Scalable Spatial Design of Electricity Access Systems

Alycia Leonard

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2022

Acknowledgements

To be perfectly honest, this is the last thing I am writing and easily the hardest to get down. There are too many people to thank. I am so lucky and grateful for all the support I have received throughout this journey. I will not be able to fit everyone here, so do forgive me if you notice any gaps. Please know that my heart is glowing with gratitude regardless of the limited number of words that fit on the page.

First, thanks to Prof. Malcolm McCulloch for your steadfast support as my supervisor. You have taught me how to wander off the map and tell a story, which is certainly not what I thought I would learn when I started a DPhil, but which has proven invaluable. Thank you for your consistent understanding and guidance.

I must also thank Dr. Scot Wheeler for his informal supervision and mentorship. Scot, when my brain was too full or going too fast down any number of rabbitholes, you offered a calm and rational ear which was indispensable in keeping me sane.

I am similarly grateful to Dr. Stephi Hirmer for her mentorship. Stephi, while you were frequently reprimanded for distracting me from writing this thing, time spent working on our innumerable side projects has been so enriching. You embody the power of women to pull one another up and I am grateful for your friendship.

Thanks to the whole Oxford Energy and Power Group for being the most interesting, brilliant, and kind group of people. Anna, Matt, Constance, and Liyang: thank you for guiding me out of the nest. Miriam, Aniq, Flora, and Claire: thank you for your friendship and solidarity when things got hard. I would not have made it across the finish line without you all.

On a personal level, there are so many people who helped me survive and thrive throughout this process. To Rose, Nick, Vicky, Sarah, and far too many other beautiful friends to name – thank you for colouring my world. To the Hertford MCR crew, thank you for holding me up and keeping me smiling. To the Oxford Basketball Club and my W1 ladies, thank you for giving me a team, the chance to be a Blue, and the opportunity to get out of the bubble.

Last but not least, thanks are due to my family for their endless support. Mom and Dad, thank you for making me who I am today. Where I am is entirely attributable to how you raised me. I could not ask for better parents. Danielle, thank you for always being a star I could follow when it was hard to find my own path. And Caitlin, thank you for providing me with a consistent oasis of safety and support.

Abstract

This thesis explores spatially specific data and methods to design community-tailored electricity access systems at scale. It is motivated by the need to close the electricity access gap in rural low- and middle-income country contexts quickly and cheaply in line with the Sustainable Development Goals.

The majority of the 760 million people currently lacking electricity access live in rural areas of sub-Saharan Africa and South Asia. Electrifying these areas is challenging given their cultural diversity, remote nature, and sensitivity to affordability. Context-specific electrical designs are required to achieve uptake in these communities; however, such specificity can come at high cost. This thesis therefore tackles the challenge of the local *specificity* and global *scale* of the electricity access problem through practical spatial design methods. Two key data gaps are identified which must be filled in order to design least-cost energy access systems: locations of potential connection points, and their anticipated demands.

Home-level location data which are publicly available for potential connection points in off-grid areas are either aggregated at inadquate resolution for topology design or contain significant gaps, particularly in off-grid areas. To address this, citizen science and computer vision are applied to accelerate accurate home detection in satellite imagery. Through a large-scale online citizen science project, approximately 1,267 km^2 of rural Kenya, Sierra Leone and Uganda was mapped at an average rate of 7 km^2/day and an estimated cost of \$20.84/km^2. Home annotations produced through this work achieve a recall of 93% and precision of 49%, which can be increased to 69% through clustering. The clustered annotations were used to train a Faster R-CNN object detection algorithm, which achieves a precision of 67% and recall of 36%; this can be increased to 57% by training on raw annotations instead of clusters. The trained detector was found to map at a rate of 42,938 km^2/day, proving the rapid mapping of rural unelectrified areas to be feasible at a global scale and low cost.

High-resolution residential demand data are similarly scarce in off-grid communities. Costly local surveys to understand electrical aspirations tend to produce inaccurate results, given the unfamiliarity of respondents with electricity. To overcome this, a bottom-up demand estimation approach rooted in existing empirical data is developed to achieve spatially-specific and realistic demand estimates for

off-grid areas. A case study application of this approach in Sierra Leone was undertaken using Multiple Indicator Cluster Survey data. The results of this work validate the underlying premise of spatial variance of demand amplitude. The load profiles generated using this approach are found to approximate a Tier 3 load, despite a lack of Tier 3 appliances, leading to a critique of the definition of the Multi-Tier Framework for Measuring Energy Access.

These location and demand data are finally applied to home-level spatial grid design. Home locations are clustered into electricity communities, grid topologies are estimated using graph theory, generation types are selected through spatial analysis, and generation and storage are optimally sized for least cost. An approach is also developed to map design pathways through modular community grid expansions which allow for demand growth and autonomy. This framework was applied in a case study region in the Northern Province of Sierra Leone. In this region, 12 local micro-grids are identified as the best electrification solution in the near-term, with 11 outlying homes receiving solar home systems. Infrastructure sizing is presented for one micro-grid, where an initial design point of 50 kW of PV and 108 kWh of battery storage is found to meet anticipated low-end and mid-term demands without energy poverty risk. Three modular expansions of 30 kW PV and 65 kWh of storage are specified for installation in this micro-grid as demands evolve.

The work in this thesis focuses on sub-Saharan Africa, with particular attention paid to Kenya, Uganda, and Sierra Leone. Case studies in the thesis primarily focus on Sierra Leone; however, the methods are purposefully intended to be practical and generalizable across low- and middle-income countries requiring electricity access system design.

Contents

List of Figures

List of Abbreviations

AC . . . . . Alternating current

ACSR . . . . . Aluminium conductor steel reinforced

AFREP . . . . African Renewable Electricity Profiles database

AI . . . . . Artificial intelligence

ANN . . . . . Artificial neural network

CIESIN . . . . Center for International Earth Science Information Network

CNN . . . . . Convolutional neural network

COCO . . . . . Common objects in context

CSV . . . . . Comma-separated values

DER-CAM . . Distributed Energy Resources Customer Adoption Model

EO . . . . . Earth observation

ESMAP . . . . Energy Sector Management Assistance Program

DBSCAN . . . Density-Based Clustering of Applications with Noise

DC . . . . . Direct current

DHS . . . . . Demographic and Health Surveys

DMC-3 . . . . Disaster Monitoring Constellation 3

DNI . . . . . Direct normal irradiation

FCN . . . . . Fully-convolutional network

GDP . . . . . Gross domestic product

GeoTIFF . . . Geographic tag image file format

GHI . . . . . Global horizontal irradiation

GIS . . . . . Geographic Information System

GPS . . . . . Global positioning system

GPU . . . . . Graphics processing unit

GPW . . . . . Gridded Population of the World

GSD . . . . . Ground sample distance

HDBSCAN* . Hierarchical Density-Based Clustering of Applications with Noise

HICs . . . . . High income countries

HOGA . . . . Hybrid Optimization by Genetic Algorithms

HOMER . . . Hybrid Optimization of Multiple Energy Resources

HRSL . . . . . High Resolution Settlement Layer

IHS . . . . . Integrated Household Survey

IRENA . . . . International Renewable Energy Agency

JSON . . . . . JavaScript object notation

KOMPSAT 3A Korea Multi-Purpose Satellite 3A

LCOE . . . . . Levelized cost of energy

LED . . . . . Light-emitting diode

Le . . . . . Leones

LMICs . . . . Low- and middle-income countries

LSMS . . . . . Living Standards Measurement Survey

MIC . . . . . Multiple Indicator Cluster Survey

MST . . . . . Minimum spanning tree

MTF . . . . . Multi-Tier Framework for Measuring Energy Access

NOAA . . . . . National Oceanic and Atmospheric Administration

NREL . . . . . National Renewable Energy Laboratory

OnSSET . . . Open Source Spatial Electrification Tool

OPTICS . . . Ordering Points to Identify the Clustering Structure

OSeMOSYS . Open Source Energy Modelling System

OSM . . . . . Open Street Map

PNG . . . . . Portable network graphics

PTTP . . . . . Power to the People citizen science project

PV . . . . . Photovoltaic

px . . . . . Pixels

RAMP . . . . Remote-Areas Multi-energy systems load Profiles model

R-CNN . . . . Regions with convolutional neural network features

R-FCN . . . . Region-based fully-convolutional networks

REM . . . . . Reference electrification model

SDG . . . . . United Nations Sustainable Development Goal

SHS . . . . . Solar home system

SPP-net . . . . Spatial pyramid pooling network architecture

UNICEF . . . United Nations Children's Fund

USD . . . . . United States Dollars

VHR . . . . . Very-high resolution

YOLO . . . . . You Only Look Once

Nomenclature

AP	Average Precision.
AR	Average Recall.
c	Number of clusters
CV	Coefficient of variance
$cycle$	Duty cycle
d_{MST}	Conductor distance in the minimum spanning tree grid estimate
d_{rad}	Conductor distance in the radial grid estimate
d_{road}	Conductor distance in the road-following grid estimate
d_{under}	Under-grid distance defined in national energy policy
E_{avg}	Average daily energy consumption
E_{day}	Average daily energy consumption while the sun is up
E_{max}	Maximum daily energy consumption
E_{med}	Median daily energy consumption
E_{min}	Minimum daily energy consumption
E_{night}	Average daily energy consumption before sunrise and after sunset
F_1	F1 score (i.e. the harmonic mean of precision and recall)
F_N	False negative
F_P	False positive
$fixed$	Constrains whether switch-on of all instances of one appliance in one household are always simultaneous.
$frac_i$	Proportion of households belonging to usertype i
$frames$	Time frames in which an appliance switch-on event can occur
$frame_{peak}$	Time frame in which peak household appliance usage is expected
h	Height of annotation bounding box exported from Zooniverse
i	Usertypes

IoU	Intersection over union
j	Number of households in a usertype
K	Number of neighbours considered in nearest neighbours
$k-dist$	Set of distances to K nearest neighbours
m	Quantity of occurences of one appliance in one household
m_{clSize}	Minimum cluster size in HDBSCAN*
$minPts$	Minimum number of points within distance ε to designate a point as a core point in DBSCAN
n	Quantity of households in a community
n_{combs}	Number of possible appliance combinations
P	Precision (Chapters 3 and 4), Power (Chapters 5 and 6)
R	Recall
T_P	True positive
t_{peak}	Peak time of appliance usage in a household
t_{min}	Minimum time appliance is kept on after switch-on event
t_{tot}	Total time of use for one appliance in one day
w	Width of annotation bounding box exported from Zooniverse
x_c	Center x coordinate of bounding box annotation
x_n	X corner coordinate of bounding box annotation ($n = 1:4$)
x_{n_u}	X corner coordinate of bounding box annotation prior to rotation ($n = 1:4$)
y_c	Center y coordinate of bounding box annotation
y_n	Y corner coordinate of bounding box annotation ($n = 1:4$)
y_{n_u}	Y corner coordinate of bounding box annotation prior to rotation ($n = 1:4$)
δ	Percentage of random variability attributed to appliance attributes in stochastic optimisation
ϵ	Local radius to identify core points and grow DBSCAN clusters
σ	Standard deviation
θ	Angle of clockwise rotation of annotations exported from Zooniverse

1

Introduction

Contents

This thesis addresses the question:

Can we design affordable electricity access systems suited to local spatial context and needs at the scale required to close the global electricity access gap?

Driven by the need to close the electricity access gap in rural developing contexts quickly and cheaply, this thesis explores spatially specific data and methods to design community-tailored electrical systems. The entire design process is examined, ranging from the sourcing of key input data to the specification of appropriate rural electrification technologies and grid topologies. This thesis seeks to contribute to knowledge by investigating spatial design approaches which can be applied with local *specificity* at a global *scale*.

1.1 Context and Motivation

Approximately 760-770 million people worldwide lack access to electricity [1–3]. The vast majority (87%) live in rural areas, mostly in South Asia and sub-Saharan Africa [4]. This thesis focuses on the sub-Saharan African context, where 53% of the population remain without access [1].

Achieving universal electrification is prioritised globally as evidenced by the seventh United Nations Sustainable Development Goal (SDG), which targets universal electrical access by 2030 [5]. This is unsurprising given the critical role that access to sustainable electrical energy plays in enabling welfare and prosperity in the modern world.

