areas outside of Kigali. The differences in satellite imagery between urban and rural regions might pull the embeddings of rural areas closer together while pushing those of rural areas further from that of urban areas. As a result, electricity consumption differences between households in the same urbanization level are harder to differentiate using the CNN-based approached. This effect coupled with the chosen loss function might encourage the model to allocate the average consumption of the neighborhood to the individual building. One strategy to addressing this is to train separate rural, peri-urban and urban models to better support the learning of relevant image features for the urbanization level. Given this ap-

proach to model training, it suggests that aggregated results may be more valuable if kilowatt-hour values are of interest to energy planners while individual predictions lend themselves better to consumption classes.

Aggregating individual predictions, offers better spatial insights to electricity consumption, with good stability across the in and out-of-sample test sets. However, while the averages of the grid cell are accurately predicted with imagery, the variation within the grid cell especially at low consumption levels is harder to capture using daytime images.

Predictive performance across multiple resolutions Thus far, we have presented performances at the individual building level and for 1 sqkm grid cells. Here we compare performances at 5 resolutions: district, 1 x 1 km, 500 x 500m, 250 x 250 m and individual buildings. The district level performances are included as the previous section on estimating overall household expenditure was performed at the district level. 500 m and 250 m grid cells are also considered as they illustrate how performance scales with increasing resolution. For the 1km grid cells with at least 5 samples, we reported the performances for both in-sample and out-of-sample test set in Table 4.3. Using those same grids, we now discuss the performance in estimating the kilowatt-hour consumption at multiple resolutions. Figure 4.8 shows the MAPE at 5 spatial resolutions for the in-sample test set. The largest MAPE is observed at the individual level, while the smallest MAPE is observed at the district. Though each level of aggregation improved the estimation of average monthly electricity consumption, the largest performance gain occurred at 1 km. Beyond 1 km the gains from further aggregation where minimal. This suggests that average monthly electricity consumption values are optimally estimated at 1 km grid cell resolutions, though the models can be trained at the individual building levels.

4.4.3 Model Transferability

A key attribute of desirable models is their ability to transfer to unseen regions. We evaluate the transferability of our predictions by comparing the performance of the in-sample test set to that

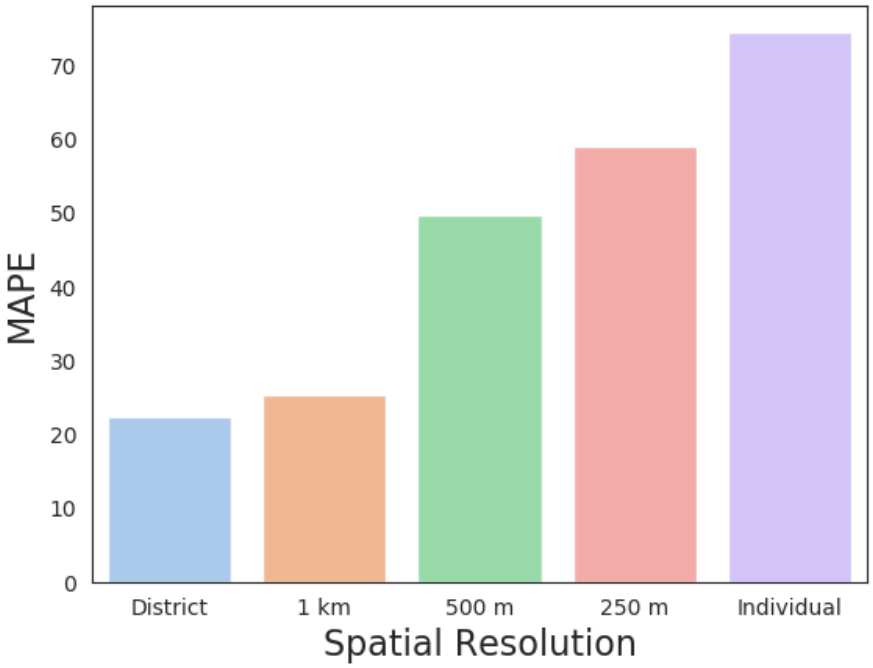


Figure 4.8: Mean Average Percentage Error (MAPE) between utility reported consumption and predicted consumption at varying resolutions. Aggregation reduces the MAPE, where predictions at individual buildings have the highest MAPE while those at the district level have the lowest MAPE. The largest gains from aggregation are observed at 1 km grid cells.

of the out-of-sample test set. The out-of-sample test set contains buildings from 1 km grid cells that were never seen by the model during training and validation. Across all metrics (classification versus regression) and at varying resolutions, the model performances for the in-sample amd out-of-sample test sets have been comparable. This suggests that the model is robust and can transfer well to new unseen regions in Rwanda. This feature is of importance as more households are being added to the grid, the model can be trusted to maintain a similar performance when used in the newly electrified regions.

4.4.4 Model Explainability

In this section we analyze the relationship between building characteristics and household electricity usage with the goal of understanding the visual features which the CNN model learns to make predictions. Specifically, we evaluate the building roof characteristics and surrounding in-

formation.

Do larger buildings consume more electricity? To start off this analysis, we first aim to understand the relationship between building roof sizes and electricity consumption. We compare the distribution of rooftop sizes for 3 electricity consumption groups. Figure 4.9 shows the distribution of building roof areas for three average monthly electricity consumption groups: < 10 kWh, 10 - 50 kWh and >50 kWh. The y-axis of the cumulative distribution curve shows the proportion of buildings with rooftop areas below a given threshold. From the Figure, we observe that households consuming > 50 kWh/month on average have a high likelihood for their building roofs to be larger than 100 square-meters. In fact, about 65 % of these buildings have roof sizes greater than 100 square-meters. On the other hand, more than 90 % of buildings using < 10 kWh/month on average have roof areas less than 100 square-meters. For the 10 - 50 kWh/month group the percentage is lower at about 80 % having roof sizes less than 100 square meters. This suggests that there are differences in roof sizes and by extension building sizes, where large electricity consumers are more likely to have larger buildings. The figure also suggests that the relationship between roof size and electricity consumption is non-linear, where rule-based functions may fall short in predicting electricity consumption from roof sizes. Thus using CNNs to learn useful non-linear relationships becomes more imperative when predicting electricity consumption from satellite imagery.

How important is the building compared to its surrounding context? Black box models provide an approach to extract patterns from the data which correlate with the indicator of interest. However, extracting human interpretable features that are relevant for predicting electricity consumption is not always evident. In this section, we provide some model explainability by comparing the relative performance of a model learning from building characteristics alone (building rooftop size and roof color) with a model learning from the building and its surrounding image pixels. To extract individual building characteristics, we leverage the Point Segmentation approach[77], which uses an indicator point to specify the building to segment. Using the GPS locations for buildings within our dataset, we specify the building to segment and can obtain the

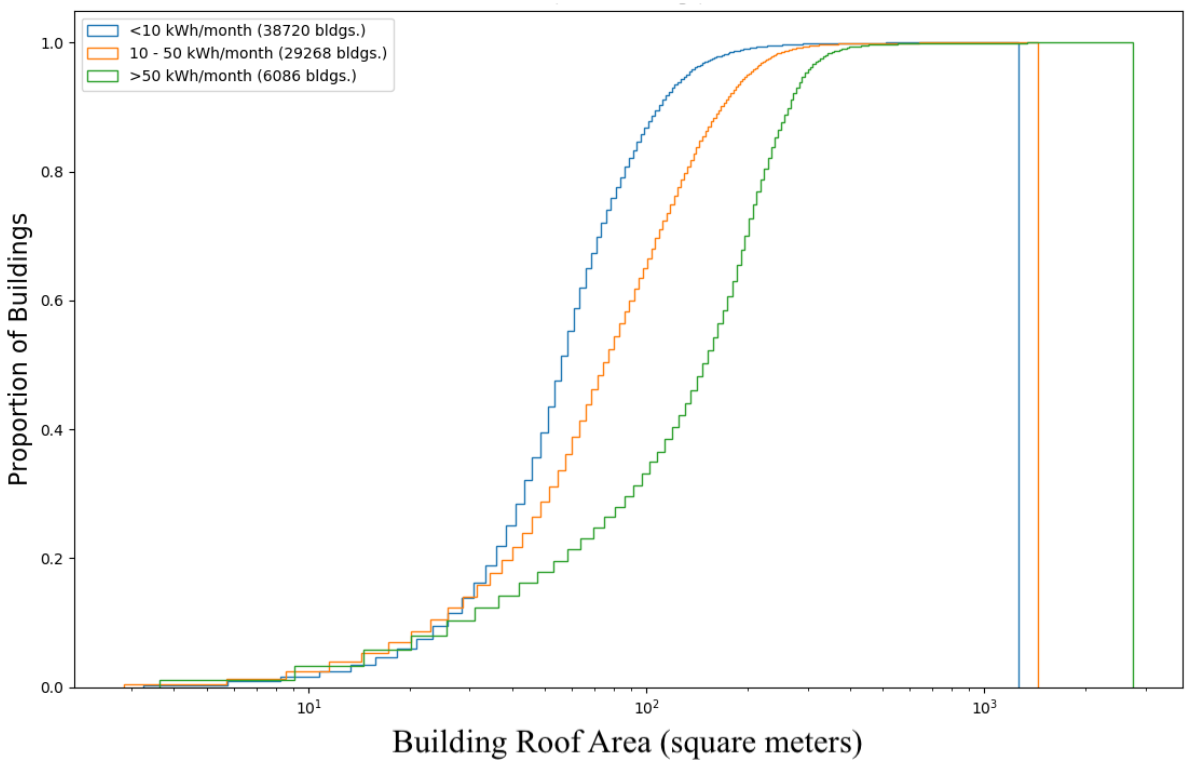


Figure 4.9: Distribution of single customer residential building roof areas as a function of 3 electricity consumption groupings, for 74K buildings. Higher electricity usage buildings also have a higher likelihood of having larger roof or building footprints and vice versa.

corresponding building roof size and roof color. Figure 4.10 shows sample buildings roofs extracted using the Point Segmentation approach. The dot or point on each building shows the input point to the segmentation model while the blue outline shows the extracted building roof footprint from the model.

Two models are compared: First is the MultiLayer Perception (MLP) binary classifier model that takes as inputs the extracted building roof size and the average roof pixel intensity for each band. The second is binary classifier CNN which takes in an image patch (64 x 64 meters) and predicts the consumption class of the household within the patch. The encodings from the previously trained regression model are used in this classification task with the only difference being that a dense layer with sigmoid activation is used as the final layer. The classification head is retrained while the encoding weights are kept frozen to ensure the same features are being used / analyzed. Classification is chosen over regression because the cross entropy loss function of the

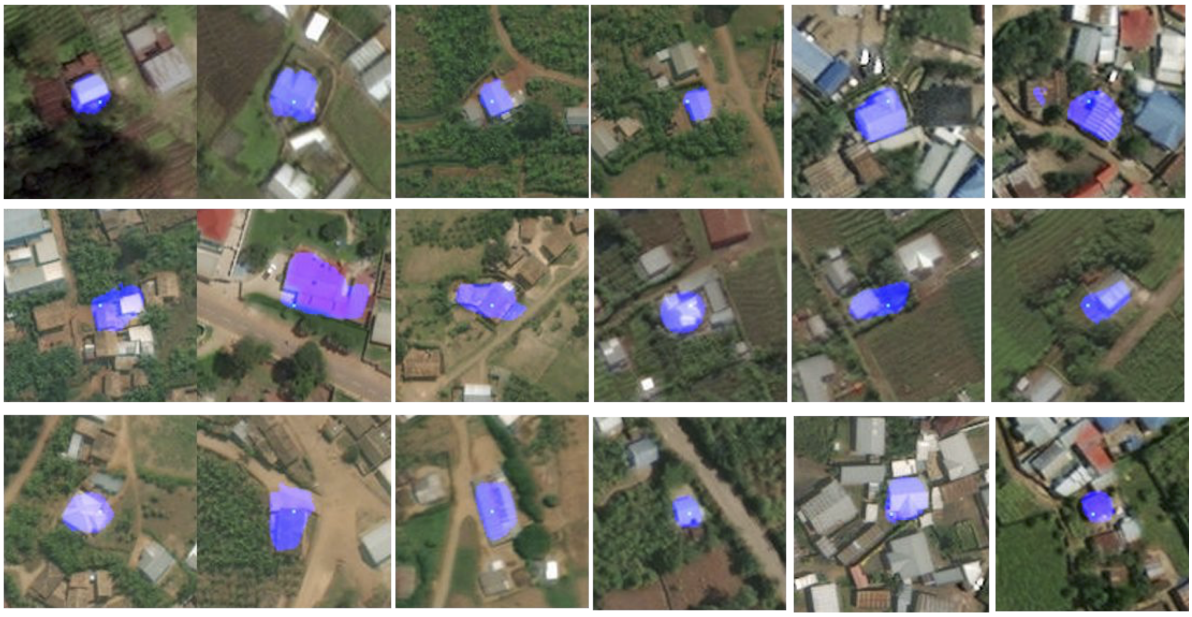


Figure 4.10: Sample building footprints extracted with the Point Segmentation method. Dots show the buildings of interest that are input into the model, while blue outlines shows model outputs as building roof footprint.

classier does not encourage averaging (an observation which was made when the regression model was used). The bottom and top quartile customers make up the low and high consumers, respectively. Specifically, low is defined as <=5 kWh/month while high is defined as >=31 kWh/month. This discontinuous class boundary is selected so as to minimize ambiguity between classes and to better tease out the contributions of building characteristics relative to those of the surroundings. Here we compare the results from the MLP with that from the CNN-based classifier.

Table 4.4: Compares classification performance when building roof size and color are used for prediction to performance when an image patch (containing the building and its surroundings) is used. Building characteristics yield a 0.77 F1-score while including surrounding context increases performance by 13 %. Results for the test sets.