The benefits of electricity access span most aspects of daily life. Electric cooking, lighting, and appliances can bring health, quality of life, and gender equality benefits. Electric cooking circumvents the harmful particulate exposure inherent to cooking with biomass or coal, reducing associated respiratory difficulties and neurological symptoms [6] amongst women and young children [7]. Lighting the home through cheap and efficient electric light-emitting diodes (LEDs) provides brighter illumination than kerosene lamps while reducing risks of structural fire and particulate exposure [8]. Appliances like washing machines and electric mixers can accelerate laborious chores and domestic work, granting women and children more time to study and socialise [9]. Electricity can also provide communal health and safety benefits and financial opportunities. Solar-powered water pumps can provide clean drinking water and irrigation to facilitate agriculture [10]. Refrigeration preserves perishable foods [4] while providing opportunities to boost income by selling cool drinks or produce. Electronic healthcare technologies [11] have concretely life-saving impacts, while night-time electric street lighting reduces gendered and sexual violence risks, allowing women to participate more safely in community activities outside the home in the evenings [12]. Electronic mobile money systems like M-PESA [13] give unbanked people financial opportunity. The diverse positive impacts of electricity access are widely thought to culminate in long-term poverty alleviation [4].

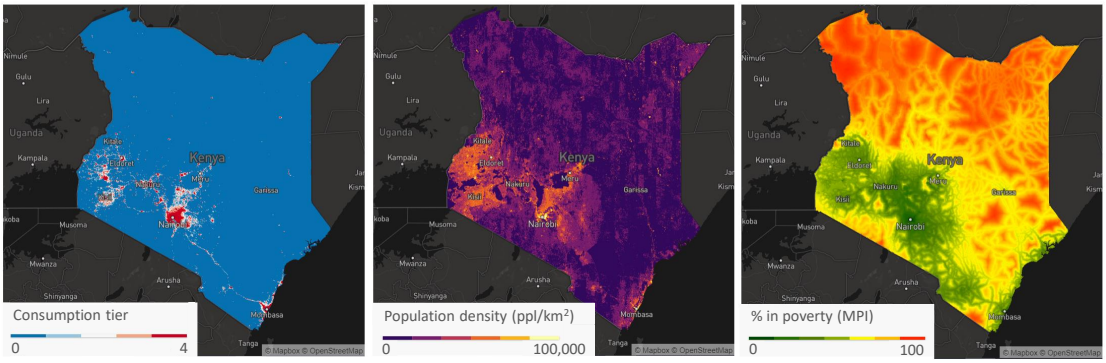


Figure 1.1: Energy consumption[1][14], population density [15], and multidimensional poverty index scores [16] in Kenya. There is strong correlation between lower energy consumption, lower density rural areas, and higher poverty.

Given the abundant development and welfare benefits of electricity access, one might wonder why about 10% of the global population remains unelectrified. Perhaps unsurprisingly, closing this access gap is deeply challenging. Most low- and middle-income countries (LMICs) containing the majority of the predominantly rural electricity gap are simultaneously trying to electrify, develop, and avoid or reduce carbon emissions under resource constraints. This is a trifecta of challenges that high-income countries (HICs) never had to face concurrently, and which are only increasing in urgency as the global population rapidly grows [17] and the climate crisis worsens [18]. LMICs are already struggling to keep up with climate change impacts [19], facing $18 billion in damage to power generation and transport infrastructure each year caused by natural disasters of accelerating frequency [20]. While it can be argued that the HICs which have caused the majority of climate change ought to compensate LMICs for such damage [21], formal mechanisms for climate reparations are scarce, so LMICs are left to foot these bills while also investing in their own green electrification infrastructure. As such, rural electrification in LMICs can become deprioritized, as rural off-grid populations have low political leverage [22] and can be expensive to electrify; it is difficult for already cash-strapped institutions to justify these costs. This compounds the vicious cycle of low incomes and poor access [23]; the correlation of these is illustrated in Figure 1.1.

[1]Note that tier 4 here represents a combination of 4 and 5 from the Multi-Tier Framework for Measuring Energy Access (MTF).

Beyond the economic challenges of closing the electricity gap, there are also technical design challenges to overcome. Poor off-grid households in LMICs are likely to have heightened sensitivity to energy affordability [24]. Electricity access system designs must therefore prioritize affordability to generate uptake; this requires a high degree of bespoke design. It seems that it would be much faster and more cost effective to electrify such areas using a uniform "plug-and-play" system design; however, such standardized designs are less likely to generate uptake, particularly given the high ethnic and cultural diversity in LMICs which can impact energy usage norms. To illustrate, while sub-Saharan Africa accounts for a quarter of all countries in the world, it accounts for 43% of its ethnic groups, with an average of eight groups per country compared to averages of three to five in other regions[2] [25]. It cannot be assumed that energy-related behaviour is the same across each country; each cultural group may have different demands, expectations, and economies of energy services impacting optimal and feasible design.

Under such resource-constrained conditions requiring a high degree of bespoke community design, traditional engineering practices for greenfield electrical planning can be difficult to scale. This thesis therefore explores alternative design methods which can negotiate the trade-off between the local specificity required in design and the global scale of the electricity access gap through data-driven spatial approaches.

1.1.1 Designing Electricity Access Systems

To begin the exploration of alternative design approaches which marry local specificity and global scale, it is useful to understand the key steps of electricity access system design. These include selecting an appropriate grid type and topology, and subsequently specifying and sizing electrical equipment. As will be discussed, these design choices are heavily dependent on the spatial characteristics of off-grid areas.

[2]This research was completed in 2003 – an updated survey would add greatly to the literature.

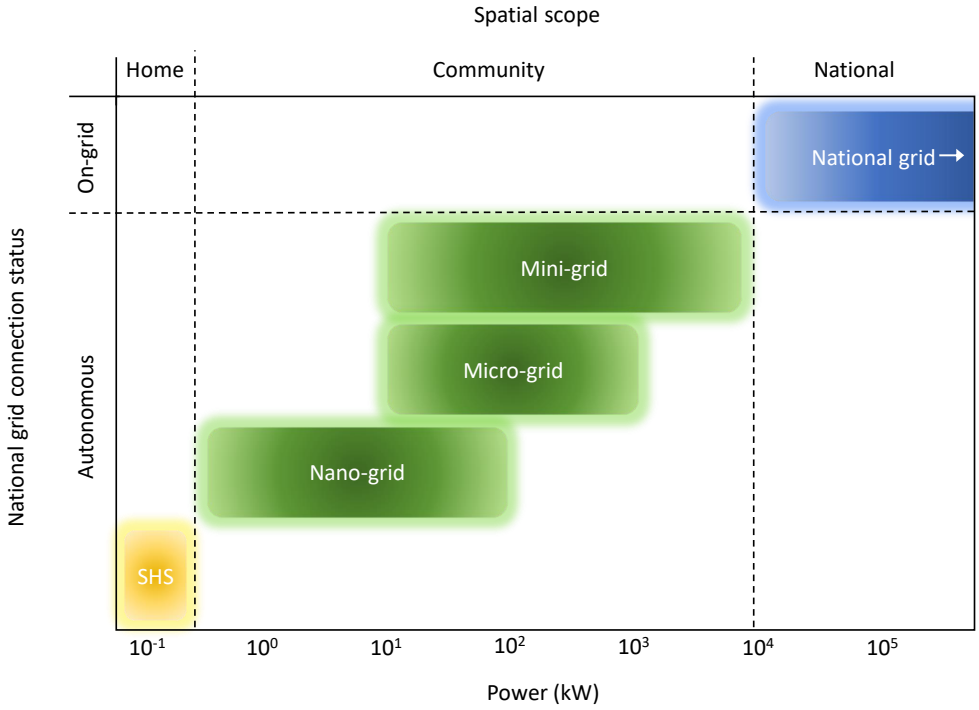


Figure 1.2: Grid technologies of different sizes for rural electrification, classified by power, spatial scope, and autonomous or on-grid status. Gradients and blurring indicate how the definitions of these technologies and the boundaries between them are not sharp.

Selecting a Grid Type or Technology

Contrary to historical precedent in HICs, it is not reasonable to assume that national grids can be extended to electrify most rural off-grid areas in LMICs. There is increasing evidence that smaller-scale electrical technologies like mini-grids may be the most cost effective solutions to connect remaining off-grid populations [26]. The palette of grid types can be generally grouped into interconnected and autonomous categories, and are summarised in Figure 1.2.

Grid extension is still often considered to be the default electrification choice. It can provide high-reliability, high-quality electricity without any practical upper capacity limit for domestic usage [27]. The cost effectiveness of grid extension depends on the proximity of those to be electrified to grid connection points such as transformers and substations. Extending the grid to reach rural and remote areas involves constructing long transmission lines with high capital costs. This

raises the cost of energy, often manifesting as prohibitively high connection charges for already poor customers [28] and low access rates despite the presence of grid infrastructure. For instance, in a study in the Western Kenyan counties of Busia and Siaya, only 5.5% of rural "under-grid" households (i.e. those within 600 m of a transformer) were found to have an electrical connection; however, 55% of these households indicated that they would connect if the connection charge were just 30% lower [29]. In some cases, remote industrial "anchor customers" such as mines or mobile service towers provide affordable or free electrical connections to surrounding communities [30], but this is the exception, not the norm.

Autonomous local community grids (i.e. mini-, micro-, and nano-grids) can be a more cost-effective option for clustered off-grid communities distant from existing grid infrastructure. These grids include generation and optionally storage to independently meet grid community energy needs. Mini-grids serve local demand in the range of 10 kW to 10 MW [31] at medium and low voltages, and are often large enough to provide grid-quality alternating current (AC) electricity [26]. Micro-grids are lower-voltage smaller-scale AC or direct current (DC) systems which serve multiple customers, while nano-grids typically serve a single building or very small area with DC power up to 100 kW [32]. The distinctions between mini-grids, micro-grids, and nano-grids are not sharp; for instance, while many consider a nano-grid to be one-building by definition, others use this term to describe extremely low-powered micro-grid-type community installations [33]. Local grids can be autonomous or optionally-islanding [34], though in the case of rural electrification, they are typically designed to be autonomous. However, the potential for their eventual interconnection with each other and with the the national grid in the future is an exciting field of research [35].

In areas where homes are sparse, stand-alone electrical technologies such as the solar home system (SHS) may provide the most affordable, albeit capacity-limited, electricity access. These modular systems include photovoltaic (PV) generation, power electronics, battery storage, and lights. They may also include sockets which can be used to charge electronics, or which can allow for the connection of appliances

such as televisions and fans. There was a push from multilateral organisations, governments, and the private sector in the 1990s and 2000s to expand the penetration of SHS [36] given their straightforward deployment (i.e. they are able to be shipped as a self-contained product) and immediate benefits (i.e. there need not be any delays between purchase and usage). However, SHS are not a silver bullet. They require adequate solar potential to charge the battery and power appliances, so their useful output varies spatially depending on local climatic conditions. Furthermore, it can be argued that SHS – particularly lower capacity versions – do not provide full electricity access, as they may not allow the household to use power-intensive appliances domestically or productively (i.e. for economic activities) [37]. SHS can be seen as complementary to other electrical services or as a stepping-stone towards higher access [38], and the possibility for SHS interconnection to create a higher capacity, modular micro-grid is another active subject of research [39, 40].

The choice between these grid types and electrical technologies can be fraught. While the electrification gap is massive and closing it is urgent, matching technologies carefully to local context, needs, and affordability is critical to generate uptake. To tackle this dissonance of specificity and scale in the electrification design process, automated decision support for grid type selection can helpfully accelerate locally-tailored design.

Designing Grid Topology

Grid topology design is quite standardized in the context of urban on-grid electrification. These networks are usually configured as radial, open- or closed-ring, or multi-radial [41], and can either be over- or under-ground, depending on the reliability required and the aesthetics and convenience of construction for each.

Topology design is, however, less routine and predictable in rural regions of LMICs, and particularly for local autonomous grids. In this context, the size and shape of communities to be electrified, as well as their reliability expectations and affordability constraints, will impact topology selection. As these areas have high sensitivity to affordability, the cost impact of topology choice based on wiring length,

gauge, and routing required must be considered. Furthermore, as new users may be hesitant to trust the electrical system, the level of reliability offered by a topology (i.e. by providing more or fewer alternative routes in the case of line failure) must also be selected to match consumer expectations or the system may fail to increase access.

When designing mini- or micro-grids, radial or hub-and-spoke topologies [42] are often applied based on their simplicity. However, loop [43] and mesh [44] micro-grids can offer higher reliability among other benefits. In either case, the precise location of connection points dictates the length of conductor required and therefore the distribution infrastructure cost. High-resolution spatial data pin-pointing potential connection point locations are therefore required. Unfortunately, such data can be difficult to find in rural off-grid areas of LMICs: most sources either suffer from large gaps or poor resolution. As such, improving the quality and availability of these spatial data, and developing methods to collect them quickly and accurately, represent avenues to accelerate community-specific grid design.

Specifying and Sizing Generation and Storage

The sizing of generation and storage technologies depends on the anticipated needs of potential users as well the availability and intermittency of the generation resource (e.g. solar, wind, etc.), both of which tend to vary geographically [45]. Additionally, local topography can be a deciding factor in the feasibility and cost of installing certain generation types. For instance, unshaded areas are needed to implement solar PV generation – where these are scarce, the costs of clearing an adequately sized area must be incorporated in the full system cost.

Generation sizing also depends on energy demands and how these may grow into the future. Demand growth is important to consider even in the first stages of electrification, as access to electricity, including via small-scale systems like SHS, has been found to stimulate loads which increase over time, particularly via social pressure and neighbourhood influence to acquire higher-consumption appliances [46]. These local forces influence system viability, and whether infrastructure becomes obsolete or stranded quickly.

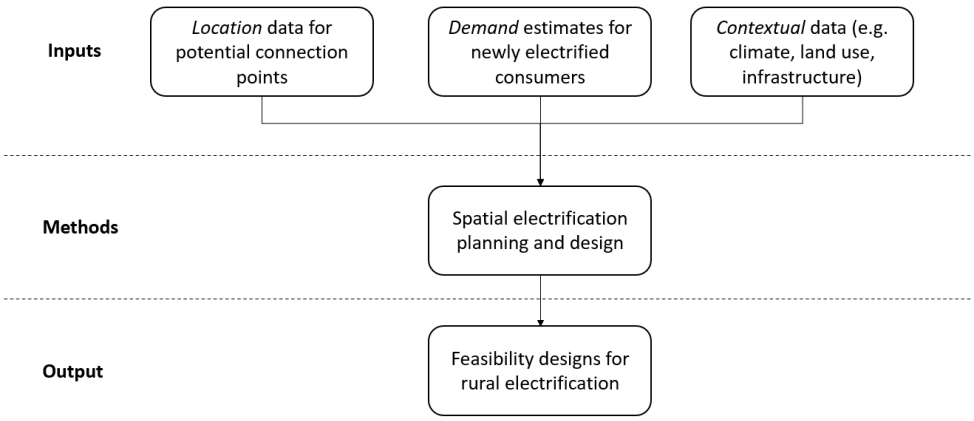


Figure 1.3: Data inputs, method, and output for spatial feasibility design of electricity access systems for off-grid rural areas.

As energy consumption depends on a complex mix of continuously evolving temporally and spatially dependent social practices [47, 48], spatially specific demand data and forecasts are required for generation sizing. Such local demand data are difficult to collect quickly and accurately; therefore, methods to better estimate site-specific present and future demands would again help to address the tension between specificity and scale in electrification design.

1.1.2 Data Needs for Spatial Electricity Access Design

Drawing from these electrification design stages, three critical data needs emerge to enable greenfield electrification in rural off-grid areas: (1) the location of potential connection points, (2) demand estimates, and (3) contextual information including climate and topography. These data inputs are illustrated in Figure 1.3. As alluded to above, the availability, resolution, accuracy, and completeness of these data vary greatly in off-grid LMIC contexts.

Locations of Potential Consumers

Accurate location data for potential connection points are required for grid type selection, system sizing, and topology design. These data need to be accurate at the dwelling level (i.e. approximately 5 m resolution). Additionally, location data need to be highly complete and accurate, ideally representing 95% or more of existing homes.

Few existing data sources meet these requirements. There are high-coverage lower-resolution census and raster datasets available, such as WorldPop, the Gridded Population of the World (GPW), and the High-Resolution Settlement Layer (HRSL), which locate population density rather than individual connection points. Alternatively, there are higher-resolution lower-coverage vector datasets such as the OpenStreetMap (OSM) which locate individual connection points, but which tend to have accuracy, completeness, and recency issues. Existing data sources will be reviewed in Chapter 2. Any existing data sources which have both adequate resolution and high accuracy tend to be either proprietary, limited in geographic scope, or both. There is a need for more accurate, complete, high-resolution and up-to-date location data for potential connection points in LMICs to facilitate electrical design in rural off-grid areas.

Demand Estimates

The demands of off-grid communities once electrified must be estimated to size electricity access systems. Load profile estimates at the connection point (i.e. household) and system (i.e. household, community, or country depending on the technology used) levels are required. These should not only represent average load, but should indicate peak demands and diversity to ensure that the system is robust to the full range of expected behaviour. For instance, to size batteries in a solar micro-grid, an understanding of whether peak load will occur in the day or after dark is needed. To size renewable-energy-based generation, seasonal variation in generation potential based on climatic conditions must be accounted for, as well as seasonal variation in energy consumption. Projections of demand growth are also useful when deciding whether and how to oversize a system to accommodate future needs.

Empirical demand data in newly-electrified rural LMIC contexts is limited. As such, bottom-up modelling is often used to estimate anticipated demands; however, the input data required to configure such models (i.e. appliance usage times and occupancy patterns) are hard to find in LMICs. Given the diversity of cultures in rural areas of LMICs, and the dependence of energy usage on cultural norms (as

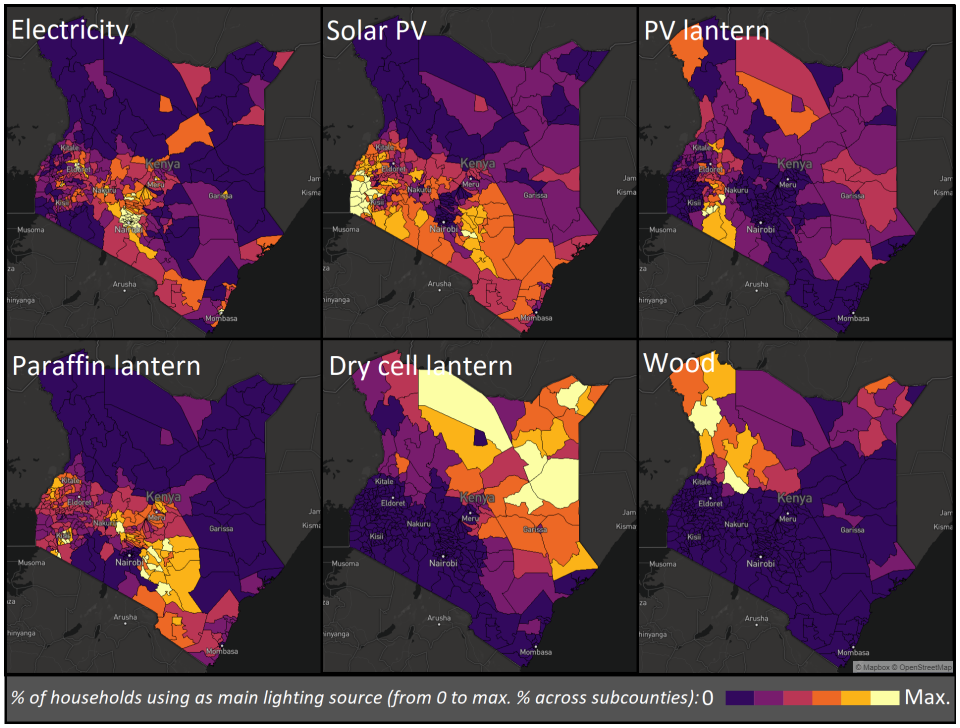


Figure 1.4: Prevalence of different lighting technologies across Kenyan sub-counties [49, 50]. Energy usage choices vary geographically, motivating the need for electrification strategies and policies which account for spatial variance.

illustrated in Figure 1.4), demands can vary greatly by region. As such, community-specific input data is best where available. Costly surveys in off-grid communities can be used to collect such data on energy aspirations; however, off-grid customers have limited exposure to electricity and therefore struggle to estimate their future demand accurately (as will be reviewed in detail in Chapter 2). There is a need for increased empirical data collection in newly-electrified areas of LMICs, and in the meantime, for novel methods to apply existing empirical data to generate site-specific and realistic demand estimates.

Local Contextual Data

Local climate and topography inform the feasibility of rural electricity access systems. These factors dictate which land is available for infrastructure construction and which types of generation and storage may be useful. Land use (e.g. deserted, farmland, built-up) and topography (e.g. elevation changes, waterways) can be used to determine, for instance, which land may be best-suited for construction of

transmission lines. Similarly, the placement of existing road infrastructure affects distribution feeder routing, as following roads generally results in lower construction costs. The relative solar and wind potentials in an area determine which generation infrastructure will be most efficient, and this in turn influences how much storage is required to fill generation gaps (e.g. at night or cloudy times in the case of PV). These considerations are particularly important in the case of autonomous micro- or mini-grids, which must be self-sufficient through their own generation and storage assets without relying on any grid backup.

Helpfully, these contextual data are available through large climate and space agencies or in academic repositories at a resolution appropriate for rural electrification design. Land use data are globally available for each year from 2015 through the Copernicus Global Land Service [51]. Larger infrastructural and topographic features such as roads and waterways are more reliably captured in the OpenStreetMap (OSM) [52] than smaller features like dwellings, and can be visualised online through platforms like the Open Infrastructure Map [53]. Additional purpose-built platforms such as Renewables.Ninja [54], GlobalSolarAtlas [55], and GlobalWindAtlas [56] provide the weather and climate data required to design electrification at high spatial resolution.

1.2 Research Sub-Questions and Scoping

This motivation leads to the main question of this thesis:

Can we design affordable electricity access systems suited to local spatial context and needs at the scale required to close the global electricity access gap?

Given the breadth of this question, it is important to specify the particular scope tackled here. As much of the remaining electricity access gap is rural, this thesis will focus on design challenges specific to rural settings. Additionally, while rural electrification design includes both spatial and temporal elements, this thesis will primarily tackle spatial design problems. Within this scope, the

work to answer the overarching question of this thesis is approached through four interconnected sub-questions:

1. *Can we accurately map rural off-grid populations for electrical system design?* This question first tackles the location data gap and seeks means to map rural off-grid populations for system design.

2. *Can we develop methodologies to rapidly map off-grid populations at a global scale?* This question explicitly addresses the scalability of mapping approaches, cognizant of the enormity of the task of global home-level mapping and the urgency of achieving sustainable electricity access in LMICs. This work is limited by the costs of required data inputs (i.e. satellite imagery), which while available, are not within the scope of the thesis budget at a global scale. Nevertheless, methodologies are investigated using a data sample with the aim to create a proof-of concept for global scalability in cost and time dimensions.

3. *Can we estimate the diverse and spatially specific energy needs of off-grid populations?* This question tackles the need to accurately estimate demands in currently off-grid communities using existing spatially disaggregated empirical and scalable data sources. This work encompasses currently available technologies; potential future uses of energy which are currently unavailable in the study context (e.g. the use of electric vehicles) are not considered.

4. *Can we design appropriate least-cost electrical systems to match community spatial context?* This question explores practical methods for the spatial, feasibility-level design of electricity access systems. The work to explore this question leverages the location and demand data developed through previous explorations. This question tackles feasibility stage electrical system design, including costing and equipment specification; final design is out of scope. Scalable yet context-specific spatial design methods are investigated for the purpose of planning and costing at the regional or national level.

While much of the work to answer these sub-questions is applicable across LMICs, this thesis tackles rural electrification particularly within the context of sub-Saharan Africa. More specifically, the countries of Kenya, Sierra Leone, and Uganda are used in case studies throughout. These countries have ongoing rural electrification efforts at various stages, and a diverse selection of rural settlement styles which make them interesting and pertinent to the spatial design problems addressed herein.