	Low (<= 4.9 kWh/month)	**High (>= 31 kWh/month)**	**Accuracy**	**F1-Score**
Building Roof Size & Band Pixel Means (Method: MLP)	0.78	0.77	0.78	0.77
64 x 64 m full image patch (Method: CNN)	0.91	0.90	0.90	0.90

Table 4.4 shows classification performance for i) building roof size and color and ii) full 64 x 64 m image patch for the test sets. Using building roof characteristics yields a binary classi-

fication balanced F1-score of 0.77, however including information about the buildings surroundings (neighboring houses, fields etc), the binary classification F1-score increased by 13 %. This suggests that while the building size and roof color are strongly correlated with electricity consumption, the increased field of view which considers features surrounding the household offers additional informational relevant for distinguishing both classes.

Incorporating surrounding information into the model can be further enhanced with a few modifications. The chosen model and training configuration in this work, removes the spatial dependence amongst buildings of close proximity. An alternative would be to predict consumption for multiple buildings within an image patch. By not predicting the electricity consumption of multiple buildings within the same image, the current model does not incorporate how consumption varies within the neighborhood. Approaches such as Mask-RCNN[131], have been used to classify and segment multiple buildings within the image. This approach allows the model to see a large view of the neighborhood. By viewing multiple buildings within an image, the model can tease out key questions about the spatial heterogeneity of consumption within the neighborhood. This might also improve the individual building prediction performance and the CoV from predictions as results are aggregated. One bottleneck to address under the Mask-RCNN architecture is that to train such a model, electricity usage data for all buildings within the patch would be needed. [77] have shown that segmentation models can be modify to handle incomplete labels by specifying or "pointing-out" the specific building for segmentation. This modification can be incorporated into the Mask-RCNN architecture to support predictions when only partial electricity consumption labels are available within the patch.

4.5 Summary

This work makes two key contributions: First we show that residential electricity consumption is well correlated with overall household expenditure and that utility electricity consumption data can be used to estimate household expenditure for grid connected households. Our results at the district level show a 0.4 % MAPE difference between survey data and utility data when esti-

mating household expenditure. In addition, utility data performs better than open source datasets such HREA, at the district level. We expect utility data to be even more valuable at the sub-administrative level, where HREA measurements may be more subject to noise. While these results highlight the value of electricity usage data for measuring other indicators, there are further opportunities to solidify this observation. Validating household expenditure from utility data at the household level could not be done as there was no way to obtain income or expenditure information for the utility customers. National bodies could include a sample of known utility customers into their repeated surveys to better capture both their longitudinal wealth and electricity consumption. This could provide an approach to develop robust and spatially explicit household expenditure models from electricity usage data. Nonetheless in the absence of household locations linked to both wealth and electricity consumption, it is well established within the literature that households on average spend about 3 % of their income on electricity, thereby providing a rule-of-thumb to roughly estimate household overall expenditure from electricity usage data. This back-of-the-envelop approach provides a first pass at obtaining household expenditure from utility-based electricity usage data.

Secondly, this work establishes a methodology to estimate household electricity usage data from high resolution 50 cm daytime satellite imagery. Here, image patches showing the building of interest (centered within the patch), are input into the CNN to predict the actual consumption. Overall, the work illustrates that satellite imagery provides valuable features for predicting individual building consumption. Significant performance gains are observed with aggregation, where stakeholders can obtain high fidelity predictions at 1 sqkm resolution while preserving household location characteristics.

Finally, this work demonstrates strong transferability where similar performances are observed between the in-sample and out-of-sample test sets. This gives confidence that as more households from unelectrified parts are given electricity, the model with a bit of fine-tuning should continue to provide useful insights on electricity usage.

4.6 Appendix

4.6.1 Relationship between utility electricity consumption and survey-based electricity expenditure

Here we present the correlation between the our utility electricity consumption data and survey-based electricity expenditure, at the district level. Agreement is validated at the district level because the survey only provides household districts and does not report household GPS locations. Figure 4.11 shows the correlation between mean survey reported monthly electricity expenditure and mean utility reported monthly electricity usage. The figure shows that utility data strongly agrees with the survey data at the district level with R^2 of 0.9. This agreement between both estimates of electricity usage gives confidence that utility data can be directly correlated with survey reported overall household expenditure.

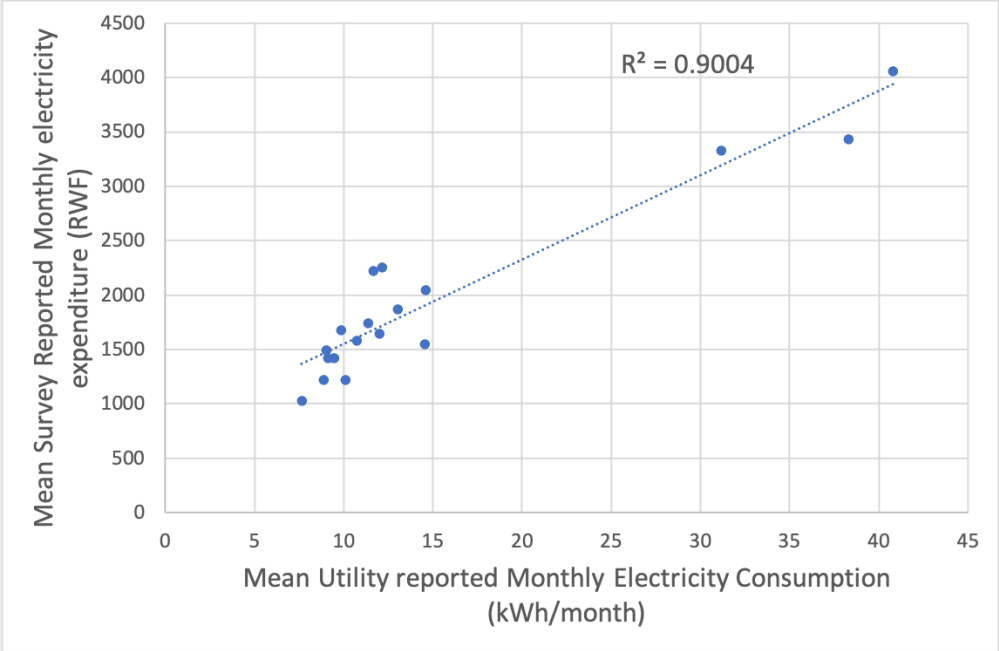


Figure 4.11: District level agreement between EICV5 survey reported electricity expenditure and utility reported electricity consumption. Each dot represents a district, for which there are 18 districts with utility and survey data

4.6.2 Relationship between electricity consumption and survey-based consumption expenditure

This section provides more insights on the agreement between electricity usage from utility data and survey reported consumption expenditure. In the main body of this chapter we report the MAPEs from linear models. Here we show the correlations and the best-fit linear models used to obtain the MAPEs. Figure 4.12 shows the agreement and best fit linear model between household electricity usage and household overall expenditure at the district level. Results from utility reported electricity consumption and satellite imagery-derived predictions, aggregated at the district level are shown. Both inputs correlate with overall household expenditure, though the predictions tend to underestimate the absolute electricity consumptions and by consequence the absolute overall expenditures. Nevertheless, the predictions preserve the relative electricity consumption and also relative household overall expenditure levels, the same districts are considered low or high in both the utility and satellite derived datatsets.

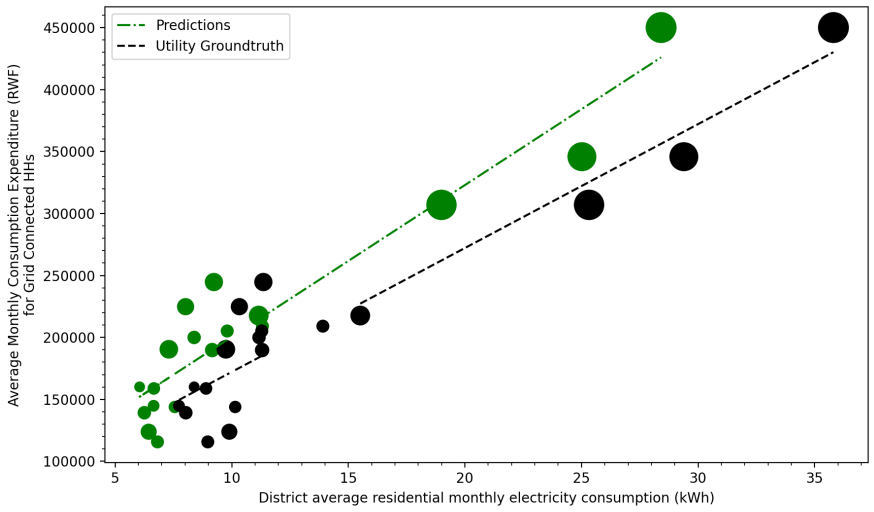


Figure 4.12: Scatter between household electricity consumption and survey reported household electricity expenditure at the district level. True utility and satellite-imagery derived predictions, aggregated at the district-level are shown. In general, absolute electricity consumption derived from satellite imagery tends to be lower than the true observed electricity consumption. Nevertheless, the relative electricity consumption levels amongst districts is preserved in the satellite imagery derived estimates.

Chapter 5: Learning to Segment from misaligned and partial labels

The previous chapters have demonstrated how utility data can be utilized to understand customers, determine optimal electrification strategies and predict household demand. CNNs were used to predict both consumption levels for unelectrified households and actual kilowatthours for already connected buildings. A recurrent theme throughout this thesis has been to uncover the relevant features driving the predictions from black-box CNN models. The main approach has been to understand building roof characteristics as seen by the model. Obtaining building roof characteristics is a largely solved problem in settings where clean label data is abundantly available. Within the context of the thesis work performed in Kenya and Rwanda, building footprint data especially in rural settings was not always available. Thus alternative approaches to segment and obtain building characteristics were needed in order to support the work. This chapter discusses a methodology to extract building footprints to support electricity access and usage studies, given household locations.

To extract information at scale, researchers increasingly apply semantic segmentation techniques to remotely-sensed imagery. While fully-supervised learning enables accurate pixel-wise segmentation, compiling the exhaustive label datasets required is often prohibitively expensive. As a result, many non-urban settings lack the ground-truth needed for accurate segmentation. Existing open source infrastructure data for these regions can be inexact and non-exhaustive. Open source infrastructure annotations like OpenStreetMaps are representative of this issue: while OpenStreetMaps labels provide global insights to road and building footprints, noisy and partial annotations limit the performance of segmentation algorithms that learn from them.

In this chapter, we present a novel and generalizable two-stage framework that enables improved pixel-wise image segmentation given misaligned and missing annotations. First, we introduce the Alignment Correction Network to rectify incorrectly registered open source labels.

Next, we demonstrate a segmentation model – the Pointer Segmentation Network – that uses corrected open source labels to predict infrastructure footprints despite missing annotations. We test sequential performance on the Aerial Imagery for Roof Segmentation dataset, achieving a mean intersection-over-union score of 0.79; more importantly, model performance remains stable as we decrease the fraction of annotations present. We demonstrate the transferability of our method to lower quality data sources, by applying the Alignment Correction Network to OpenStreetMaps labels to correct building footprints; we also demonstrate the accuracy of the Pointer Segmentation Network in predicting cropland boundaries in California from medium resolution data. Overall, our methodology is robust for multiple applications with varied amounts of training data present, thus offering a method to extract reliable information from noisy, partial data.

5.1 Introduction

Processing remotely-sensed imagery is a promising approach to evaluate ground conditions at scale for little cost. Algorithms that intake satellite imagery have accurately measured crop type [132],[133], cropped area [134], building coverage [135] [136], urbanization [137], and road networks [61] [138]. However, successful implementation of image segmentation algorithms for remote sensing applications depends on large amounts of data and high-quality annotations. Wealthy, urbanized settings can more readily apply segmentation algorithms, due to either the presence of or the ability to collect significant amounts of carefully annotated data. In contrast, more rural regions often lack the means to exhaustively collect ground truth data. Some open source datasets exist for such settings, and by successfully coupling these annotations with remotely sensed imagery, researchers can gain insights into the status of infrastructure and development where well-curated sources of these data do not exist. [139] [140].

Although these global open source ground truth datasets – e.g. OpenStreetMaps (OSM) – offer large amounts of labels for use at no cost, the annotations within suffer from multiple types of noise [141] [142]: *missing or omitted annotations*, defined as objects being present in the image and not existing in the label [141]; *misaligned annotations* occur when annotations are translated and/or

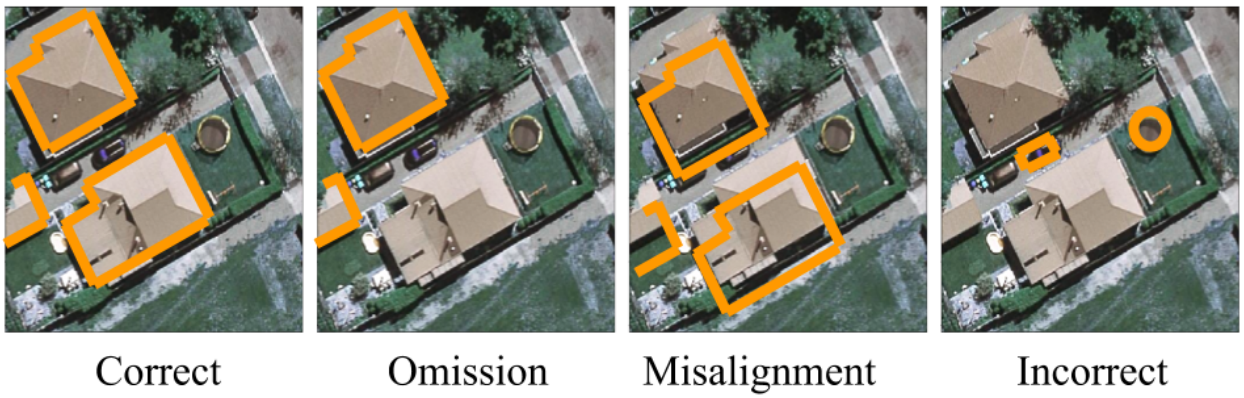


Figure 5.1: Types of label noise present in open source data. Building footprints are the class of interest.

rotated from its true position [143]; and *incorrect annotations* – annotations that do not directly correspond to the object of interest in the image. Figure 5.1 presents examples of these three types of label noise.

Noisy datasets present a training challenge when using traditional segmentation algorithms, as the model cannot learn to associate image features and target labels when the relationship is obscured by noise. To address the issues of misaligned and omitted annotations, and in order to extract information from imperfect data, we present a simple and generalizable method for pixel-wise image segmentation. First, we address annotation misalignment by proposing an Alignment Correction Network (ACN). With a small number of images and human verified ground truth annotations, the ACN learns to correct misaligned labels. Next, the corrected open source annotations are used to train the Pointer Segmentation Network (PSN), a model which takes in a point location and identifies the object containing that point. Learning associations from a representative point is a widely acknowledged method of object detection: [144] notes that an intuitive way for humans to refer to an object is through the action of pointing. By '*pointing-out*' the object instance of interest, our network ignores other instances that may not have corresponding annotations, therefore preventing performance degradation caused by annotation-less instances within the image. As a result, our sequential approach presents a method for handling misaligned data as well as

varying levels of label completeness without explicitly changing the loss function to compensate for noise. While our approach cannot replace large amounts of carefully annotated outlines, it can complement existing open source datasets and algorithms, reduce the cost of obtaining large amounts of full annotations, and allow researchers to extract information from imperfect datasets. This chapter's key contributions are as follows:

- We introduce the Alignment Correction Network (ACN), a means to verify and correct misaligned annotations using a small amount of human verified ground truth labeled data.
- We propose the Pointer Segmentation Network (PSN), a model that can reliably predict polygon boundaries on remotely-sensed imagery despite omitted training annotations and without requiring any bespoke loss functions.
- We demonstrate the applicability of our methodology to three different segmentation problems: building footprint detection with a highly-accurate dataset, building footprint detection with noisier training data, and cropland boundary prediction.

Taken as a whole, our approach enables resource constrained actors to use large amounts of misaligned and partial labels – coupled with a very small amount of human verified ground truth annotations – to train image segmentation algorithms for a variety of tasks. The rest of the chapter is organized as follows: In *Related Work*, we discuss related literature; in *Methods*, we describe our novel methodological contributions; in *Results*, we present results for the ACN and the PSN for all segmentation tasks; and in *Conclusion*, we restate our most salient findings.

5.2 Related Work

Computer vision researchers have recently made numerous advances in semantic segmentation, in applying state-of-the art techniques to remote sensed imagery, and in learning from noisy datasets; we discuss some important contributions to the literature below.

Existing Segmentation Approaches

Primarily based on improvements to deep convolutional neural networks (DCNN) architectures,

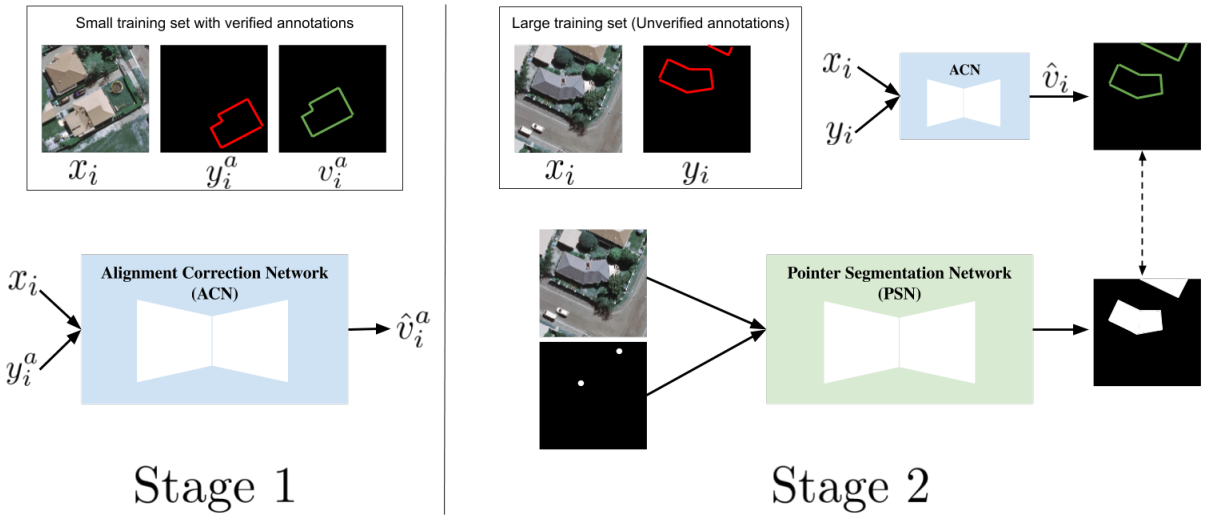


Figure 5.2: Summary of our two-stage approach to segment from noisy annotations. Stage 1: The ACN uses an image (x_i) and label (y_i^a) with a single misaligned annotation to predict a corrected annotation $\hat{v}_i^a$ containing the realigned annotation. Random shifts between ± 10 pixels are applied to v_i^a to obtain y_i^a. The network is trained with a small set of images (x) and verified ground truth annotations (v). Stage 2: A large noisy training set is first realigned with the ACN. Realigned, incomplete annotations are used for supervision. The PSN uses selected points from available instances, x_i and $\hat{v}_i$ to learn the segmentation task.

researchers have achieved record performances for a variety of different segmentation tasks. Fully convolutional encoder-decoder type architectures – one type of DCNN – take in an image and output a per-pixel prediction for the class of interest [145]. Some architectures use symmetric networks with skip connections to perform pixel-wise predictions [146] [147]. Alternatively, two-stage detection algorithms first perform region proposal – areas that have a high likelihood of containing the object of interest – and then detect objects within the identified regions [148] [149] [150]. Modifications to two-stage detection algorithms have enabled semantic segmentation of images, whereby individual pixels in an image are placed into one of a number of classes [131] [151]. Development of these segmentation architectures has been facilitated by large, comprehensive datasets which enable the implementation of these algorithms in a fully supervised approach: here, every object in the image and its corresponding annotation are used in the learning process [152] [153] [154].

Applying Deep Learning to Remote Sensed Imagery

Multiple projects have leveraged satellite imagery to answer various questions on land use, road quality, object detection, consumption expenditure: by linking sparse ground truth with abundant imagery, researchers can extrapolate trends in existing data to areas where labeled data do not exist [155], [156], [157]. Alternatively, some works have proposed neural network architectures that sidestep training data constraints and the relative lack of labeled ground-truth in remote areas [158] [159]. Jean et al. combine Google maps daytime images (provided by DigitalGlobe), nighttime lighting, and survey data to estimate poverty for multiple African countries [123]. High resolution daytime images were used to train a model to predict nighttime lights as measured by DMSP-OLS; features extracted from the last layer of the model were then used to estimate household expenditure or wealth. Results from this chapter suggest that predictions about economic development can be made from remote sensed data using features derived from imagery; this insight provides additional motivation for developing methods that extract information from noisy imagery datasets.

Learning From Noisy Annotations

The problem of poor-quality training data, especially in rural areas, for segmentation tasks is well

known: [160] acknowledge the variability in coverage of open source data in Kenya and observe significant degradation of coverage as one moves away from urban settings. Coverage degradation from urban to rural areas is also seen in South Africa[161], Brazil[162] and Botswana[163]. [164] estimates the effects of multiple types of training data noise, including misalignment and missing annotations, finding that as noise levels increase, both precision and recall decrease. For applications such as measuring building or field area which are useful in downstream analysis of wealth, crop yield and more, high noise levels decrease the ability to successfully use segmentation algorithms. Several works tackle the problem of learning from imperfect labels. [141] propose new loss functions to address noisy labels in aerial images. [143] [165] both focus on the issue of misalignment: [165] uses a self-supervised approach to align cadaster maps, and while the method proposed in [143] maximizes the correlation between annotations and outputs from a building prediction CNN, it assumes buildings in small groups have the same alignment error. Our two-stage approach builds upon existing convolutional frameworks common to many noise correction approaches. However our approach relies on the well-known binary cross entropy loss function, addresses both misalignment and omitted annotation, and does not require that all misalignments are identical. Thus serving as an attractive alternative when noisy labels are present.

5.3 Methods

Traditional segmentation methods take an image input x_i and aim to learn a function $f(\mathbf{x})$ that predicts a single channel label $\hat{v}_i$ containing all building instances present in the image. Equation 5.1 shows the learned function given x_i, where v_i^a is the single channel label of instance a in image x_i and there are a total of A instances in that image:

$$\begin{aligned} & f(x_i) \rightarrow \hat{v}_i \\ & s.t. \quad \hat{v}_i = \hat{v}_i^1 \cup \hat{v}_i^2 ... \cup \hat{v}_i^A \end{aligned} \tag{5.1}$$

5.3.1 Alignment Correction Network

Misalignment occurs when there is a registration difference between an object in an image and its annotation. In remote sensing, misaligned annotations may occur for a number of reasons, including human error and imprecise projections of the image [165]. There are two types of annotation alignment errors: 1) translation errors, where the annotation is shifted relative to the object, and 2) rotation errors, where the annotation is rotated relative to the object. [143] suggest that translation errors are more frequent for OpenStreetMaps in rural areas. Thus in this chapter, we only address translation errors present in open source data. We propose an Alignment Correction Network (ACN) that takes in an image x_i and a label y_i^a containing one misaligned instance a. The ACN outputs a label $\hat{v}_i^a$ containing the predicted, corrected annotation. $\hat{v}_i^a$ is compared to v_i^a to learn optimal weights for the network. During training, the misaligned label y_i^a is obtained by applying random x-y shifts, between ±10 pixels to v_i^a. Sensitivity to the ±10 pixels translation shift is discussed in the results.

When multiple misaligned instances are present in an image, the instances are corrected independently. This approach is chosen for two reasons: it allows instances within an image to have varying degrees of translation error and it also enables the network to be robust to incomplete labels with missing instances. Here, a small dataset of images (x) and carefully verified ground truth labels (v) are used to train the ACN as shown in Stage 1 of Figure 5.2.

5.3.2 Pointer Segmentation Network

Assuming m available annotations – $v_i^1 ... v_i^m$, where $m < A$ – common algorithms will struggle to implement Equation 5.1, as some predicted object instances will not have corresponding true labels for comparison during training. To address this issue, we introduce the PSN, a network that learns to segment an image using only m available annotations. The PSN takes as inputs an image x_i and a single channel of points specifying selected instances to be segmented, and it outputs a segmentation mask only for the selected instances. We specify the fraction of instances to be used for training using a parameter α, where α is the number of selected instances divided by the number

of available instances. Equation 5.2 shows this formulation, where $p_i(\alpha)$ specifies a point within each selected instance, and $\hat{v}_i(\alpha)$ denotes the predicted label for instances specified by $p_i(\alpha)$:

$$f(x_i, p_i(\alpha)) \rightarrow \hat{v}_i(\alpha) \tag{5.2}$$

By including a single channel containing points $p_i(\alpha)$, our PSN segments only instances that are associated with the points. This offers two benefits: first, we simplify the learning task to specify instances of interest, and second, the network can be trained with common binary cross entropy loss. To handle varying extents of missing annotations, the model is trained by randomly picking α for every image in each epoch; at inference time, all instances of interest are specified using points.

In the sequential training configuration, the ACN is used to correct a training dataset that is then inputted to the PSN for object segmentation; this process is shown in Stage 2 of Figure 5.2. Binary cross-entropy loss is used for all networks. Both ACN and PSN use the same baseline architecture (lightUNet) shown in Appendix 5.7.1 and further explained in the results, albeit modified by the number of input channels.

5.4 Data

Three separate datasets are used to train and test the performance of the ACN and the PSN, all described below. During training and testing, we only use images that contain labels.

5.4.1 Aerial Imagery for Roof Segmentation

We use the Aerial Imagery for Roof Segmentation (AIRS) dataset to establish baseline performances for both the ACN and PSN. The AIRS dataset covers most of Christchurch ($457km^2$), New Zealand and consists of orthorectified aerial images (RGB) at a spatial resolution of 7.5 cm with over 220,000 building annotations, split into a training set (T_{set}) and a validation set (V_{set}). The AIRS dataset provides all building footprints within the dataset coverage area. To mimic more

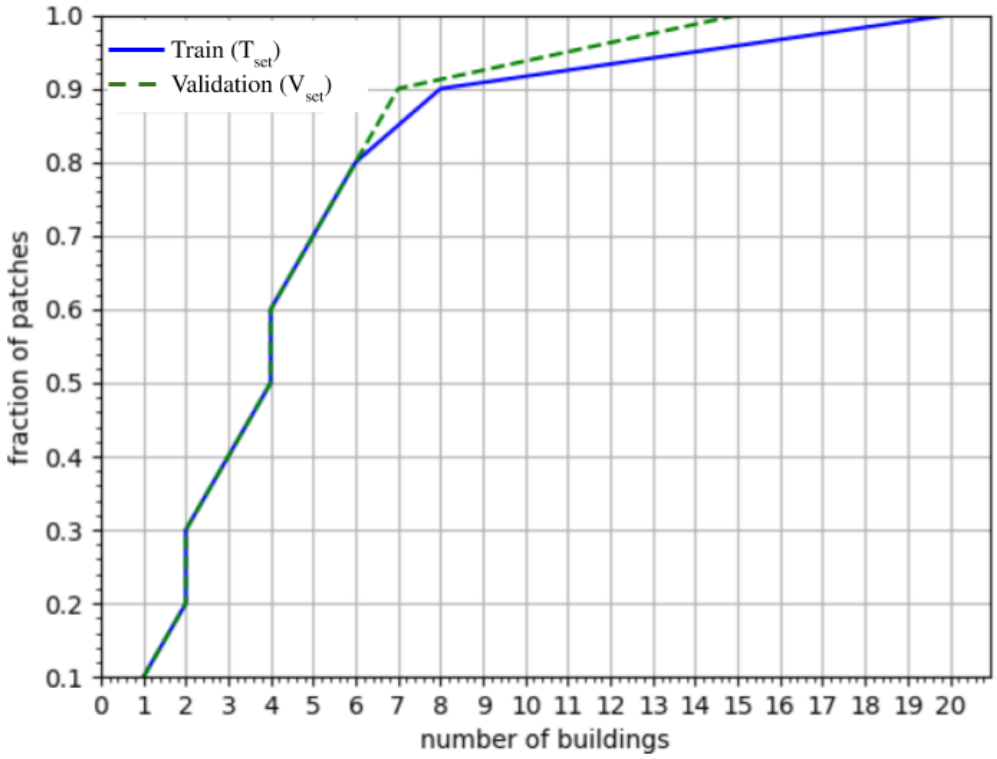


Figure 5.3: CDF of the number of buildings present in 128x128 patches of the 30cm-resampled AIRS dataset.

readily-available data, we resample the imagery to 30 cm, an approach which creates imagery more similar to that provided by Google Earth. Next, we slice the resampled images into 128 by 128 pixel patches and discard all patches in which the area occupied by buildings is less than 10 % of the total area – this methodology ensures that patches with multiple buildings are selected. Other than this basic filtering, we preserve T_{set} and V_{set}.