Finally, it must be noted that this thesis prioritizes renewable electrification. Given the ongoing climate crisis and its devastating effects on LMICs, as well as the increasing cost-competitiveness of renewable energy technologies, it is assumed that it will be economically and politically advantageous for LMICs with abundant renewable energy potential to "leapfrog" directly to climate-compatible development and renewable-based electricity access.

1.3 Thesis Structure

To begin the work, state-of-the-art datasets, literature, and methods for the spatial design of electricity access systems for rural off-grid regions of LMICs are first reviewed in Chapter 2. Two key data gaps are identified: (1) accurate, complete, and high-resolution locations of potential connection points; and (2) accurate site-specific demand estimates in off-grid areas. A methodological gap is also identified for practical and scalable spatial design methods which account for home-level settlement distribution at low computational expense. The location data gap is first addressed in Chapter 3, which uses citizen science to map off-grid regions at home-level using satellite imagery as input data. Subsequently in Chapter 4, mapping is accelerated through computer vision and is proven to be scalable to a global level in terms of cost and time-investment. With potential connection points located, the demand data gap is next addressed. Chapter 5 uses existing large-scale socioeconomic datasets to quickly generate spatially specific demand estimates at low cost. These are shown to better incorporate regional context and intra-community variability than alternatives like access-tier-targeting. Finally, with locations and energy needs determined, a spatial feasibility-level system design framework is

developed in Chapter 6. This framework matches electrical technologies and topologies to regional needs and settlement context while charting modular system expansion pathways to accommodate future demand growth. The thesis concludes in Chapter 7, which discusses the implications of this work as a practical spatial design process for rural electrification and outlines areas requiring additional study.

2

Literature Review

Contents

This chapter reviews state-of-the-art data and methods for the scalable design of context-appropriate rural electrification and identifies gaps which motivate the work of the thesis. Spatial population data useful in electrification design, including census, raster, and vector data, are first reviewed in Section 2.1 with a focus on accuracy, completeness, recency, and cost. Methods to collect these data where existing sources are incomplete are also reviewed. Next, demand data availability and estimation methods for LMICs are reviewed in Section 2.2. Finally, geographic information systems (GIS) and spatial design methods for electrification design are examined in Section 2.3, including local and large-scale design models. Research gaps identified in this chapter, highlighted in **boldface** throughout, are discussed in Section 2.4.

2.1 Locating Potential Connection Points

To expand electrical access, off-grid populations must first be located. It is impossible to design an electrification system without first knowing where the people who need access to electricity services are. To size and specify grid topologies and technologies in detail, the precise locations of all potential connection points (i.e. homes and businesses) are needed.

Unfortunately, **existing georeferenced data which locate off-grid populations in LMICs have accuracy, completeness, resolution, and recency issues (Gap 1)**. Generally, three types of georeferenced data can be used to locate and enumerate off-grid populations in rural areas: (1) census data aggregated by political district or enumeration area; (2) population counts or densities aggregated by raster grid cell; or (3) vector-based electronic topographic maps which include precise building footprints. Each has certain issues when locating potential electrical connection points in off-grid areas of LMICs, as discussed below.

2.1.1 Census Data

While population counts are often available in national census datasets, their usefulness in grid design can be hindered by long gaps between census rounds and the irregular sizes and shapes of political districts. Though the United Nations recommends collecting census data at least once every ten years [57], some countries (especially those which are very low income or unstable) have longer gaps between rounds. For instance, the last census conducted in the Democratic Republic of Congo was in 1984 [58]. This 38-year-old data is unhelpful in electrification design, as populations grow, change, and relocate over time. Additionally, the political borders used to aggregate census data are frequently arbitrary colonial relics, the usefulness of which can be questioned [59], as they do not necessarily account for deeper underlying tribal and cultural differences which are likely to influence energy usage norms. While population data from a recent census can be used to estimate the total electrical capacity required to serve a region based on its total population, they offer little information about how demand may be dispersed at

the household level, or where potential connection points might be. As such, they cannot enable the detailed design of distribution topologies.

2.1.2 Population Density Rasters

Raster datasets capture population density over uniformly sized grid cells, eliminating the issues associated with arbitrary political districts used in census population data aggregation. They typically extrapolate on empirical input data to provide more spatially uniform and comprehensive coverage. For instance, the GPW dataset produced by the Center for International Earth Science Information Network (CIESIN) extrapolates upon census data with minimal modelling to estimate world population at 30″ resolution and five-year intervals from 2000 to 2020 [60]. While GPW intentionally minimizes the modelling involved in their population estimates to preserve the integrity of the original empirical inputs, other initiatives use much more intensive modelling to achieve higher resolution and differing data types. For instance, WorldPop draws on census, survey, satellite, and cell phone data, leveraging machine learning to produce gridded population datasets with resolutions up to 3″ [61]. The HRSL developed by CIESIN and Facebook [15] achieves even higher resolution using machine learning and computer vision alongside census and OSM training data inputs to produce population density rasters at 1″ resolution. These data are visualized for comparison in Figure 2.1. While predictive and machine learning approaches to population density estimation can augment resolution and fill data gaps, they also introduce increased potential for error, and so should be used with this caveat in mind. Furthermore, though their resolutions are very detailed compared with the size of the earth, each of the datasets mentioned could still misrepresent the spatial configuration of households in communities smaller than grid cell size, or those whose homes are unevenly distributed throughout grid cells. They cannot precisely locate potential connection points for distribution grid design and detailed electrical planning.

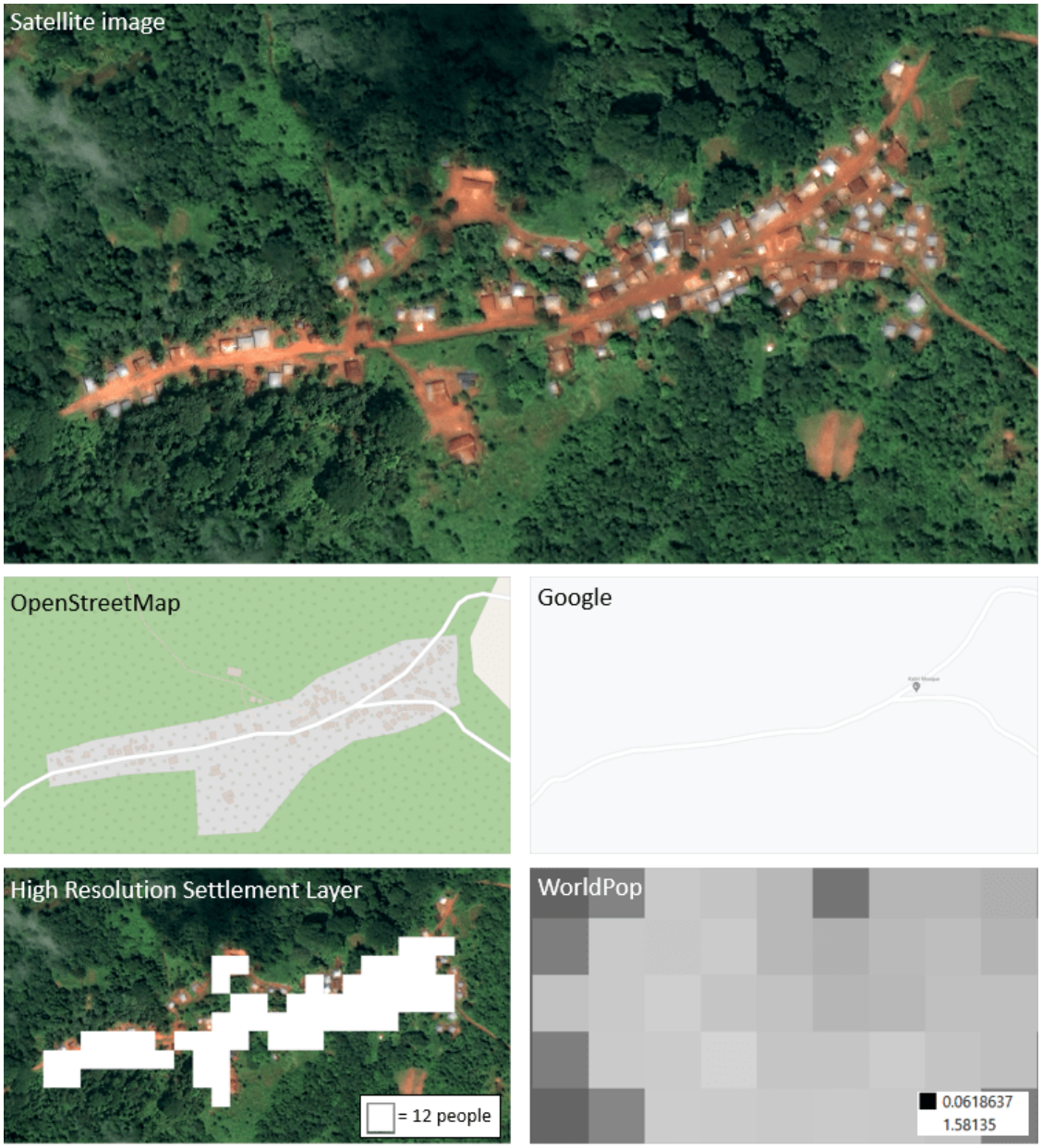


Figure 2.1: Satellite data for a sample community in Northern Sierra Leone (9°32'47.823", -12°10'9.611") alongside OpenStreetMap [52], Google Maps [62], High Resolution Settlement Layer [63], and WorldPop [16] data. Note that the OpenStreetMap does not have all dwellings indicated, Google Maps misses all buildings besides one, the High Resolution Settlement Layer misses some buildings, and WorldPop is too low-resolution to give any indication of potential connection points. The Gridded Population of the World dataset is not visualised because the grid cell size is larger than this sample, and simply shows a count of 137 people in the cell containing this community.

2.1.3 Vector Maps

Vector maps can offer higher resolution population data than census or raster data, as their features are precisely georeferenced instead of being aggregated at grid-cell level. Indeed, open-source topographic maps which include precise vectors for home locations, such as the OSM [52], are an ideal data source for detailed distribution grid planning. However, these data tend to have gaps and lack uniformity. For instance, a sample of OSM building locations is shown in Figure 2.2, and a large data collection gap is evident. Unhelpfully, there is a higher prevalence of these data gaps in remote and rural regions with low energy access. Furthermore, "non-Western" housing styles (e.g. small, thatched-roof homesteads far from paved roads) tend to be under-represented in these vector data sources. While this issue is being tackled in specific areas, especially for humanitarian purposes [64], many gaps and missed homes persist. Additionally, fully open-source efforts like OSM generally have no regulated update frequency or temporal data uniformity, which again makes it difficult to be certain that they are up-to-date and can be used to produce grid designs relevant to present day community configurations.