After resampling and filtering, we obtain 99,501 and 10,108 patches from the T_{set} and V_{set}, respectively. We further split T_{set} into 80:20 fractions, where 80% is used for training and 20% for validation. V_{set} is withheld and used as a test set to evaluate performance. Figure 5.3 shows the fraction of patches for a given number of buildings in T_{set} and V_{set}. Note that some patches contain partial buildings.

5.4.2 OpenStreetMaps

Humanitarian OpenStreetMaps (OSM), through free, community-driven annotation efforts, provides building footprints by country on their Humanitarian Data Exchange (HDX) platform. While this data provides the best (and only) ground truth for many parts of the world, label quality is highly heterogeneous, both in terms of footprint alignment and coverage. In order to test the performance of the ACN on these incomplete and misaligned building footprints, we pair OSM annotations for Kenya [166] with selected DigitalGlobe tiles from Western Kenya (a box enclosed by 0.176 S, 0.263 S, 34.365 E, and 34.453 E) and closer to Nairobi (a box enclosed by 1.230 S, 1.318 S, 36.738 E, and 36.826 E). The DigitalGlobe tiles have a 50 cm spatial resolution and were collected between 2013 and 2016. Slices measuring 128 by 128 pixels were generated from the DigitalGlobe images, which we then couple with overlapping OSM building labels. We generated human verified ground truth annotations for 500 of the image patches.

5.4.3 California Statewide Cropping Map

We also use crop maps and decameter imagery to demonstrate the flexibility of the PSN. The California Department of Water Resources provides a Statewide Cropping Map for 2016 [167]; we pair this shapefile with Sentinel-2 satellite imagery to learn to extract crop extents [168]. Red, blue, green, and near-infrared bands – all at 10m resolution – are acquired from a satellite pass on August 30, 2016; the bands cover the same spatial extent as Sentinel tile 11SKA (a box enclosed by 37.027 N, 36.011 N, 120.371 W, and 119.112W). Cropped polygons larger than $500m^2$ are taken from the California cropping map and are eroded by 5m on all sides to ensure that field boundaries are distinct at a 10m spatial resolution. We split the 110km x 110km tile into images patches measuring 128 by 128 pixels and remove all slices that do not cover any cropped areas, leaving a total of 5,681 patches containing an average of 17 fields per patch; these images are split into training, validation, and test sets at a ratio of 60/20/20.

Table 5.1: mIOU of Base-UNet[76] and lightUNet for routine segmentation with complete and well-aligned labels. Both models are trained on 30 cm resampled AIRS imagery.

Models	**mIOU**
Base-UNet	0.86
lightUNet	0.85

5.5 Results

For all model testing, we report the mean intersection-over-union (mIOU), defined as the intersection of the predicted and true label footprints divided by the union of the same footprints, averaged across the testing dataset.

5.5.1 Baseline Model

We establish the performance of the baseline model (lightUNet) used for both the ACN and PSN by comparing the lightUNet to the UNet architecture proposed by DeepSenseAI [76]. The lightUNet [1] architecture is modified from [76] to perform segmentation with fewer parameters. We refer to the model proposed by [76] as Base-UNet; we train both the Base-UNet and lightUNet models for 30 epochs on the 30 cm resampled AIRS dataset [169], and we report their mIOU. Table 5.1 shows that our lightUNet model achieves comparable performance to the Base-UNet when performing routine building segmentation. Our lightUNet model has about half the number of parameters as the Base-UNet and therefore takes less time to train.

5.5.2 Alignment Correction Network

V_{set} in the AIRS dataset is used to evaluate the performance of the ACN. Random translations were generated between ± 10 pixels for the xy-axis and applied to ground truth AIRS annotations, resulting in unique translation shifts for each object in an image. The introduction of noise through random translation yields a baseline mIOU of 0.55 for comparison. The shifted annotations together with the images are fed into the ACN, and the corrected annotations are compared to the

[1] See Appendix 5.7.1 for details about the convolutions.

true annotations to drive the learning process. We report the mIOU on V_{set} when varying amounts of T_{set} data are used for training. Random translations between ± 10 pixel are applied to all objects in V_{set}. When the ACN is trained with 800, 400 and 240 images, the corresponding mIOU on all images in V_{set} are 0.81, 0.77 and 0.67 respectively, compared to the baseline of 0.55. This suggests that the ACN performs better when more images are used but can learn with only a couple hundred training images.

Table 5.2: mIOU before and after ACN correction.

	mIOU	
Translation Shift (± pixels)	**Before ACN**	**After ACN**
0 to 5	0.63	0.81
5 to 10	0.40	0.73
10 to 15	0.26	0.46
15 to 20	0.18	0.28

Using the ACN model trained with 400 images and random translation shifts between ± 10 pixels, we evaluate the robustness of the ACN to varying levels of translation shifts. Table 5.2 shows mIOU before and after ACN correct, when different ranges of translations shifts are applied to V_{set}. Across all translation shifts the ACN is able to perform some realignment of annotations, even for translations (>10 pixels) which the model was never trained on.

We observe two types of alignment correction as outputs from the ACN: translations and translations plus infilling. Infilling occurs when the misaligned annotation area is less than the building area. In the translation plus infilling case, the model both shifts the annotation and fills the missing portion of the annotation. Overflow is sometimes observed upon correcting the label, resulting in the corrected annotation exceeding the building outline. Figure 5.4 shows examples of both types of corrections when training on 800 images. This figure demonstrates how the ACN learns over time: green outlines show predictions from the ACN and blue outlines show misaligned annotations which the ACN takes as input.

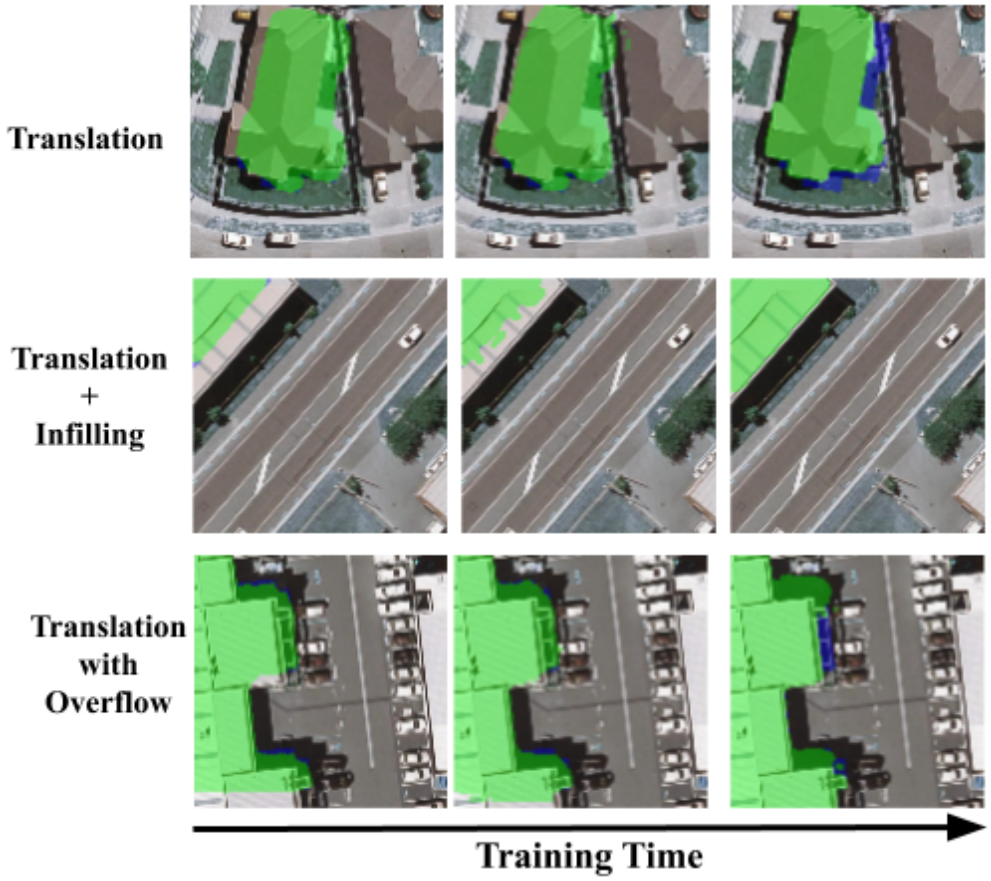


Figure 5.4: Types of annotation corrections performed by the ACN when trained with 800 images. Green shows corrected annotations. Blue shows misaligned annotations.

5.5.3 Pointer Segmentation Network

As an alternative to traditional segmentation models, we propose the Pointer Segmentation Network (PSN), a network that takes in an additional channel with points of interest and returns a single channel output with annotations. The PSN was evaluated separately from the Alignment Correction Network (ACN); this section focuses on reporting segmentation performance on the AIRS dataset when partial – but well-aligned – labels are used. To appropriately compare the PSN with the lightUNet, we evaluate model performance using all annotations in every image of V_{set}. Here, we compare the ability of both networks to segment every building instance in the image, having learned with missing annotations. Table 5.3 reports the performance of the lightUNet and

the PSN with varying fractions of selected annotations (α): As α decreases, performance of the PSN remains robust, indicating that the network still learns the segmentation task despite missing annotations. By specifying the points of interest, the PSN outperforms the lightUNet model.

Table 5.3 also presents results for two different methods of acquiring the required building points: using building centroids versus using a randomly generated point from within the corresponding annotation. By comparing the performance of the PSN using centroids with that of randomly generated points, the best annotation strategy to be used at inference can be determined. We find that the PSN performs better when centroids are used to train the model: This suggests that annotators should strive to extract points near the center of buildings to ensure better segmentation outcomes during inference. Additionally, because the extent of missing annotations may not be known *a priori* for datasets, we evaluate how the network handles heterogeneous (Het.) amounts of label completeness by sampling α from a random uniform distribution between 0 and 1. The uniform distribution ensures an equal chance for alpha to take on any value between 0 and 1. α is resampled for each image during every training epoch. Table 5.3 shows that the PSN remains robust at performing segmentation and works for a heterogeneous α that varies across images. Although α will likely differ across images but remain constant for a given image at a particular time, during training we allow α to change over every training epoch for a given image, enabling our approach to be robust against images taken at different times where new construction may have occurred.

Figure 5.5 shows how the PSN learns – and where non-PSN type networks fail – when learning with missing annotations. The figure shows some outputs of the PSN and the lightUNet model when both are trained with $\alpha = 0.7$ and used to predict all building instances present within the image. Although both networks are trained with missing annotations, generated annotations from the PSN are more visually accurate.

Table 5.3: mIOU of PSN and lightUNet for all buildings in V_{set} images, when trained with varying α.

		mIOU
$\alpha = 1$	PSN (centroid)	0.90
	lightUNet (centroid)	0.85
$\alpha = 0.7$	PSN (centroid)	0.89
	PSN (non-centroid)	0.83
	lightUNet (centroid)	0.53
$\alpha = 0.5$	PSN (centroid)	0.87
	lightUNet (centroid)	0.18
$\alpha =$ Het.	PSN (centroid)	0.87
	lightUNet (centroid)	0.71

5.5.4 Sequential Testing

The AIRS dataset is used to evaluate the sequential performance of our two-stage methodology shown in Stage 2 of Figure 5.2, whereby the ACN and PSN are trained and tested sequentially. Using T_{set}, we establish two training datasets for the sequential process: T1, containing misaligned labels generated from the true T_{set}; and T2, containing ACN-corrected T1 labels. The ACN model trained with 400 training images is used to generate T2. The noise present in both training datasets is captured by the mIOU listed in Table 5.4. The PSN and lightUNet models are trained on T1 and T2 using $\alpha = Het$ with an identical implementation of label withholding to that described in the previous section. The trained models are used to segment V_{set} images; we compare predicted annotations to the true annotations to attain the performance metrics reported in Table 5.4.

Table 5.4 shows that, with $\alpha = Het$, the PSN performs significantly better than the lightUNet when trained on either misaligned labels (T1) or ACN-corrected labels (T2). Again, we find that with incomplete labels, regardless of alignment quality, the PSN outperforms the lightUNet. Moreover, in both training configurations, PSN mIOU performance nears that of the training dataset. As a result, we conclude that the PSN is able to predict object extents at a similar accuracy to that of the training dataset.

Figure 5.6 presents outputs from the PSN when trained with ACN-corrected annotations: cor-

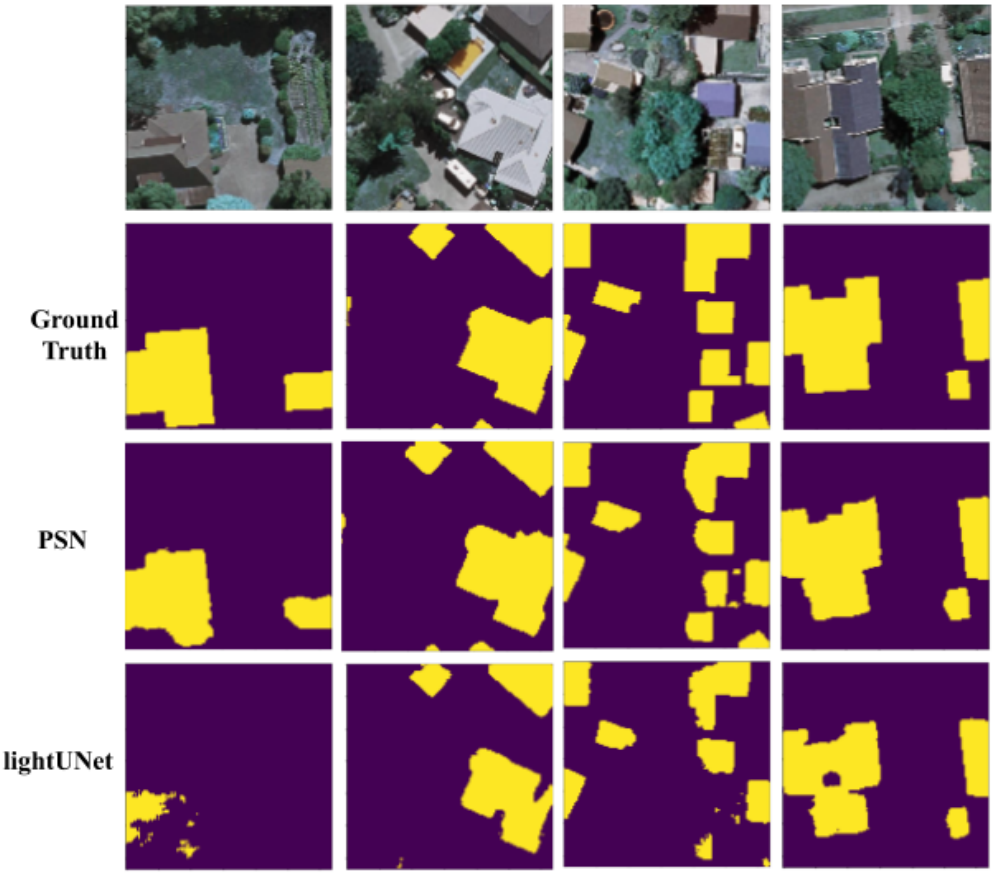


Figure 5.5: Annotations from PSN and lightUNet models when trained with $\alpha = 0.7$. Predictions are made for all building instances in the image and are compared to the ground truth.

rected annotations from the ACN are shown in blue and predicted outputs from the PSN are shown in green. In the left half of Figure 5.6, we present properly corrected ACN-labels and demonstrate that the PSN is able to predict building footprints accurately when corrected annotations are accurate. The right half of the figure shows poorly corrected annotations: These corrected annotations fall on roads, grass, or across the actual building extent. In these cases, the PSN tries to predict a building footprint where there is no building. Accordingly, we conclude that improvements to the ACN can further improve PSN performance, as more accurate training labels will allow for better label prediction. Nonetheless, in the presence of misaligned annotations and partial labels, we are able to achieve better performance with our sequential architecture than with traditional segmentation approaches.

Table 5.4: Performance of the segmentation architectures. The ACN is trained with 400 images; both segmentation networks are trained with $\alpha = Het.$ available annotations.

	mIOU
T1: Misaligned train dataset	0.57
PSN (trained on T1)	0.54
lightUNet (trained on T1)	0.17
T2: ACN-corrected train dataset	0.81
PSN (trained on T2)	0.79
lightUNet (trained on T2)	0.74

5.5.5 ACN Application: Realignment of OSM Annotations

In many parts of the world, ground truth is rare or nonexistent; moreover, what resources do exist often have significant accuracy issues. Despite potential shortcomings, these datasets can provide unique insight into conditions on the ground, and if their quality can be improved, they offer immense value to researchers. To confirm the performance of our realignment method on noisier images and labels, we tested the ACN on OSM building polygons in Kenya, a dataset containing considerable amounts of label misalignment. Of the 500 human-verified ground truth image labels generated for Kenya, 400 are used to train the ACN and 100 to validate. The extent of noise in OSM labels is measured by comparing the labels to the human-verified ground truth labels. mIOUs of 0.30 and 0.31 for the train and validation data respectively were recorded, when comparing OSM labels to their ground truth counterparts. OSM training labels are used to train the ACN and the trained model is ran on the 100 validation labels. A 50 % improvement in mIOU from 0.31 to 0.47 is observed on the 100 validation images. This suggests that our approach is transferable to open source labels and offers gains even with noiser images and labels, using a small dataset.

Figure 5.7 shows a sampling of ACN-corrected OSM annotations for images in the validation dataset: Hand-labelled annotation are shown in blue, OSM annotations are shown in red and corrected annotations are shown in green. Overall, we find that the ACN is able to correct misaligned OSM annotations both in rural and urban regions. In rural Western Kenya, where buildings tend to

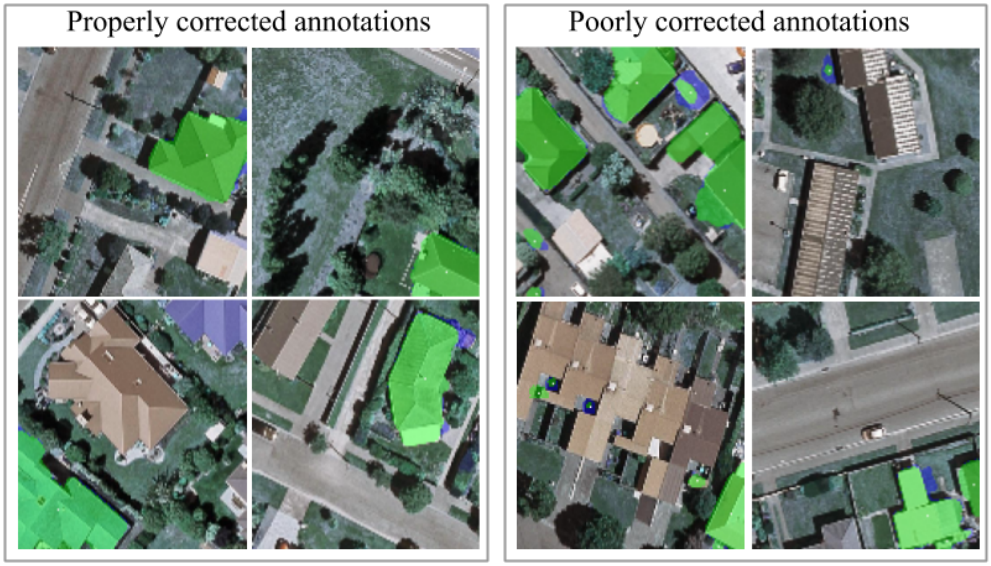


Figure 5.6: Sample images showing PSN performance when trained with corrected annotations. Blue footprints show ACN-corrected annotations. Green footprints show PSN-predicted annotations trained with $\alpha = Het.$ and 400 ACN-corrected labels. PSN performance is dependent on the quality of corrected annotations.

be smaller, the ACN shifts OSM footprints to better align with the buildings. We observe that the noisier image quality makes it more difficult for the ACN to identify extremely small buildings. In more urbanized Nairobi, the ACN also improves the alignment of OSM annotations, albeit with some failure cases.

5.5.6 PSN Application: Cropland Segmentation

Next, we apply the PSN to the task of cropland segmentation using Sentinel-2 imagery and a 2016 California cropping map. Knowing exact field outlines provides valuable information to farmers, planners, and governments; however, a lack of reliable, location-specific ground truth often hampers these efforts. We demonstrate the ability to accurately learn cropland extents using only a subset of fields, instead of requiring the comprehensive set of training polygons that would be necessary for traditional segmentation networks. Similar to previously described tests, we quantify the performance of the PSN in recreating these field boundaries as we select a certain fraction

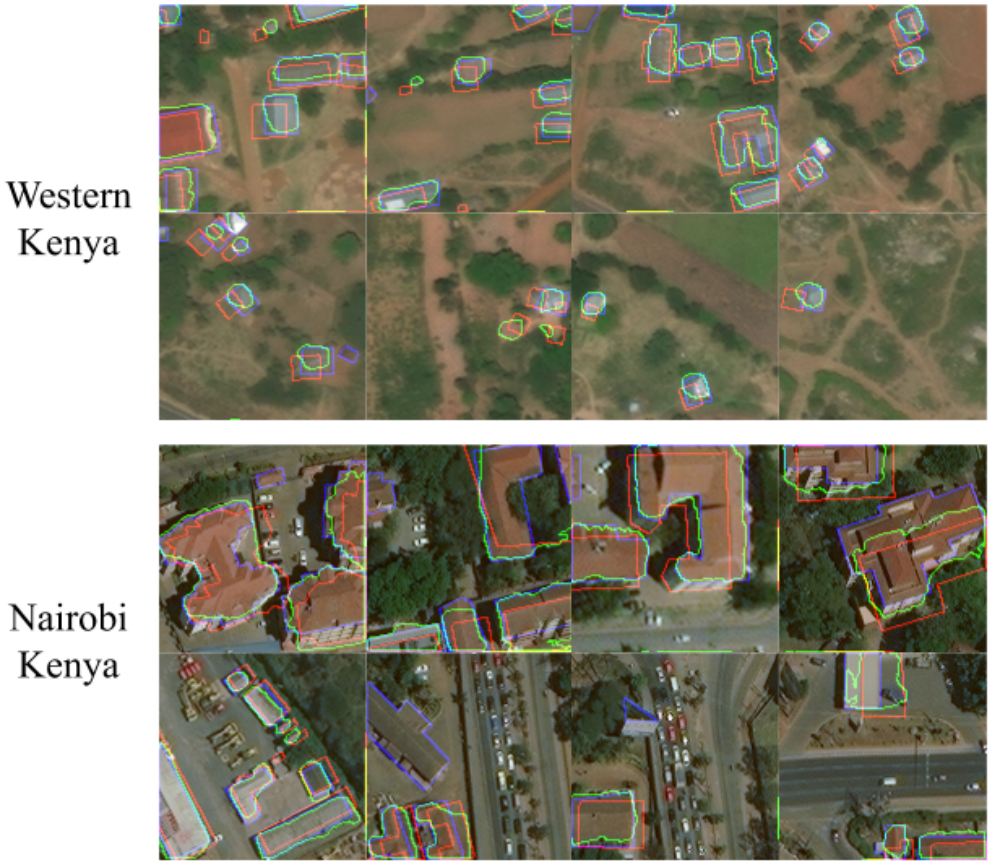


Figure 5.7: Hand-labelled annotations, OSM annotations and ACN-corrected annotations. The ACN is trained on 400 images from Western Kenya and Nairobi, and improves label quality despite the noisier training data.

of the annotations, comparing results to those of the lightUNet. Table 5.5 presents these results.

At all fractions of available training data shown in the table, the PSN outperforms the lightUNet in segmenting croplands. After 40 training epochs, the PSN is able to predict all field boundaries for the test set across both values of α. When trained with all annotations ($\alpha = 1$), the PSN achieves a mIOU of 0.92. In contrast, the lightUNet only reaches a mIOU of 0.75 when $\alpha = 1$, and sees its performance significantly diminish as field boundaries are withheld. Figure 5.8 shows the PSN- and lightUNet - recreated field polygons when the models are trained with $\alpha = 0.75$ and are asked to predict all polygons within an image. The true cropland polygons are shown in blue while the predicted polygons are shown in green; all examples shown come from the test set.

These results demonstrate the viability of the PSN in delineating field boundaries and the preferability of our method over a baseline alternative, when the acquisition of field boundaries is expensive. In locations with low data availability and smaller, non-uniform field boundaries, the PSN provides a reliable method for determining cropped area polygons.

5.6 Summary

As the demand for extracting information from satellite imagery increases, the value of reliable, transferable object segmentation methodologies – especially ones that compensate for noise and inaccuracies in training data – increases in parallel. In this chapter, we present a novel and generalizable two-stage segmentation approach that address common issues in applying deep learning approaches to remotely-sensed imagery. First, we present the Alignment Correction Network (ACN), a model which learns to correct misaligned object annotations. We test the ACN on a set of alignment errors, including i) misalignment of the AIRS dataset, ii) existing and substantial misalignment errors within the OSM Kenyan building footprint dataset. Overall, we find that the ACN significantly improves annotation alignment accuracy.

We also introduce the Pointer Segmentation Network (PSN), a model which reliably predicts an object's extent using only a point from the object's interior. The value of the PSN lies in learning to segment objects within an image despite incomplete or missing annotations, an issue which both hinders traditional segmentation efforts and is common in many ground-truth datasets. We train and test the PSN on the AIRS dataset and find that the model can accurately predict building extent regardless of the fraction of available annotations present or where the training point resides within

Table 5.5: mIOU for all field boundaries in test set, for varying α values.

		mIOU
$\alpha = 1$	PSN	0.92
	lightUNet	0.75
$\alpha = 0.75$	PSN	0.91
	lightUNet	0.69

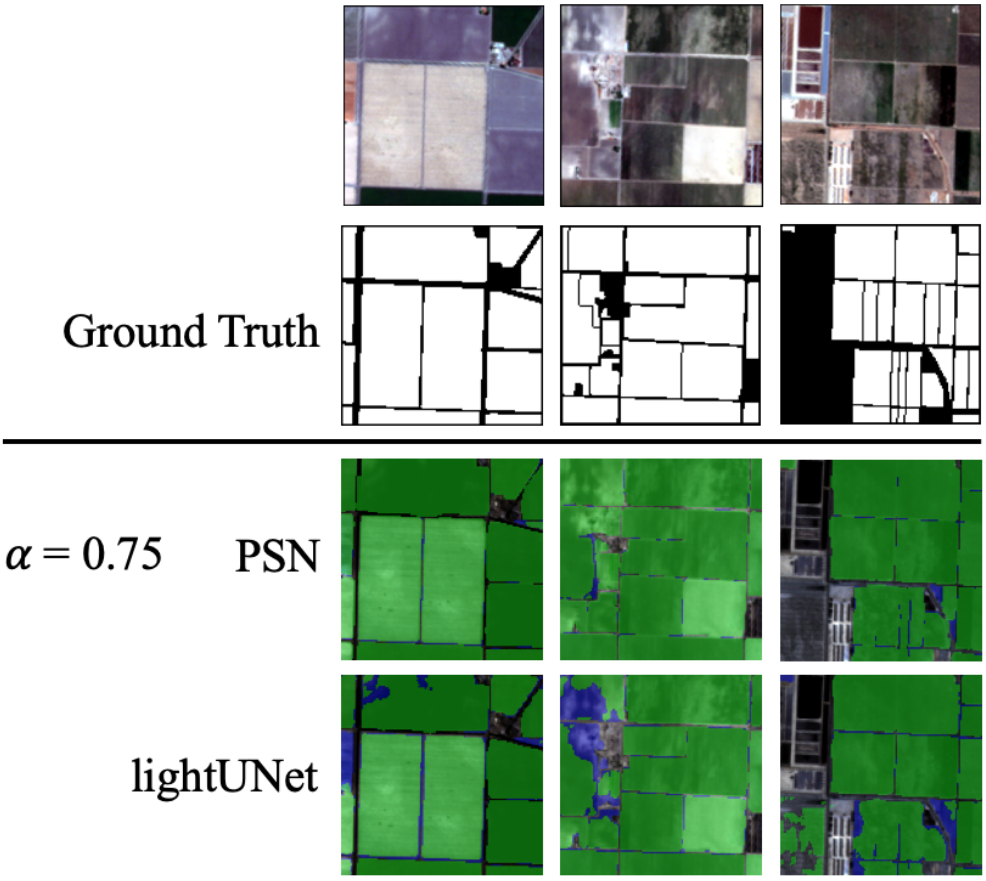


Figure 5.8: Sample images and ground truth labels showing cropland extent in California; also shown in green are PSN and lightUNet predicted footprints $\alpha = 0.75$, overlaid on true cropland polygons, shown in blue. PSN predictions remain highly accurate. Comparatively, the lightUNet predicts only a portion of the crop extents correctly

the object. We also evaluate the performance of the PSN for cropland segmentation using Sentinel imagery and a 2016 California cropland map as inputs, demonstrating that the model can reliably learn cropland polygons regardless of the fraction of available annotations. Overall, for all testing configurations – those which vary the fraction of available training annotations and those which change the location of where the training point lies– and for both object segmentation applications presented – building footprint and cropland extent predictions – the PSN outperforms a baseline segmentation model.