2.1.4 Collecting Location Data

Where there are gaps in existing population location data, data must be collected to enable electrical design. Typical methods such as in-person site surveying can be used. However, these **traditional survey-based data collection methods are costly, inconvenient, and difficult to scale, particularly in rural areas of LMICs (Gap 2)**. Reaching remote communities can require lengthy, dangerous, or expensive travel on poor infrastructure [67]. The strong power structures in many rural communities, such as the chieftaincy structures in many African communities and associated hierarchical land relations [68], add cultural complications which must be navigated carefully when entering a community to perform surveying (e.g. when demarking land ownership for electrical infrastructure placement during surveys). Additionally, off-grid communities are (rather obviously) likely to have minimal electrical access and possibly limited connection to wireless or mobile

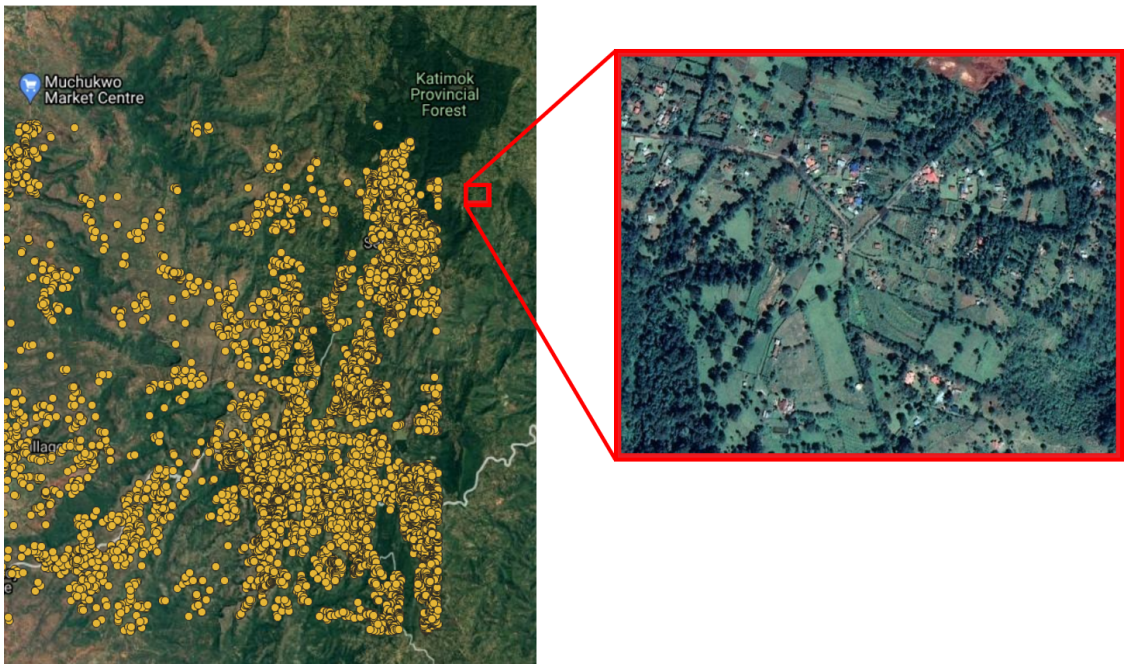


Figure 2.2: Sample of OpenStreetMap building locations in Baringo County, Kenya (0°34'01"N 35°47'37"E) [65] overlaid on Google Earth imagery [66]. Note the sharp edge to the right of which there are no buildings identified. However, as shown in the expanded area, there are buildings present. Such data gaps are common in rural areas.

networks, making efficient electronic data collection methods (e.g. mobile-phone-based geo-tagging, survey collection, and mapping) less practical.

Given these difficulties with site surveying, remote data collection methods can be useful in mapping potential connection points. Remotely sensed aerial or satellite imagery can facilitate cheap and effective community surveying: the International Federation of Surveyors finds land administration to be three-to-five times cheaper using annotation of remotely sensed imagery than using traditional field surveys [69]. At a local level, such data can be annotated by experts, local citizens, or engineers to map individual communities. To accelerate this, methods like citizen science and computer vision might be applied to achieve mapping at scale.

Satellite Imagery

Satellite imagery is a key data input to accelerated mapping. Though earth observation (EO) imaging has steadily increased since the inception of aerial

photography in the 1850s[1], this field has accelerated dramatically in the past decade, making very-high resolution (VHR) data ever cheaper and easier to access. Remote sensing was first popularised for military purposes during the world wars [70] and through the Cold War with the launch of the Sputnik and Explorer satellites [71]. However, public oriented EO programmes such as Landsat [72] soon opened up remote sensing for scientific research and resource evaluation. Other governmental (e.g. Sentinel) and commercial (e.g. IKONOS, Pléaides, and WorldView) EO programmes followed, producing data used in diverse research linked to geography and the natural world. EO data have, for instance, been used to characterise vegetation phenology [73], analyse cyclone intensities [74], forecast malaria [75], determine ocean water depths [76], and predict poverty [77], to name but a few applications.

Satellite EO data may be openly-accessible or restricted for use by paying clients or in government applications, depending on their resolution and potential applications. Platforms like United States Geological Survey EarthExplorer [78] and the European Space Agency Copernicus Hub [79], for instance, make EO data openly available online. EO images which are made openly and freely accessible typically have resolutions in the range of tens of meters; for instance, 10-60 m Sentinel-2 imagery is available on the Copernicus Hub [79]. Meanwhile, VHR data as precise as 0.3 m per pixel is available commercially via providers like Planet, Airbus, or Maxar. While imagery prices are continuously decreasing as more satellites are launched, they still may present a hurdle for researchers and electrification planners depending on the volume of imagery required. Lacking the resources to buy imagery, one can turn to samples of VHR EO imagery which are made available freely for competitions such as SpaceNet [80] or humanitarian purposes via initiatives like the Maxar Open Data programme [81]. These samples are aggregated in resources like the OpenAerialMap [82]. VHR EO imagery can also sometimes be accessed through platforms like Google Earth, though such platforms tend to only allow the data to be manipulated through proprietary interfaces. As open spatial data infrastructures

[1]While these early photographs have been lost, examples as early as the 1860s still survive. See: `https://www.metmuseum.org/art/collection/search/283189`.

become more normalized [83], the accessibility of spatial data, including VHR EO satellite imagery, should continue to increase.

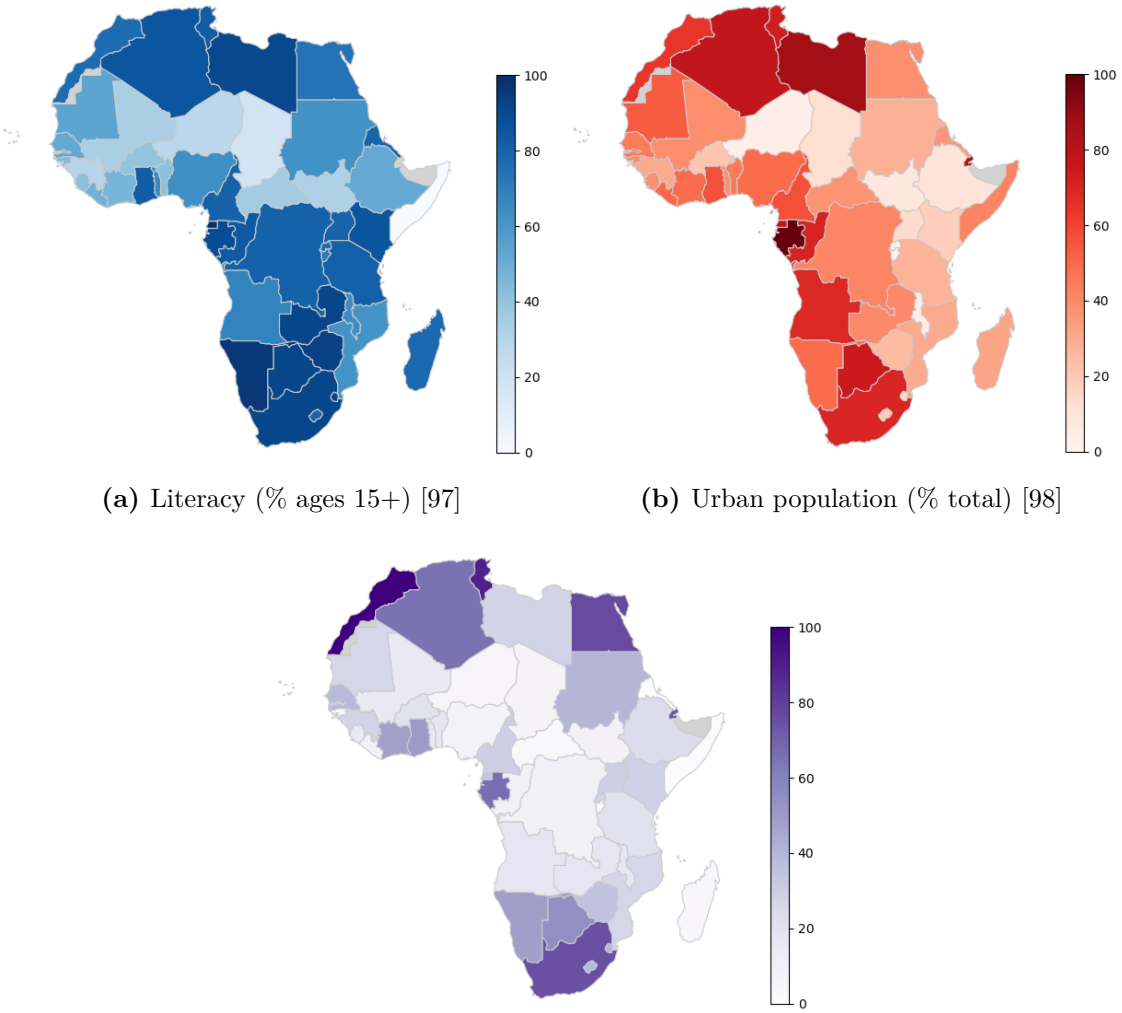
Citizen Science

Citizen science, or the active participation or collaboration of the public in scientific research [84], is one possible route to map rural off-grid areas based on satellite imagery at scale. The degree of public involvement in citizen science can vary from contributory to co-creative or co-led; contributory approaches are typically the most popular, wherein citizen scientists primarily assist in data acquisition or annotation.

One of the core principles of citizen science is mutual benefit to both the professional and the citizen scientist [85]. For the professional, this method can accelerate research where extensive data collection or annotation are required, and can diversify the data collected. For the citizen scientist, it offers a gateway into the often-obscure world of academic research, access to cutting-edge data, access to professional researchers (virtually or in-person) for guidance and support, learning opportunities, and a sense of community and satisfaction.

While citizen science has traditionally been most useful in ecological and astronomical research [86], this approach now spans diverse disciplines including social sciences [87] and engineering [88]. Citizen science could accelerate the mapping of rural off-grid areas in EO satellite imagery by leveraging the enthusiasm of citizen researchers to annotate satellite data.

Mapping research which uses citizen science can be on-the-ground or remote. On-the-ground mapping involves citizens recording ground-truth locations using a mobile phone camera or global positioning system (GPS). This approach is employed, for instance, in the popular iNaturalist project to create geotagged photographic records of natural features or wildlife [89–91]. Remote mapping involves citizens observing and annotating an existing map or data source. This is more commonly used to map larger-scale features such as buildings and roads, as in the Missing Maps initiative [92–94]. In rural regions of LMICs, an on-the-ground approach would rely on a limited contributor base likely to have lower-than-average availability of mobile



(a) Literacy (% ages 15+) [97]

(b) Urban population (% total) [98]

(c) Internet use (% total) [99]

Figure 2.3: Literacy, urban population, and internet usage in African countries[2]. Note the correlation between countries with more rural populations, less literacy, and fewer internet users. The rural areas needing electrification are likely to have less people on-the-ground able to contribute to mapping via an online text-based citizen science interface. Countries with no data available shown in grey.

phones, internet access, and literacy rates, as illustrated in Figure 2.3. Furthermore, it may be inappropriate to expect vulnerable communities to contribute voluntarily to research when they already face disproportionate (and gender-disparate) time poverty [95, 96]. A remote citizen science approach could be applied to leverage the spare time of those in more privileged circumstances with an existing desire to contribute to the mapping of LMICs for electrification.

[2]Figure published in [100], created by author.

Computer Vision

Computer vision is a field of artificial intelligence (AI) which uses computers to extract meaningful information from images. It can perform tasks – like satellite-imagery-based mapping – which humans can accomplish by eye. Emerging as a topic of research interest from the 1960s onward, computer vision suffered the effects of the AI winter [101] before experiencing a boom in the 2010s onward.

While classic computer vision methods including Canny edge detection, the Hough transform [102], Scale Invariant Feature Transform [103], active contours [104, 105], and morphology [106] have been applied to home mapping in satellite imagery, these approaches perform poorly in rural contexts with complex backgrounds (i.e. forests, farmland) and varied home shapes (i.e. rounded roof edges) [107]. This is because they generally leverage the building regularity, rectangularity, and shadows appearing in uniform urban contexts to detect homes. In more varied rural contexts, modern computer vision algorithms underpinned by artificial neural networks (ANNs) can map homes more effectively than classic alternatives.

Using webs of synthetic neurons, ANNs seek replicate human thought processes including prediction, pattern detection, and optimisation [108]. The ANNs used in modern computer vision algorithms tend to be discriminative (i.e. they learn the boundaries between classes of data) and deep (i.e. multiple layers of neurons exist between the input and output [109]). Furthermore, they are typically convolutional: they convolve sets of filters with input data to produce feature maps. Convolutional neural networks (CNNs) operate over data volumes instead of vectors, allowing them to process three-dimensional data like images (with dimensions of width, height, and colour). While CNN-driven computer vision can be incredibly powerful and robust, performance is dependent on the selection of a well-suited algorithm and the availability of appropriate training data.

In applying computer vision to remote mapping in satellite imagery, three major CNN-driven algorithm types are most useful: classification, object detection, and segmentation. The differences between each of these tasks are illustrated in Figure 2.4.

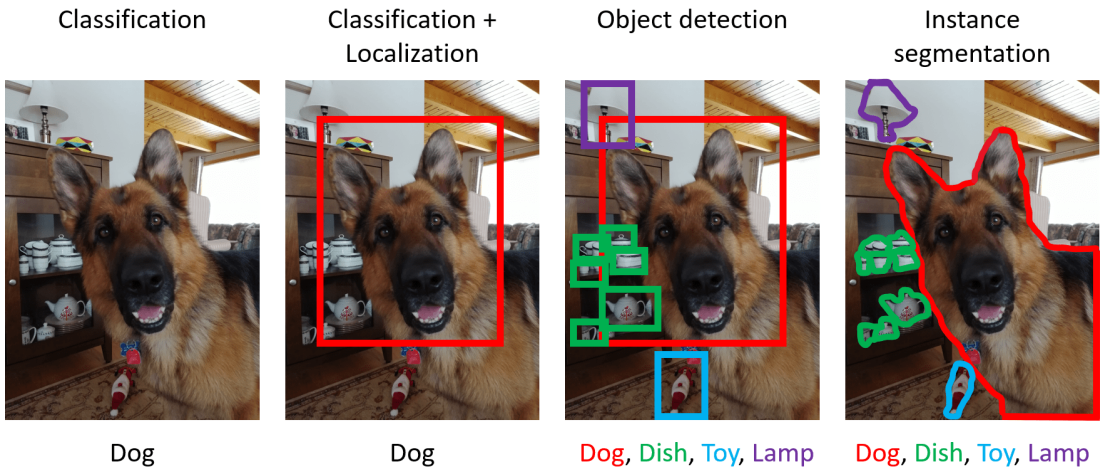


Figure 2.4: Image classification, object detection, and instance segmentation visualized on a sample image. Object detection and instance segmentation are used to find four feature categories (i.e. dog, dish, toy, and lamp). The combination of classification and localization is also included to illustrate the difference between this and object detection. Object detection identifies the features in bounding boxes, while instance segmentation identifies them at the pixel level. Note that semantic segmentation would not identify specific instances of each type (e.g. there would be no differentiation between each dish).

Classification categorizes images based on their contents. While work on CNN-based classification began with the LeNet digit recognition model in 1998 [110], it was the unprecedented high performance of the 2012 AlexNet CNN model [111] that spurred the modern explosion of CNN-driven image classification, including competitors like VGGnet [112], GoogLeNet [113], ResNet [114], and Densenet [115]. While classification is less applicable to mapping specific connection points in satellite imagery, it can be used to pre-filter satellite images into subsets which are likely to be populated or unpopulated. For instance, a classifier or CNN can be used to classify image tiles as likely to contain a village or not, as applied by Sun in image subsets denoted "superpixels" [107].

Object detection identifies specific classes of objects in images and indicates their locations in bounding boxes. Early CNN-driven object detection models, such as Regions with CNN features (R-CNN) [116], simply applied a classifier to a certain number of subsets of an image at various scales. However, this was quite slow, so subsequent iterations (e.g. Fast R-CNN [117], Faster R-CNN [118]) used more efficient algorithms to propose regions for classifier application. Other concurrent

work took a regression-driven approach. For instance, the You Only Look Once (YOLO) model detects objects in real-time through a single forward pass of a CNN. While incredibly efficient, YOLO struggles with small and densely-packed objects [119]. Object detection can be used to identify features such as roads and homes in satellite imagery. For instance, the You Only Look Twice algorithm detects cars and homes amongst other features on Planet and DigitalGlobe satellite images alongside aerial images with GSD as high as 0.15 m, achieving an F1 score (i.e. a harmonic mean of precision and recall) of 0.6 to 0.9 depending on the category detected [120]. The SpaceNet [80] and Cars Overhead with Context [121] datasets are used as training data alongside researcher-generated annotations. Others have used the Xview dataset [122] and super-resolution techniques or the UC Merced land use dataset[3] and a two-stage classification and object detection pipeline [123] to achieve similar and better results.

Segmentation also identifies image features and classes, but in arbitrary shapes at the pixel level. This can be semantic (i.e. each pixel is assigned a class, but no differentiation is made between instances of the same class) or instance-based (i.e. each instance of a class is individually identified). Following the first major CNN-based segmentation model, the Fully-Covolutional Network (FCN) [124], the U-net [125] and DeepLab [126] models built on this concept through the addition of skip connections and spatial pyramid pooling among other improvements. Semantic segmentation has been proven as a feasible method for building detection. For instance, it has been used to generate three-channel (e.g. road, building, or background) labelled images [127] using the 1 m ground sample distance (GSD) Mnih dataset [128] and for pixel-wise building prediction from 0.3 m GSD imagery based on GIS data for home locations in Washington DC [129]. Note, however, that these applications are largely based in urban contexts and urban building detection, where plentiful, open, and complete data are more likely to be available, such as the Mnih dataset [128] or the Inria dataset covering urban contexts in the United States and Austria [130]. Other similar works have leveraged 0.075 m aerial

[3]http://weegee.vision.ucmerced.edu/datasets/landuse.html

data openly available in Christchurch, New Zealand [131, 132] and 0.8 m urban imagery in China [132]. While the algorithmic improvements in such studies are valuable, without the availability of high-quality rural training data, they cannot be applied to map buildings in off-grid LMIC regions.

Barriers to Computer Vision

There are certain barriers which prevent practitioners from using computer vision algorithms to accelerate connection point mapping. First, modern computer vision algorithms often do not have user-friendly interfaces, and some level of programming ability is typically required to use them. For instance, OpenCV, Keras, TensorFlow, and PyTorch – which are frequently used to implement state-of-the-art computer vision algorithms – each require the use of a command line interface and/or writing a custom Python script to be used, depending on the task. This can be intimidating to less programming-savvy practitioners. Companies like Picterra [133] and Petuum [134] offer guidance and user interfaces to make computer vision based mapping and AI accessible, but do so as a paid service or platform which may not be accessible in low-resourced LMIC contexts. The open-source community is endeavouring to create similar platforms such as Mapwith.ai [135], but these are still in their infancy.

Additionally, these powerful methods can only be used to their full potential when high-quality data are available to train them. While training data availability for computer vision has exploded in recent years, including data which can allow detection or segmentation of buildings from EO imagery, **there are minimal training data for building detection in rural areas of LMICs (Gap 3)**. For instance, SpaceNet – perhaps the most well-known training data available for building detection – lacks coverage of rural areas. SpaceNet data include: 685,235 building footprints labelled on satellite imagery in Las Vegas (USA), Paris (France), Shanghai (China), Khartoum (Sudan), and Rio de Janeiro (Brazil) at 0.3-0.5 m resolution [80]; 126,747 footprints of off-nadir buildings on 0.5 m satellite imagery in Atlanta (USA) [136]; 48,000 building labels on 0.5 m imagery from Rotterdam (Netherlands) [137]; and 24 monthly images from 100 urban locations

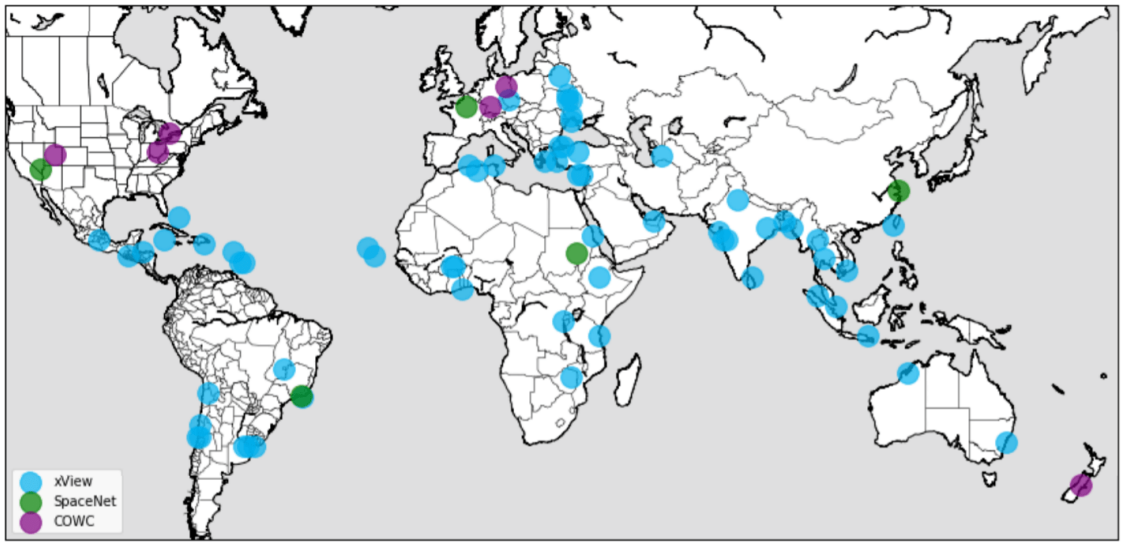


Figure 2.5: Satellite image sample locations from the XView training dataset [122], alongside locations from the SpaceNet and Cars Overhead With Context (COWC) datasets for comparison. Even datasets such as this one with relatively high geographic diversity tend to have minimal coverage in low- and middle-income countries (particularly in Africa) and and more urban coverage than rural coverage.

at 4 m resolution with approximately 11 million building annotations representing approximately 500,000 unique buildings [138]. These data are frequently used in building detection research and competitions, including the DeepGlobe competition [139] and the AIcrowd Missing Maps competition [140]. However, they are not particularly useful in rural dwelling detection, as rural buildings are often surrounded by entirely different contexts than urban buildings.

Indeed, most available training data for building detection tend to be urban, HIC, or both. Training datasets which only cover non-LMIC countries include the Landcover.ai dataset in Poland [141], the SkyScapes dataset in Munich (Germany) [142], and the Mnih dataset in Massachusetts (USA) [128]. Datasets which include LMICs but only in urban areas include SpaceNet and the OpenCities AI Challenge dataset, which covers ten African cities with aerial imagery and OSM building footprints [143]. Some datasets are beginning to tackle this data gap; for instance, OpenAI Tanzania used OpenAerialMap [82] drone imagery over Zanzibar for their building mapping competition. Additionally, the XView dataset offers buildings on 0.3 m resolution imagery over a 1,400 km^2 area across multiple continents, and

includes LMIC areas, as shown in Figure 2.5. However, there is minimal coverage in Africa [122]. More training data coverage in rural, off-grid areas of LMICs is needed.

2.2 Understanding Energy Demands

Energy demand data is also required alongside location data to design electricity access systems. These can include industrial, productive use, and domestic energy uses. As this thesis primarily addresses rural electrification for energy access, energy demands within the household are the focus.

2.2.1 Household Energy Uses

Household energy use can be generally divided into the four categories: (1) lighting, (2) cooking, (3) thermal comfort, and (4) other appliances. The relative importance of these categories, and the preferred energy vector for each, varies based on context.

As discussed in Section 1.1, lighting is typically an early priority following electrification, as electrical lighting provides better illumination [144] than alternatives like kerosene while also lowering health risks [145]. Energy-efficient LED bulbs make electric lighting cheap and convenient, as they can be powered by small home-level PV systems such as SHS at low cost.

Conversely, electric cooking is generally a lower priority in newly electrified rural communities. This is evidenced by the fact that 2.6 billion people worldwide remain without access to clean cooking despite the devastating health impacts of cooking indoors with dirty fuels, which causes 2.5 million premature deaths each year [3]. Electric cooking may be perceived as more expensive than biomass or fuel-based cooking. As shown by Rhodes through studies in Peru, Nepal, and Kenya, it can also be unsuitable to local cooking culture [146], depressing uptake. It is therefore unlikely amongst newly-electrified populations unless some strong awareness-raising effort or technology subsidy is implemented to encourage it.