Lastly, we sequentially link the ACN and PSN to demonstrate the ability of the combined networks to accurately segment objects having learnt from misaligned and incomplete training

data. Taken together, we envision our proposed networks providing value to the community of researchers and scientists looking to extract information from widely-available satellite imagery and unreliable ground-truth datasets.

5.7 Appendix

5.7.1 Architecture

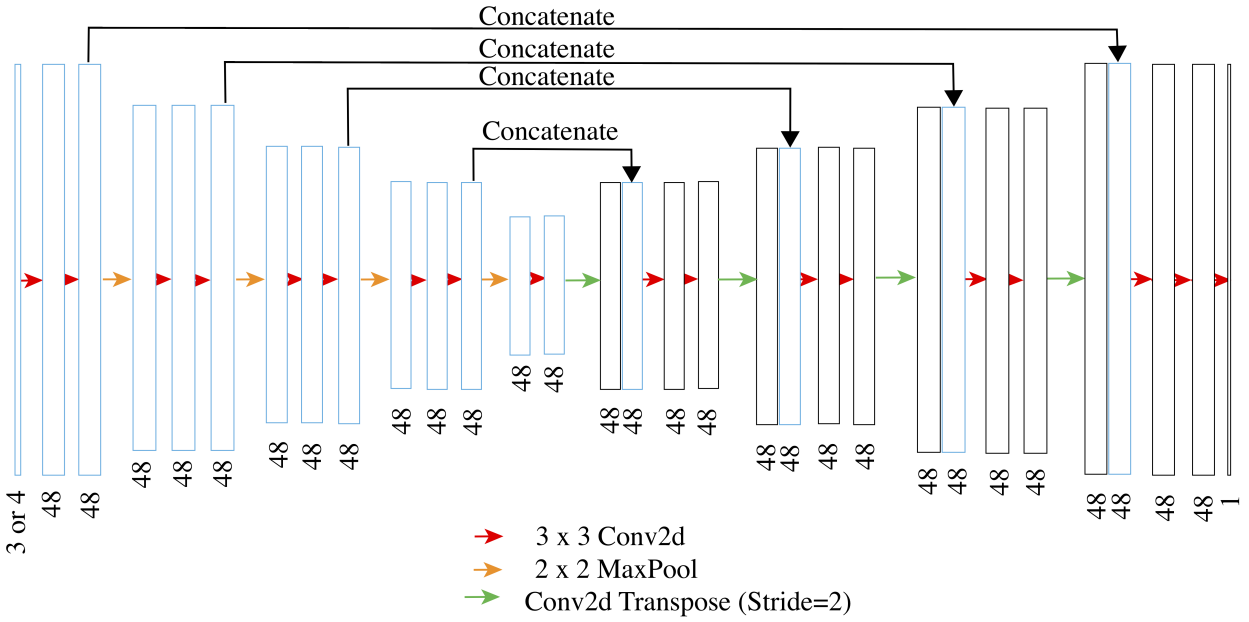


Figure 5.9: Architecture used for both the Alignment Correction Network (ACN) and the Pointer Segmentation Network (PSN). Four input channels are used for both ACN and PSN, while three are used for the lightUNet. This network is modified from [76] by reducing the number of filters to 48 and maintaining the same filter size through out the network. In addition, the network uses dropout in addition to batch normalization after every epoch.

Conclusion

Achieving affordable and reliable universal electricity access and usage requires multiple contributing advancements. This thesis makes key contributions towards i) increased electricity access - by performing electricity demand-side analysis / prediction, and ii) least cost grid access studies given settlement patterns. Beyond electricity access this thesis also contributes towards rapid and large scale assessment/monitoring given investment decisions- by providing methodological approaches to measure welfare indicators and buildings characteristics. Below we outline some novel contributions and key takeaways from this dissertation.

Results from the utility data reveal some relevant patterns about electricity consumption especially for the newly grid-connected homes. As utilities bring in millions of households to the grid, these household tend to use less electricity, peak sooner and plateau at electricity consumption levels lower than their peers who were connected earlier on. This pattern is observed both in Kenya and Rwanda, albeit median monthly consumptions in Kenya ranging in the 30 kWhs while those in Rwanda are much lower, around 5 kWhs. Preliminary evidence from surveys, show that these households are poorer and as a result allocate a smaller proportion of their income to electricity consumption. Concurrently, the capital cost of grid connections, especially in more rural areas remains high reaching up to $1500/connection. At such high capital costs and low consumptions, if energy providers could adequately anticipate consumption, they could better provide more cost-effective systems (e.g. a $200 30W solar panel) to meet such low demand, while simultaneously investing in avenues that could boost income levels of households (e.g. irrigation pumps for improved agriculture). Overall, these insights emphasizes that off-grid technologies remain an important rung in the access ladder.

Deciding the unit price for electricity while meeting the dual objectives of cost recovery and encouraging consumption amongst poorer household, remains a challenge. By analyzing utility revenue and electricity consumption pre and post tariff policies, this work shows that increasing the tariff suppresses consumption especially for lower consumers who also tend to be poorer. Conversely, while the implementation of a "lifeline" tariff might encourage consumption especially in

lower consumers, utility revenues are significantly impacted, further limiting its ability to recover costs in the low consumers. Thus, to adequately support the "lifeline" tariffs and cost recovery, government interventions in the form of subsidies might be required.

Anticipating consumption levels of future customers remains a difficult problem even for utilities. Chapter 2 shows that rapid large scale evaluations of latent electricity demand can be performed using satellite imagery. Results from the work shows that predictions from satellite imagery, provide a better starting point than assuming that newly connected households will consume the same as their older peers. This chapter also shows that improved performances can be obtained by learning about buildings characteristics prior to predicting electricity consumption levels. Approaches that extract household characteristics would better support predictions and human interpretation of the learnt models. In addition to daytime satellite imagery, this chapter also shows that multi-modal models that learn from multiple data sources better capture the variability in household consumption. Further extensions and deployment of the prediction models should perform a thorough examination of other input datasets as they may yet hold the key to improved prediction performance.

Estimating the cost of grid connections is well documented within the literature. However, few approaches demonstrate how cost estimates can be carried out at scale for millions of buildings. Instead, cost metrics based on population density are utilized to determine places that are easier or more difficult to connect via grid extension. In chapter 3, through our abstraction of the grid, we offer a set of unique metrics that allows energy planners to estimate the ease of electrifying every individual structure. This is especially valuable in places of similar population density but different settlement patterns. For example, we show that for two places of similar population density, where one has clustered structures and the other has more spread-out structures, the cost of a grid connection is much lower for the clustered settlement as less wire is needed to grid connect these households. This perspective is completely missed in the absence of our proposed metrics where population density is used as the indicator for cost. We show that this novel indicator of grid connection costs, can be computed across the whole landscape of a country (Kenya in our case)

and is especially relevant in places where the is no pre-existing grid.

Large investments are being made to support electricity bill recovery, one of which is the installation of at home electric meters. As result, the utility is already collecting large amount of electricity usage data, though mainly using it to support energy related endeavours. Chapter 4 shows that electricity usage data can be re-purposed to answer new questions relevant for sustainable development. This application is particularly interesting because it does not require additional large scale investments into data collection efforts. Specifically, this work provides a methodology for estimating household overall consumption expenditure using electricity usage data. We show that electricity usage remains a good proxy for household overall expenditure. This chapter also shows how satellite imagery can be used to estimate the actual average monthly electricity consumption of a building. This methodology can empower governments who are already resource constraint, with the ability to approximate other relevant indicators without deploying large investments to collect new data.

Decision making for planning purposes can occur at multiple resolutions, from the individual household level to the administrative level. High resolution data provides good variability of the indicator but may not preserve the privacy of household. Low resolution data, at the administrative level may support individual privacy but might miss the spatial heterogeneity of the indicator. We show through the work done in chapter 4 that aggregating electricity consumption predictions in Rwanda to the 1 sqkm resolution maximizes both objectives of providing high resolution insights about electricity usage while preserving the privacy of households. In reality, stakeholders tend to be more interested in areas that might be in need for prioritization of a service, rather than the individual households. Our experiments on aggregation provide insights to how quickly performance improves with aggregation and an approach that other stakeholders can apply to evaluate their data or indicator of choice.

Transferable and generalizable electricity prediction models remain a desirable feature especially when multitudes of households remain unconnected. The training and parameter optimizing schemes exemplified in this work show that robust models can be developed to support electricity

consumption studies. In both chapter 2 and 4, the proposed models showed stable and comparable performances from the train to in-sample and out-of-sample test sets. These model tuning strategies are critical to ensuring that the models transfer to unseen regions of the country and perform well for new customers.

Convolutional Neural Networks offer a pathway for extraction of non-linear features or patterns from data. However, they tend to be black-box models with little to no insight about what is driving the predictions. Understanding the tunable knobs within the model enables the planner to first detect spurious results and secondly make informed decisions. Throughout this dissertation, the discovery of explanatory features driving predictions have been prioritized. Building roof sizes and types, road quality, agriculture land have emerged as key variables that correlate with electricity consumption levels. Through the three approaches for model explainability demonstrated in this work, we provide a multi-view lens on the features driving electricity consumption predictions. Such novel explanatory analysis gives confidence that the models are learning relevant features over spurious patterns within the data.

Despite the poor building footprint quality in the regions of this work, a smart modification to traditional building segmentation was developed in order to extract building characteristics. This model can be leverage and coupled with already detected buildings to develop other household indicators such as roof top size and quality. This contribution supports the application of building segmentation to areas where data quality and quantity may be a bottleneck.

Electric utilities are themselves sitting on large volumes of data and as grid-connections continue to be a priority, the volume of this data will grow. This dissertation shows how utilities can analyze their own data to reveal insights about their customers and perform predictions about potential future grid customers. Rather than assuming that newly connected household will consume the same amounts of electricity as older customers, this thesis provides methods to support similar analysis by the data-holders, to inform and support their planning of electricity services.

Numerous contributions have been made throughout this dissertation. While the work could be extended in mutiple ways, here we highlight three critical extensions that may better support

application and deployment of the work:

Predictions from lower resolution satellite imagery: Chapter 4 proposes the use of high resolution 50 cm daytime satellite imagery as an input to the deep learning models to make predictions about average monthly electricity consumption. These images were one-time purchases from Maxar. The provided Maxar imagery from varying satellites (WorldView2, WorldView3, GEOS) is a tapestry of the best images and this collection is selected from different acquisition years. To better support recurrent predictions from remote sensed imagery, this work could be extended to evaluate the predictive power of freely available products such as 10m Sentinel data, that have higher temporal candence. While predictions might not be carried out at the individual building level, predictions at 1sqkm grid can still be supported with the medium resolution product. The freely available imagery will reduce the cost that stakeholders may face when applying this work to new customers. In addition, predictions from medium resolution products might better incorporate the spatial inter-dependence of households' consumption, when multiple households are considered in the same image patch.

Learning from masked data: This work was made possible due to the carefully built relationships with stakeholders such as the utilities in Kenya and Rwanda. On one hand, the utilities hold the data needed to perform this work, while on the other hand sits the energy and machine learning expertise needed to develop the methodologies. By providing masked datasets, utilities can catalyze research and innovation needed to improve their planning and operations. One approach from literature that has shown promising results in protecting data privacy while supporting learning is federated learning methods. Through federated learning approaches, machine learning models can still leverage supervisory signals from relevant training data, while the data remains on the energy planner's server. With such privacy preservation approaches (data masking or federated learning), more robust methods to demand-side analysis and prediction can be developed.

Utility Analytics Toolkits: A key observation from this work is that while utilities have access to large volumes of electricity usage data fewer have access to the necessary human resources and expertise to conduct electricity consumption studies and predictions. This dissertation presents

methodological approaches to studying and predicting electricity usage. For these methods to be deployed and utilized recurrently, there is a need to translate such methods into simple and usable software packages and toolkits that are regularly maintained. Such toolkits can enable utilities to better understand their customers (current and future) and provide the necessary analysis needed to improve their electricity access and reliability plans.

This dissertation has extensively demonstrated multiple contributions in the form of key results and novel methodologies. While the focus of the thesis has been on electricity usage in Kenya and Rwanda, the work can equally be applied to more countries and to other domains seeking to increase access, measure the impact of investments and provide useful insights to planners. We hope that it catalyzes research and deployment beyond the electricity domain.

References

[1] International Energy Agency, *World energy investment 2021*, https://iea.blob.core.windows.net/assets/5e6b3821-bb8f-4df4-a88b-e891cd8251e3/WorldEnergyInvestment2021.pdf, Accessed: 2022-04-04.