While newly electrified rural households in LMICs are likely to desire electric thermal comfort technologies as the climate crisis worsens [147], high-powered air conditioners will be difficult for many poor households to access. The second-hand

units typically available in LMIC markets are not very energy efficient and may create demands which cannot be managed by decentralized electrical systems [148]. Indigenous architectures in LMICs (e.g. earthen roundhouses in sub-Saharan Africa) can offer strong insulation, thus passively providing thermal comfort [149]; however, they may not be conducive to the installation of electric air conditioners and other modern thermal comfort technologies. As such, lower-powered air *circulation* technologies are more likely to be used than air conditioners in the near-term, though this is liable to change as thermal comfort technologies better-suited to the rural LMIC context are developed.

Other household appliances used for communication, entertainment, or information (e.g. radios, televisions, computers, mobile phones) and to facilitate domestic tasks (e.g. irons, washing machines, refrigerators, blenders) are prioritized variably by newly electrified households. Whether and when these appliances are acquired depends on household demographics as investigated by Rao [150] and income as shown by Poblete-Cazenave [151] among other factors. For instance, while a high-income household with school-aged children may prioritize acquiring a computer for educational purposes, an elderly couple with low income may be perfectly happy to obtain information via a radio and never purchase a computer.

As argued by Shove, Walker, and Hui, the specific combination of these demands present in a household or across households in a community will necessarily arise from a complex mix of continuously evolving temporally and spatially dependent social practices [47, 48]. Furthermore, as common appliances such as refrigerators, fans, and televisions become increasingly energy efficient [152], this balance of energy priorities will continue to evolve. To model the anticipated demands of currently off-grid households, approaches which integrate temporal and spatial specificity are therefore required. This is particularly important when designing small-scale electricity access systems (e.g. mini-grids and solar home systems), which may be the most affordable option in many rural off-grid contexts [26], and whose designs are more tightly bound to the spatial distribution of demands than a larger-scale national grid.

2.2.2 Demand Estimation Approaches

Generally, two classes of methods can be used to estimate domestic demands: top-down approaches, and bottom-up approaches. Whereas the top-down approach considers residential energy use as a unified energy sink (i.e. individual consumer behaviour is not considered), the bottom-up approach instead builds aggregated profiles from the consumer level to the community or national levels [153]. The macroeconomic data used to construct top-down models, such as country-wide gross domestic product (GDP) and unemployment rates, cannot capture the diversity of living conditions influencing early demands in newly electrified communities on-the-ground. Indeed, top-down approaches tend to produce less benefit and produce systems which fail in the mid- to long-term [154]. They are therefore not considered appropriate here to estimate demands for rural electrification.

Considering bottom-up methods, there are again two methodological subsets, namely statistical and engineering methods [153]. Whereas statistical methods rely on historical information and regression analysis to attribute consumption to particular end uses, engineering methods build consumption profiles based on expected equipment power ratings and physical behaviour. Again, considering the rural off-grid context, historical demand data are likely to be unavailable; this leaves bottom-up engineering methods as the most appropriate methodological subset. Typically, such methods incorporate stochastic variation in appliance usage to represent anticipated demand diversity.

In HICs where accurate empirical demand data are plentiful, stochastic bottom-up models are a popular and effective approach for domestic demand estimation. These models build demand profiles from the appliance-level to the household and community levels using detailed behavioural and appliance use surveys as inputs, such as time-of-use surveys, home occupancy patterns, and appliance data. Such data are applied by Capasso in modelling for Italy [155] and Yao for the United Kingdom [156], among other classic HIC studies. In the United Kingdom, the 10-minute resolution Time Use Survey data is usually taken as a modelling input [157–159]. Similar high-resolution time use surveys are used in Sweden and Germany

[160]. Where behavioural data are unavailable, Markov chains are frequently used to determine household activities for demand estimation [158, 159, 161].

Regrettably, **high-resolution and accurate demand estimation input data are typically not available in LMICs, particularly in off-grid or newly electrified areas (Gap 4)**. There are huge data deficiencies in LMICs (e.g. as studied for small island developing states [162], with regards to transport [163], and with regards to the Millennium Development Goals [164]) which limit understanding of energy demands. In the face of the demand data gap, simplified top-down demand estimation approaches can be tempting, as these can be applied more quickly than bottom-up approaches requiring numerous data inputs. As highlighted in the World Bank's guidance on demand forecasting, there is a "paucity of data" which leads practitioners to employ simple, crude heuristics [165] which fail to represent load accurately at the community level. While estimating the energy demands of rural off-grid communities may be difficult, it should not be overlooked or avoided; as Wolfram notes, domestic demand increases of the poor and near-poor are likely to be an important driver in medium-term growth in energy consumption globally [166].

The shortcomings of both HIC-developed estimation models and alternative heuristics have led to the development of demand estimation models specifically tailored to off-grid remote areas of LMICs. As previously discussed, most HIC-tailored models require data inputs unavailable in LMICs Additionally, many western-developed demand models struggle to capture the poverty-related, migratory, and cultural factors influencing energy demand in LMICs [167]. LMIC-tailored models are therefore constructed to overcome these challenges.

Stochastic bottom-up models built for use in rural LMIC contexts typically use energy and appliance use aspiration data collected from local site surveys to estimate anticipated demand. The 2015 open-source Python-based ESCoBox model by Boait [168], for instance, uses these parameters to create a Monte Carlo-based demand estimation tool. ESCoBox complements the widely-used Hybrid Optimization of Multiple Energy Resources (HOMER) software for mini-grid planning, and is applied by Sandwell for electrical demand simulation in Uttar Pradesh, India [169]. The

2016 LoadProGen model, introduced by Mandelli through a case study in Cameroon [170] and more formally defined with a Tanzanian case study in [171], provides a user-friendly graphical user interface for a similar LMIC-tailored approach, which is implemented in MATLAB and openly available for use. This model was soft-linked with a system dynamics model in [172] to improve long-term projections, and expanded in [173] with the motivation to model techno-economic feasibility of expanded electric cooking in Tanzania. These expansions were implemented in Python as Lombardi's lightweight and open-source "Remote-Areas Multi-energy systems load Profiles" (RAMP) model [174], which adds variability in cooking behaviour, thermal appliance power, and multi-energy vector analysis options.

While the RAMP model begins to incorporate mechanisms to evaluate temporal demand growth, **there are few studies in the literature which account for demand evolution in newly electrified rural areas (Gap 5)** [175]. This is problematic, since as highlighted by Opiyo, access to electricity (including via small-scale systems like SHS) stimulates loads which increase over time, particularly via social pressure and neighbourhood influence to acquire higher-consumption appliances [46]. Efforts to include demand evolution in modelling include the previously-mentioned systems dynamic approaches soft-linked with LoadProGen [172] as well as the combination of appliance projection with LoadProGen soft-linked with the Open Source Energy Modelling System (OSeMOSYS) long-term planning model [176].

2.2.3 Sourcing Data Inputs for Demand Estimation

Local surveys are frequently used to understand energy aspirations in rural off-grid areas of LMICs, which can be used as inputs to the demand estimation approaches detailed above. However, surveying does not always generate accurate input data, as people who have limited experience with electricity tend to overestimate their future use. Survey-based load profiles can fail to provide an accurate energy estimate compared to measured data [177], with the largest discrepancies in the load occurring at night. For instance, Blodgett's study employing energy-use surveys for eight

solar mini-grid projects in Kenya showed a mean absolute error of 426 Wh on 113 Wh consumption [178]. By instead using a proxy village method (i.e. where demand is modelled on load profiles of a similar recently-electrified community) mean absolute error was reduced to 42.4 Wh. While the proxy village approach performs well, it involves locating an adequately similar electrified village to the off-grid village for which demand is estimated; this can be difficult in LMICs with high cultural diversity or very low electrification rates, where the availability of electrified communities to serve as a proxy may be minimal. Similarly, demand estimates from energy-use surveys were found to overshoot by by 305% for seven off-grid solar systems in Louie's study in Malawi [179]. Service- or value-based surveying, as undertaken by Clements in Kenya and Bangladesh [180] and Hirmer in Uganda [181] respectively, can provide increased detail on user demands over traditional survey methods. Services (e.g. lighting and refrigeration), appliances (e.g. sewing machine, cooker), and values (e.g. security, family, religion) are easier for end-users to understand than kilowatt-hours. However, these approaches typically require a higher time investment to gather data, which can result in higher costs and increased delays to achieving electricity access.

Large-scale socioeconomic and demographic surveys – such as the Demographic and Health Surveys (DHS), the World Bank Living Standard Measurement Surveys (LSMS), and the United Nations Children's Fund (UNICEF) Multi-Indicator Cluster Surveys (MICS) – can also be used to fill demand estimation input data gaps at lower expense, as explored by Fabini [182] and Zeyringer [183] in the Kenyan context. The MICS programme, for instance, has completed 346 surveys in 118 countries on key indicators of women's and children's health since its inception in the mid-1990s [184]; its questions on appliance ownership, household characteristics, wealth, and other demographic markers can be leveraged in demand estimation [185].

While these large-scale surveys provide useful data to inform demand estimation, their questions on electricity access are neither as detailed nor as standardized as they could be, limiting their usefulness. For instance, the MICS and DHS questionnaires phrase their questions on electrical access simply as "Does your household have

electricity?" – no details on the duration or quality of the electrical connection are collected, although they do ask whether the household owns various electrical appliances [186, 187]. The LSMS asks more questions on electricity, including uses of electricity in the household, the duration and frequency of power cuts, the level of power consumption, and the presence of metering [188]. These questions can enable more detailed understanding of energy service and usage to facilitate demand estimation. The Multi-Tier Framework for Measuring Energy Access (MTF) represents the state-of-the-art benchmark for nuanced electricity access measurement, covering aspects such as affordability, reliability, and capacity [189]. Surveys using this framework are ongoing from the Energy Sector Management and Assistance Program (ESMAP) and the World Bank, though data coverage as of yet is minimal.

It must be noted that large-scale survey datasets are susceptible to overlooking local and indigenous knowledge. These surveys are often produced by wealthy Western institutions; their content and structure may be influenced by unconscious biases or Western values. For instance, while many large-scale surveys take ownership of high-powered appliances as an indicator of high wealth and energy access, Rao has argued that the expectation of eventual ownership of "white goods" appliances (e.g. dishwashers, washing machines) is a primarily white and Western phenomenon, and that appliance ownership aspirations actually vary based on local and familial wealth, class, race, and religion [150]. While local and indigenous knowledge is widely integrated in environmental management [190], ecological systems research, and conservation [191], it is less often incorporated in electricity demand and planning research. Demand estimation might benefit from the integration of indigenous and local knowledge, given the heavy ties between energy and the environment, where indigenous populations hold deep traditional knowledge.

2.3 Spatial Design for Electricity Access

Key design parameters in technical electrical design, such as renewable resource potentials and connection point locations, vary spatially. Furthermore, sociopolitical, cultural, and economic factors which affect energy demands also show geographic

variance. Given these spatially varying determinants of appropriate electrical design, automated spatial system planning is an exciting avenue to accelerate electrification. The intention is to design more contextually-appropriate systems at lower cost; ESMAP has already found that remote site assessment for electrification using GIS can decrease costs by an order of magnitude compared to traditional in-person methods [26].

GIS can be used to evaluate renewable energy resources such as wind [192–194] and solar [195–197] potentials to facilitate generation type selection and sizing for electrification. Such analyses are facilitated by increasingly available and high-resolution large-scale open spatial datasets on renewable resource potential. Online platforms such as Renewables.Ninja [54], GlobalSolarAtlas [55], GlobalWindAtlas [56] make such data easily available online via attractive graphical user interfaces. They can also be used for high-level planning to increase electricity access. For instance, GIS-driven analyses have been used to evaluate electrification strategies throughout Africa by differentiating generation types and grid architectures, as in Szabo's cost analysis [198] and Mentis' assessment of African wind potential [199]. In combination with night-time satellite imagery, GIS can also be used to estimate the proportion of rural people without access to electricity, as shown by Doll [200] and developed by numerous studies thereafter.