[2] Sustainable Energy for All, *Energizing finance: Understanding the landscape 2021*, https://www.seforall.org/publications/energizing-finance-understanding-the-landscape-2021, Accessed: 2022-04-04.

[3] R. M. Desai, H. Kato, H. Kharas, and J. W. McArthur, "From summits to solutions: Innovations in implementing the sustainable development goals," in Brookings Institution Press, 2018.

[4] World Bank, *More people have access to electricity than ever before, but world is falling short of sustainable energy goals*, https://www.worldbank.org/en/news/press-release/2019/05/22/tracking-sdg7-the-energy-progress-report-2019, Accessed 2019-12-22, 2019.

[5] International Energy Agency, *The pandemic continues to slow progress towards universal energy access*, https://www.iea.org/commentaries/the-pandemic-continues-to-slow-progress-towards-universal-energy-access, Accessed: 2022-05-23.

[6] T. Sablik, *Electrifying rural america*, https://www.richmondfed.org/publications/research/econ_focus/2020/q1/economic_history, Accessed: 2022-05-23.

[7] K. Lee, E. Miguel, and C. Wolfram, "Does household electrification supercharge economic development?" *Journal of Economic Perspectives*, vol. 34, no. 1, pp. 122–44, 2020.

[8] K. Lee, E. Miguel, and C. Wolfram, "Experimental evidence on the economics of rural electrification," *Journal of Political Economy*, vol. 128, no. 4, pp. 1523–1565, 2020.

[9] A. Castellano, A. Kendall, M. Nikomarov, and T. Swemmer, "Brighter africa: The growth potential of the sub-saharan electricity sector," *McKinsey Company*,

[10] Energy Information Agency, *How much electricity does an american home use?* https://www.eia.gov/tools/faqs/faq.php?id=97&t=3, Accessed: 2022-05-23.

[11] Energy Information Administration, *During 2021, U.S. retail electricity prices rose at fastest rate since 2008*, https://www.eia.gov/todayinenergy/detail.php

?id=51438#:~:text=In%202021%2C%20the%20average%20nominal,our%20latest%20Electric%20Power%20Monthly., Accessed: 2022-05-23.

[12] C. Wolfram, O. Shelef, and P. Gertler, "How will energy demand develop in the developing world?" *Journal of Economic Perspectives*, 2012.

[13] H. Louie and P. Dauenhauer, "Effects of load estimation error on small-scale off-grid photovoltaic system design, cost and reliability," *Energy for Sustainable Development*, vol. 34, pp. 30–43, 2016.

[14] R. Jones, A. Fuertes, and K. Lomas, "The socio-economic dwelling and appliance related factors affecting electricity consumption in domestic buildings," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 901–917, 2015.

[15] M. J. C. Villareal and J. M. L. Moreira, "Household consumption of electricity in brazil between 1985 and 2013," *Energy Policy*, vol. 96, pp. 251–259, 2016.

[16] J. T. Mensah, G. Marbuah, and A. Amoah, "Energy demand in ghana: A disaggregated analysis," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 924–935, 2016.

[17] P. Esmaeilimoakher, T. Urmee, T. Pryor, and G. Baverstock, "Identifying the determinants of residential electricity consumption for social housing in perth, western australia," *Energy and Buildings*, vol. 133, pp. 403–413, 2016.

[18] S. Amarala, G. Câmaraa, A. M. V. Monteiroa, J. A. Quintanilhab, and C. D. Elvidge, "Estimating population and energy consumption in brazilian amazonia using dmsp nighttime satellite data," *Computers, Environment and Urban Systems*, vol. 29, pp. 179–195, 2005.

[19] Y. Xie and Q. Weng, "Detecting urban-scale dynamics of electricity consumption at chinese cities using time-series dmsp-ols (defense meteorological satellite program-operational linescan system) nighttime light imageries," *Energy*, vol. 100, pp. 177–189, 2016.

[20] C. Elvidge, K. Baugh, E. Kihn, H. Kroehl, E. Davis, and C. Davis, "Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption," *International Journal of Remote Sensing*, vol. 18, pp. 1373–1379, 1997.

[21] P. Chévez, D. Barbero, I. Martini, and C. Discoli, "Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the great la plata region, buenos aires, argentina," *Sustainable Cities and Society*, vol. 32, pp. 115–129, 2017.

[22] Kenya Power, "Kenya power annual report and financial statements for the year ended 30 june 2016.," 2016.

[23] Kenya National Bureau of Statistics, *The 2009 Kenya Population and Housing Census*, http://statistics.knbs.or.ke/nada/index.php/catalog/55, Accessed 2021-01-21, 2021.

[24] AfriPop. "Kenya 100m population." http://www.worldpop.org.uk/data/summary/?contselect=Africa&countselect=Kenya&typeselect=Population, Accessed: 2022-04-04. (2010).

[25] Energy Resource Commision. "Gazette notice no.280: The advocates act, the complaints commission, 92nd quarterly report." http://www.erc.go.ke/images/docs/GazettedRetailElectricityTariffs.pdf. ().

[26] E. Christenson, R. Bain, J. Wright, S. Aondoakaa, R. Hossain, and J. Bartram, "Examining the influence of urban definition when assessing relative safety of drinking-water in nigeria," *Science of the Total Environment*, no. 490, pp. 301–312, 2014.

[27] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 577–584, 2001.

[28] World Bank Global Electrification Database. "Access to electricity (% of population) - rwanda." https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?locations=RW,Accessed: 2022-04-24. ().

[29] Rwanda Energy Group, *Reg annual report (2019-2020)*, https://www.reg.rw/fileadmin/user_upload/FINAL_REG_ANNUAL_REPORT_19-_2020.pdf, Accessed: 2021-08-14, 2020.

[30] Rwanda Energy Group, *Electricity access*, https://www.reg.rw/what-we-do/access/, Accessed: 2021-08-17, 2021.

[31] Rwanda Energy Group, *History of REG*, https://www.reg.rw/about-us/history/, Accessed: 2021-08-14.

[32] Rwanda Utilities Regulatory Authority, *RURA Strategic Plan (2013-2018)*, https://www.rura.rw/fileadmin/docs/RURA_Strategic_Plan_2013-2018.pdf, Accessed: 2021-08-14, 2020.

[33] The NewTimes, *REG chief explains increase in electricity tariff*, https://www.newtimes.co.rw/section/read/191529, Accessed: 2021-08-14, 2015.

[34] World Bank, "Second rwanda energy sector development policy financing," The World Bank, Washington, DC, Tech. Rep. PGD26, Oct. 2018.

[35] Rwanda Energy Group, *Rwanda Energy Group Strategic Plan(2019-2024)*, https://www.reg.rw/fileadmin/user_upload/REG_Strategic_plan.pdf, Accessed: 2022-04-24, 2020.

[36] S. Fobi, V. Deshpande, S. Ondiek, V. Modi, and J. Taneja, "A longitudinal study of electricity consumption growth in kenya," *Energy Policy*, vol. 123, pp. 569–578, 2018.

[37] B. Muhwezi, N. J. Williams, and J. Taneja, "Ingredients for growth: Examining electricity consumption and complementary infrastructure for small and medium enterprises in kenya," *Development Engineering*, vol. 6, p. 100 072, 2021.

[38] World Bank, *Gdp per capita*, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=RW-KE/, Accessed: 2022-04-11.

[39] M. Kojima, X. Zhou, J. Han, J. F. De Wit, R. Bacon, and C. P. Trimble, "Who uses electricity in sub-saharan africa? findings from household surveys," *Findings from Household Surveys (August 9, 2016). World Bank Policy Research Working Paper*, no. 7789, 2016.

[40] S. Fobi, V. Deshpande, S. Ondiek, V. Modi, and J. Taneja, "A longitudinal study of electricity consumption growth in kenya," *Energy Policy*, vol. 123, pp. 569–578, 2018.

[41] P. Alstone, D. Gershenson, and D. M. Kammen, "Decentralized energy systems for clean electricity access," *Nature climate change*, vol. 5, no. 4, pp. 305–314, 2015.

[42] National Institute of Statistics Rwanda, *Integrated household living conditions survey 5 (eicv 5)*, https://www.statistics.gov.rw/datasource/integrated-household-living-conditions-survey-5-eicv-5, Accessed: 2022-05-24.

[43] A. Zvoleff, A. S. Kocaman, W. T. Huh, and V. Modi, "The impact of geography on energy infrastructure costs," *Energy Policy*, vol. 37, no. 10, pp. 4066–4078, 2009.

[44] L. Parshall, D. Pillai, S. Mohan, A. Sanoh, and V. Modi, "National electricity planning in settings with low pre-existing grid coverage: Development of a spatial model and case study of kenya," *Energy Policy*, vol. 37, no. 6, pp. 2395–2410, 2009.

[45] S. Fobi, A. S. Kocaman, J. Taneja, and V. Modi, "A scalable framework to measure the impact of spatial heterogeneity on electrification," *Energy for Sustainable Development*, vol. 60, pp. 67–81, 2021.

[46] Division of Energy Systems - KTH Royal Institute of Technology, *OnSSET - Open Source Spatial Electrification Tool*,
, Accessed 2021-06-04, 2021.

[47] A. Streltsov, J. M. Malof, B. Huang, and K. Bradbury, "Estimating residential building energy consumption using overhead imagery," *Applied Energy*, vol. 280, p. 116 018, 2020.

[48] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 1819 –1835, 2009.

[49] S. C. Bhattacharyya and G. R. Timilsina, "Modelling energy demand of developing countries: Are the specific features adequately captured?" *Energy Policy*, vol. 38, no. 4, pp. 1979 –1990, 2010, Energy Security - Concepts and Indicators with regular papers.

[50] K. Gajowniczek and T. Ząbkowski, "Electricity forecasting on the individual household level enhanced based on activity patterns," *PloS one*, vol. 12, no. 4, e0174098, 2017.

[51] A. Ushakova and S. J. Mikhaylov, "Big data to the rescue? challenges in analysing granular household electricity consumption in the united kingdom," *Energy Research & Social Science*, vol. 64, p. 101 428, 2020.

[52] M. Sajjad *et al.*, "A novel cnn-gru-based hybrid approach for short-term residential load forecasting," *IEEE Access*, vol. 8, pp. 143 759–143 768, 2020.

[53] Y. Hong, Y. Zhou, Q. Li, W. Xu, and X. Zheng, "A deep learning method for short-term residential load forecasting in smart grid," *IEEE Access*, vol. 8, pp. 55 785–55 797, 2020.

[54] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017.

[55] U. H. Syed, *Estimation of un-electrified households & electricity demand for planning electrification of un-electrified areas: Using south africa as case*, 2013.

[56] M. Shen, H. Sun, and Y. Lu, "Household electricity consumption prediction under multiple behavioural intervention strategies using support vector regression," *Energy Procedia*, vol. 142, pp. 2734–2739, 2017.

[57] A. H. Pandyaswargo *et al.*, "Estimating the energy demand and growth in off-grid villages: Case studies from myanmar, indonesia, and laos," *Energies*, vol. 13, no. 20, p. 5313, 2020.

[58] K. Olaniyan, B. C. McLellan, S. Ogata, and T. Tezuka, "Estimating residential electricity consumption in nigeria to support energy transitions," *Sustainability*, vol. 10, no. 5, p. 1440, 2018.

[59] A. Allee, N. J. Williams, A. Davis, and P. Jaramillo, "Predicting initial electricity demand in off-grid tanzanian communities using customer survey data and machine learning models," *Energy for Sustainable Development*, vol. 62, pp. 56–66, 2021.

[60] Y. Shen *et al.*, "Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[61] G. Cadamuro, A. Muhebwa, and J. Taneja, "Street smarts: Measuring intercity road quality using deep learning on satellite imagery," in *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2019.

[62] X. Hou, B. Wang, W. Hu, L. Yin, and H. Wu, "Solarnet: A deep learning framework to map solar power plants in china from satellite imagery," *arXiv preprint arXiv:1912.03685*, 2019.

[63] I. Demir *et al.*, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[64] K. R. Varshney *et al.*, "Targeting villages for rural development using satellite image analysis," *Big Data*, vol. 3, no. 1, pp. 41–53, 2015.

[65] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[66] A. Head, M. Manguin, N. Tran, and J. E. Blumenstock, "Can human development be measured with satellite imagery?" In *Ictd*, 2017, pp. 8–1.

[67] C. Yeh *et al.*, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[68] P. S. Das, H. Chhabra, and S. K. Dubey, "Socio economic analysis of india with high resolution satellite imagery to predict poverty,"

[69] I. Tingzon *et al.*, "MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION,"

[70] C. Ledesma, O. L. Garonita, L. J. Flores, I. Tingzon, and D. Dalisay, "Interpretable poverty mapping using social media data, satellite images, and geospatial information," *arXiv preprint arXiv:2011.13563*, 2020.

[71] K. Ayush, B. Uzkent, K. T. M. B. D. Lobell, and S. Ermon, "Efficient poverty mapping from high resolution remote sensing images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 12–20.

[72] S. Han *et al.*, "Learning to score economic development from satellite imagery," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2970–2979.

[73] G. Falchetta, S. Pachauri, S. Parkinson, and E. Byers, "A high-resolution gridded dataset to assess electrification in sub-saharan africa," *Scientific data*, vol. 6, no. 1, pp. 1–9, 2019.

[74] Microsoft, *Uganda-Tanzania-Building-Footprints*, https://github.com/microsoft/Uganda-Tanzania-Building-Footprints, Accessed 2020-12-27, 2020.

[75] Google, *Mapping Africa's Buildings with Satellite Imagery*, https://ai.googleblog.com/2021/07/mapping-africas-buildings-with.html?m=1, Accessed 2021-08-11.

[76] deepsense.ai, *Deep learning for satellite imagery via image segmentation*, https://deepsense.ai/deep-learning-for-satellite-imagery-via-image-segmentation/, 2020.

[77] S. Fobi, T. Conlon, J. Taneja, and V. Modi, "Learning to segment from misaligned and partial labels," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 286–290.

[78] Earth Observation Group Payne Institute for Public Policy, *Version 1 VIIRS Day/Night Band Nighttime Lights*, https://eogdata.mines.edu/download_dnb_composites.html, Accessed 2020-07-01.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[81] P. Khosla *et al.*, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[83] P. Samangouei, A. Saeedi, L. Nakagawa, and N. Silberman, "Explaingan: Model explanation via decision boundary crossing transformations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[84] S. Fobi and T. Conlon, *Explaining residential electricity consumption with satellite imagery*, https://github.com/tconlon/tconlon.github.io/blob/maste

r/files/Explaining_Electricity_Consumption_Final.pdf, Accessed: 2022-04-24.

[85] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[86] Energy Sector Management Assistance Program, World Bank, *Multi-Tier Framework for Measuring Energy Access - Kenya Household Survey (2018)*, https://datacatalog.worldbank.org/dataset/multi-tier-energy-access-tracking-framework-global-survey-2016-2018, Accessed 2020-08-15, 2018.

[87] UNDP, *Goal 7: Affordable and Clean Energy*, https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-7-affordable-and-clean-energy.html, Accessed 2019-12-13, 2019.

[88] J.-P. Carvallo, J. Taneja, D. Callaway, and D. M. Kammen, "Distributed Resources Shift Paradigms on Power System Design, Planning, and Operation: An Application of the GAP Model," *Proceedings of the IEEE*, vol. 107, no. 9, pp. 1906–1922, 2019.

[89] P. Ciller and S. Lumbreras, "Electricity for all: The contribution of large-scale planning tools to the energy-access problem," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109 624, 2020.

[90] K. B. Debnath and M. Mourshed, "Challenges and gaps for energy planning models in the developing-world context," *Nature Energy*, vol. 3, no. 3, pp. 172–184, 2018.

[91] M. Zeyringer, S. Pachauri, E. Schmid, J. Schmidt, E. Worrell, and U. B. Morawetz, "Analyzing grid extension and stand-alone photovoltaic systems for the cost-effective electrification of Kenya," *Energy for Sustainable Development*, vol. 25, pp. 75–86, 2015.

[92] M. Moner-Girona, K. Bódis, T. Huld, I. Kougias, and S. Szabó, "Universal access to electricity in Burkina Faso: scaling-up renewable energy technologies," *Environmental Research Letters*, vol. 11, no. 8, p. 084 010, 2016.

[93] S. Mahapatra and S Dasappa, "Rural electrification: Optimising the choice between decentralised renewable energy sources and grid extension," *Energy for Sustainable Development*, vol. 16, no. 2, pp. 146–154, 2012.

[94] L. Parshall, D. Pillai, S. Mohan, A. Sanoh, and V. Modi, "National electricity planning in settings with low pre-existing grid coverage: Development of a spatial model and case study of Kenya," *Energy Policy*, vol. 37, no. 6, 2395—2410, 2009.

[95] ModiLabs, *Networkplanner*, http://optimus.modilabs.org/, Accessed: 2019-03-04.

[96] A. Sanoh, L. Parshall, O. F. Sarr, S. Kum, and V. Modi, "Local and national electricity planning in Senegal: Scenarios and policies," *Energy for Sustainable Development*, vol. 16, no. 1, pp. 13–25, 2012.

[97] F. Kemausuor, E. Adkins, I. Adu-Poku, A. Brew-Hammond, and V. Modi, "Electrification planning using Network Planner tool: The case of Ghana," *Energy for Sustainable Development*, vol. 19, pp. 92–101, 2014.

[98] U. Akpan, "Technology options for increasing electricity access in areas with low electricity access rate in Nigeria," *Socio-economic Planning Sciences*, vol. 51, pp. 1–12, 2015.

[99] Y. Abdul-Salam and E. Phimister, "The politico-economics of electricity planning in developing countries: A case study of Ghana," *Energy Policy*, vol. 88, pp. 299–309, 2016.

[100] G. Bolukbasi and A. S. Kocaman, "A prize collecting Steiner tree approach to least cost evaluation of grid and off-grid electrification systems," *Energy*, vol. 160, pp. 536–543, 2018.

[101] MIT, *Reference Electrification Model: A tool for Rural Electrification Planning*, `https://tatacenter.mit.edu/portfolio/reference-electrification-model-a-tool-for-rural-electrification-planning/`, Accessed 2020-01-20.

[102] A. Zvoleff, A. S. Kocaman, W. T. Huh, and V. Modi, "The impact of geography on energy infrastructure costs," *Energy Policy*, vol. 37, no. 10, pp. 4066–4078, 2009.

[103] A. S. Kocaman, W. T. Huh, and V. Modi, "Initial layout of power distribution systems for rural electrification: A heuristic algorithm for multilevel network design," *Applied Energy*, vol. 96, pp. 302–315, 2012.

[104] J. E. Adkins *et al.*, "A geospatial framework for electrification planning in developing countries," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, IEEE, 2017, pp. 1–10.

[105] Kenya National Bureau of Statistics, *Kenya Demographic and Health Survey 2014*, `https://dhsprogram.com/pubs/pdf/fr308/fr308.pdf`.

[106] Facebook and CIESIN-ColumbiaUniversity, *High Resolution Settlement Layer (HRSL) 2016*, `https://www.ciesin.columbia.edu/data/hrsl/`.

[107] M. R. Garey and D. S. Johnson, *Computers and Intractability – A guide to NP-completeness.* 1979.

[108] V. Chvatal, "A greedy heuristic for the set-covering problem," *Mathematics of operations research*, vol. 4, no. 3, pp. 233–235, 1979.

[109] N. Megiddo and K. J. Supowit, "On the complexity of some common geometric location problems," *SIAM journal on computing*, vol. 13, no. 1, 182—196, 1984.

[110] R. C. Prim, "Shortest connection networks and some generalizations," *The bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[111] L. R. Esau and K. C. Williams, "On teleprocessing system design, Part II: A method for approximating the optimal network," *IBM Systems Journal*, vol. 5, no. 3, pp. 142–147, 1966.

[112] A. Navarro and H. Rudnick, "Large-scale distribution planning—Part I: Simultaneous network and transformer optimization," *IEEE Transaction On Power Systems*, vol. 24, no. 2, pp. 744–751, 2009.

[113] U. Pape, N. Yoshida, and S. Malgioglio. "Data-driven tools can support decision-making and improve implementation – especially in crises like covid-19." `https://blogs.worldbank.org/opendata/data-driven-tools-can-support-decision-making-and-improve-implementation-especially-crises`,Accessed: 2022-05-18. ().

[114] L. See, S. Fritz, I. Moorthy, O. Danylo, M. van Dijk, and B. Ryan, *Using remote sensing and geospatial information for sustainable development*. Brookings Institution Press Washington, DC, USA, 2018.

[115] World bank Group. "Data for better lives." `https://www.worldbank.org/en/publication/wdr2021`, Accessed: 2022-05-18. ().

[116] E. Aiken, S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock, "Machine learning and phone data can improve targeting of humanitarian aid," *Nature*, vol. 603, no. 7903, pp. 864–870, 2022.

[117] S. O. Rutstein and K. Johnson. "The dhs wealth index." `https://dhsprogram.com/pubs/pdf/cr6/cr6.pdf`,Accessed: 2022-05-18. ().

[118] S. R. Khandker, D. F. Barnes, and H. A. Samad, "Are the energy poor also income poor? evidence from india," *Energy policy*, vol. 47, pp. 1–12, 2012.

[119] M. Barron and M. Torero, "Household electrification and indoor air pollution," *Journal of Environmental Economics and Management*, vol. 86, pp. 81–92, 2017.

[120] S. R. Khandker, H. A. Samad, R. Ali, and D. F. Barnes, "Who benefits most from rural electrification? evidence in india," *The Energy Journal*, vol. 35, no. 2, 2014.

[121] U. Chakravorty, K. Emerick, and M.-L. Ravago, "Lighting up the last mile: The benefits and costs of extending electricity to the rural poor," *Resources for the Future Discussion Paper*, pp. 16–22, 2016.

[122] P. J. Burke, D. I. Stern, and S. B. Bruns, "The impact of electricity on economic development: A macroeconomic perspective," *International Review of Environmental and Resource Economics*, vol. 12, no. 1, pp. 85–127, 2018.

[123] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell,Stefano Ermon, "Combining satellite imagery and machine learning to predict poverty," *Journal of Science*, vol. 353, no. 6301, pp. 790 –794, 2016.

[124] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[125] J. E. Blumenstock, "Estimating economic characteristics with phone data," in *AEA papers and proceedings*, vol. 108, 2018, pp. 72–76.

[126] M. Hernandez, L. Hong, V. Frias-Martinez, A. Whitby, and E. Frias-Martinez, "Estimating poverty using cell phone data: Evidence from guatemala," *World Bank Policy Research Working Paper*, no. 7969, 2017.

[127] J. E. Steele *et al.*, "Mapping poverty using mobile phone and satellite data," *Journal of The Royal Society Interface*, vol. 14, no. 127, p. 20 160 690, 2017.

[128] M. Fatehkia *et al.*, "Mapping socioeconomic indicators using social media advertising data," *EPJ Data Science*, vol. 9, no. 1, p. 22, 2020.

[129] S. Fobi, J. Mugyenyi, N. J. Williams, V. Modi, and J. Taneja, "Predicting levels of household electricity consumption in low-access settings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3902–3911.

[130] M. Rosenfelder, M. Wussow, G. Gust, R. Cremades, and D. Neumann, "Predicting residential electricity consumption using aerial and street view images," *Applied Energy*, vol. 301, p. 117 407, 2021.

[131] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[132] R. M. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell, "Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 75–82, 2019.

[133] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[134] Z. Du, J. Yang, C. Ou, and T. Zhang, "Smallholder crop area mapped with a semantic segmentation deep learning method," *Remote Sensing*, vol. 11, no. 7, p. 888, 2019.

[135] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[136] G. Wu and Z. Guo, "Geoseg: A computer vision package for automatic building segmentation and outline extraction," *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[137] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.

[138] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, "Road extraction from high-resolution remote sensing imagery using deep learning," *Remote Sensing*, vol. 10, no. 9, p. 1461, 2018.

[139] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[140] N. Audebert, B. L. Saux, and S. Lefevre, "Joint learning from earth observation and openstreetmap data to get faster better semantic maps," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 67–75, 2017.

[141] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proceedings of the 29th International conference on machine learning*, 2012.

[142] A. Basiri *et al.*, "Quality assessment of openstreetmap data using trajectory mining.," *Geospatial information science*, vol. 19, pp. 56–68, 2016.

[143] J. E. Vargas-Munoz, S. Lobry, A. X. Falcao, and D. Tuia, "Correcting rural building annotations in openstreetmap using convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 283–293, 2019.

[144] A. Bearman, V. F. Olga Russakovsky, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," *European conference on computer vision*, 2016.

[145] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3431–3440, 2015.

[146] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation.," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[147] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *International Conference on Medical image computing and computer-assisted intervention.*, vol. 39, no. 12, pp. 2482–2495, 2017.

[148] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation.," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[149] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[150] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015.

[151] Z. Li, C. Peng, G. Yu, X. Zhang, and Y. D. andJian Sun, "Light-head r-cnn: In defence of two-stage object detector," in *arXiv preprint arXiv:1711.07264*, 2017.

[152] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[153] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, 2001, pp. 416–423.

[154] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," *European conference on computer vision*, 2014.

[155] Safyan, "Overview of the planet labs constellation of earth imaging satellites.," 2015.

[156] C. N. Doll, J.-P. Muller, and J. G. Morely, "Mapping regional economic activity from night-time light satellite imagery," *Ecological Economics*, 2005.

[157] B. Joshi, H. Baluyan, A. A. Hinai, and W. L. Woon, "Automatic rooftop detection using a two-stage classification," in *Proceedings of the 2014 UKSim-AMSS 16th International*

Conference on Computer Modelling and Simulation, ser. UKSIM '14, USA: IEEE Computer Society, 2014, 286–291, ISBN: 9781479949229.

[158] J.-Q. Liu, Z. Wang, and K. Cheng, "An improved algorithm for semantic segmentation of remote sensing images based on deeplabv3+," in *ICCIP '19*, 2019.

[159] A. Perez, S. Ganguli, S. Ermon, G. Azzari, M. Burke, and D. B. Lobell, "Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty," *CoRR*, vol. abs/1902.11110, 2019. arXiv: 1902.11110.

[160] R. Mahabir, A. Stefanidis, A. Croitoru, A. T. Crooks, and P. Agouris, "Authoritative and volunteered geographical information in a developing country: A comparative case study of road datasets in nairobi, kenya," *ISPRS International Journal of Geo-Information*, vol. 6, no. 1, p. 24, 2017.

[161] L.-A. Siebritz and G. Sithole, "Assessing the quality of openstreetmap data in south africa in reference to national mapping standards," in *Proceedings of the Second AfricaGEO Conference, Cape Town, South Africa*, 2014, pp. 1–3.

[162] S. P. Camboim, J. V. M. Bravo, and C. R. Sluter, "An investigation into the completeness of, and the updates to, openstreetmap data in a heterogeneous area in brazil," *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, pp. 1366–1388, 2015.

[163] A. Wright, *Map completeness estimation and experimental analytics for health*, Retrieved March 6, 2020 from https://www.hotosm.org/updates/experimenting-with-analytics-for-health/, 2020.

[164] B. Swan, M. Laverdiere, and H. L. Yang, "How good is good enough? quantifying the effects of training set quality," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, ser. GeoAI'18, Seattle, WA, USA: Association for Computing Machinery, 2018, 47–51, ISBN: 9781450360364.

[165] N. Girard, G. Charpiat, and Y. Tarabalka, "Noisy supervision for correcting misaligned cadaster maps without perfect ground truth data," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2019, pp. 10 103–10 106.

[166] Humanitarian Data Exchange, *Hotosm kenya buildings (openstreetmap export)*, Retrieved February 27, 2020 from https://data.humdata.org/dataset/hotosm_ken_buildings, 2020.

[167] California Department Of Water Resources, *2016 california statewide agricultural land use map*, Retrieved February 27, 2020 from https://gis.water.ca.gov/app/CADWRLandUseViewer/, 2020.

[168] F. Gascon *et al.*, "Copernicus sentinel-2a calibration and products validation status," *Remote Sensing*, vol. 9, no. 6, 2017.

[169] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings.," *arXiv preprint arXiv:1807.09532*, 2018.