Increasingly, spatial design models for electricity access embed GIS into sophisticated algorithms which incorporate technology selection, system optimisation, and financial feasibility analysis. These electrical design models can be generally divided into "local" and "large-scale" categories.

2.3.1 Local Planning Models

Local planning models tackle the problems of local generation and storage sizing and distribution network design. Linear programming can be used to determine a cost-optimal generation system design given local constraints such as demands and weather, for instance as implemented by Huneke in the General Algebraic Modeling System for India and Colombia [201]. Alternatively, metaheuristics such as genetic

algorithms [202, 203], particle swarm optimisation [204, 205], simulated annealing, and tabu search [206] can provide more efficient solutions for generation design than linear programming, albeit at the expense of potentially finding non-globally optimal solutions. Local distribution network planning involves feeder routing [207, 208], locating substations, and siting any distributed generation within a distribution system. This can be again tackled via optimisation, as demonstrated by Paiva using mixed integer linear programming [209], or metaheuristics [210, 211].

The most popular tool for local system planning is HOMER [212, 213] developed by the United States National Renewable Energy Laboratory. This model simulates grid operation for a defined load and location over a range of infrastructure types and allowed sizes as defined by the user. It then selects an optimal infrastructure from those simulated based on selected criteria which can include cost and reliability. Similar but less popular alternatives include the Distributed Energy Resources Customer Adoption Model (DER-CAM) [214] developed at the Lawrence Berkeley National Laboratory, which provides decision support in identifying optimal renewable energy investments; the Hybrid Optimization by Genetic Algorithms model (HOGA) [215] developed at the University of Zaragoza, which simulates and cost-optimises renewable energy generation systems; and the RETScreen Clean Energy Management Software [216, 217] developed National Resources Canada to enable decision support for renewable energy development, implementation, monitoring, and reporting. While these models can be useful in optimal generation storage and sizing to suit local climate and costs, they do not tackle topology specification in detail and assume that the type and scope of grid implementation is known to the user. Notably, less well-known software options exist to plan distribution grid topologies, and designs are frequently conceived manually. The Reference Network Model [218] is one of the few relevant options available, which plans high-, medium- and low-voltage networks including substations and feeders, though its focus is on urban areas and not the rural LMIC context.

2.3.2 Large-scale Planning Models

Large-scale spatial electrical planning methods generally aim to establish the best-fit electrification technology (i.e. grid extension, micro-grid, or SHS) for any given area based on some defined criteria. This involves synthesizing some combination of rasterized environmental, social, and economic data through simulation and optimising for a least-cost system type. Early large-scale models include Monteiro's SolarGIS [219], which was further improved by Amador [220] and whose methods (or similar approaches) are applied in numerous subsequent works [198, 221–226]. As in the local planning models, optimisation and linear programming methods are often applied to this problem [183, 227, 228].

It has become academically popular, particularly as large-scale raster data and GIS have become increasingly accessible, to try and build new one-size-fits-all user-friendly large-scale modelling tools in the style of SolarGIS. For instance, the Python-based Network Planner [229] by Columbia's Quadracci Sustainable Engineering Laboratory computes costs for a limited number of electrification options (i.e. grid extension, SHS with diesel generator back-up, diesel-based mini-grids) across demand centers along a specified time horizon. This has been applied in the literature Ghana [230], Nigeria [231], and Liberia [232]. The Open Source Spatial Electrification Tool (OnSSET) [233] developed by Mentis and a large team from KTH Royal Institute of Technology and other partner institutions is another popular open-source large-scale planning framework. OnSSET identifies least-cost electrification technologies amongst grid extension, mini-grid, and stand-alone options in order to achieve a particular MTF tier of access. It has been applied in the literature in many contexts, including Malawi [234], Burkina Faso, and Cote D'Ivoire [235], and has been coupled with other energy simulation software including OSeMOSYS [236] and RAMP [237]. While the outputs of models like OnSSET can be useful in preliminary electrification design when scoping the coverage of different technologies, without consideration the distribution of homes and potential interconnections, they miss a significant chunk of the implementation picture.

Some models do aim to tackle electrical design at large-scale while still accounting for local settlement style. For instance, the Reference Electrification Model (REM) [238] optimises to find best fit technologies and also generates preliminary distribution grid designs. Though the initial REM work had a more large-scale focus [239], subsequent developments integrated local grid design [240–243]. However, this fully-integrated approach comes at the expense of increased input data needs and computational complexity. Additionally, REM is not open-source. There is therefore a gap for **scalable open-source models which account for local settlement distribution and diversity using existing empirical data at low computational expense (Gap 6)**.

Many large-scale models are now either hosted online or provide their results in online graphical user interfaces for easy access by practitioners. Such online tools include the Off-Grid Market Opportunity tool developed by the International Finance Corporation [244], the ECOWREX2 Map Viewer featuring data from the ECOWAS Regional Centre for Renewable Energy and Energy Efficiency [245], the Energy Access Explorer led by the World Resources Institute [246], and the Global Electrification Platform which hosts OnSSET simulation results [247].

2.4 Gap Analysis

Through this review of the state-of-the art literature in spatial electrification design, including data requirements and methodologies, the following gaps have been identified.

Gap 1 **Existing georeferenced data which locate off-grid populations have accuracy, completeness, resolution, and recency issues.** Census data can be out of date, particularly in the least developed LMICs, and their usefulness in grid design is limited by the often arbitrary and colonial borders used to aggregate them. Raster population datasets are frequently too low in resolution to be used to identify potential connection points for electrification design; when their resolution is adequate, they are more likely to include

modelling-based error. Vector topographic datasets representing dwellings tend to have accuracy issues, lack global coverage, and are not collected at a uniform time causing recency and temporal specificity problems. This data gap is tackled in this thesis in Chapters 3 and 4, which leverage citizen science and computer vision for scalable and accurate home mapping for grid design.

Gap 2 **Traditional survey-based data collection methods are costly, inconvenient, and difficult to scale, particularly in rural areas of LMICs.** In-person site surveys are costly and time-consuming. There can be cultural barriers to entering rural communities in LMICs and power dynamics which must be negotiated. These challenges are present whether collecting spatial data or demand data in off-grid communities. In Chapters 3 and 4, methods are proposed to collect location data for potential connection points entirely remotely, without intrusive, expensive, and lengthy site survey. In Chapter 5, existing large-scale datasets collected in appropriate and well-funded programmes are leveraged to generate spatially specific demand estimates without re-surveying.

Gap 3 **There are minimal training data for building detection in rural areas of LMICs.** Training datasets available for segmentation or detection of buildings in satellite imagery are usually focused on urban settlements, HICs, or both. Coverage in rural LMIC areas is low, which hinders the potential to use computer vision to map off-grid communities. This gap is tackled through the generation and open publication of a training dataset focused on rural LMIC areas, as discussed in Chapter 3 and created through the Power to the People citizen science project.

Gap 4 **High-resolution and accurate demand estimation input data are typically not available in LMICs, particularly in off-grid or newly electrified areas.** Detailed appliance use, behaviour, and occupancy datasets available in HICs are not available in off-grid areas of LMICs. High resolution data on energy use in LMICs, and particularly in newly electrified rural areas,

would best fill this gap. This, however, is beyond the scope of the thesis, as it requires national policy efforts and programmes. Instead, this gap is tackled in Chapter 5 by improving upon existing demand estimation methods by using spatially specific data.

Gap 5 **There are few studies in the literature which account for demand evolution in newly electrified rural areas.** Demand is typically either estimated based on current needs, surveyed aspirations, or a specified tier of access to be achieved. This gap is tackled in the thesis in Chapter 6, which maps demand evolution pathways to plan modular grid expansion and to help select a best-fit affordable design point to prevent stranded infrastructure.

Gap 6 **Scalable open-source models which account for local settlement distribution and diversity using existing empirical data at low computational expense are scarce.** Existing models are typically either focused at the local level, simulating and optimising demand and storage infrastructure given specific local demands, or large-scale, selecting a broad grid type over a raster grid cell. Local models tend to assume that the grid type and load is known, whilst large-scale models do not account for settlement style. Models that endeavour to do both are incredibly complex. This gap is tackled in Chapter 6, which proposes a spatial design framework which leverages existing empirical data to propose feasible grid designs in rural off-grid LMIC areas.

3

Mapping Off-Grid Homes With Citizen Science

Contents

As identified in Chapter 2, two data gaps must be filled to enable spatial feasibility design for rural electrification: location data and demand data for potential connection points. This chapter first addresses the question of location: *Can we accurately map rural off-grid populations for electrical system design?* It is argued that rural off-grid populations can be effectively mapped at home level using citizen science and satellite imagery to enable energy system design.

Location data for potential connection points are required to select an appropriate technology and topology for electricity access, and to size system components (e.g. generation and storage). The precise location, quantity, and density of potential connection points influences whether SHS, mini-grid, grid extension, or other technologies will be able to reach potential consumers at lowest cost. As the

location and quantity of connections are also required to estimate demands, these data also inform infrastructure sizing and thus the capital costs of the system. This ultimately influences connection costs and tariffs experienced by the end-user, and thus the accessibility and affordability of the energy service.

The locations of potential electrical connection points can be identified from population density datasets (i.e. rasters and census data) or vector topographic maps. Population density raster datasets such as the GPW [248] and the HRSL [15] have resolutions ranging from 30" to 1" (i.e. 1 km to 30 m at equator) which can misrepresent communities smaller than grid cell size, or those whose homes are unevenly distributed throughout grid cells. Census data suffer similar resolution limitations to rasters, and are further hindered by the irregular size and shape of political districts. These density-based population data do not capture the locations of individual homes or buildings, and therefore do not identify the precise configuration of potential electrical connection points required for topology design. While vector datasets exist which can offer the precise building locations required for electrical topology design, those available to the public tend to have gaps and inconsistencies. For instance, as discussed in Chapter 2 and shown in Figure 2.2, the OSM [52] has recency, accuracy, and completeness issues. Without a coordinated update frequency or temporal data uniformity, one cannot be sure that buildings captured in the OSM are up-to-date. Additionally, large gaps in building coverage unhelpfully tend to correspond with remote and rural regions with low energy access. There are also subtler data gaps; non-Western housing styles (e.g. thatched-roof homesteads) tend to be underrepresented, and businesses are more consistently recorded than non-commercial dwellings. The issues in available data sources lead to **Gap 1** as identified in Chapter 2.

To illustrate the importance of home-level location data in electricity access design, consider three communities, each with a population of 300 people and a total area of less than 1 km^2: a clustered community far from roads, a community linearly dispersed along a major highway, and homes dotted on plots of agricultural land. An electrical engineer using the GPW dataset to plan electrical systems

for these communities would perceive them identically (i.e. as 300 people in one grid cell). They may be able to specify a general least cost grid type for each using a large-scale system planning tool like OnSSET [233], but they will have no way to design or cost the distribution system in each case. The potential demands of these communities may also vary based on their configuration and context. For instance, the community along a major highway is likely to engage in different kinds of trade than the agricultural village, requiring different productive use loads. Even if the designer consults the HRSL to understand the shape of each community, detailed satellite imagery or vector maps will be needed to site infrastructure based on roads, buildings, and land-use patterns. The precise location of community features, and particularly of buildings, are critical in generating complete context-specific electrification designs.

The gaps in existing home-level location data must therefore be filled to design electricity access systems. Traditional approaches for data collection, such as in-person site surveying, are inconvenient, costly, intrusive, and difficult to scale in remote off-grid regions. This is particularly true given the poor infrastructure connections in rural areas of many LMICs which make off-grid communities time-consuming and costly to reach. However, this is not the only approach available for home mapping: innovative participatory, crowd-sourcing, and computing methods can extract this location data from existing global EO satellite imagery. For instance, citizen science can be applied to integrate enthusiastic non-specialists into the research and mapping process. This creates opportunities for accessible engagement in science and adheres to principles of mutual benefit and open data access [85]. Citizen science has been proven effective in collecting and processing substantial geolocated data at low-cost through projects like iNaturalist [89–91] and Missing Maps [92–94] *inter alios*. Citizens have the knowledge and ability to map diverse spatial features, including potential electrical connection points in satellite imagery, and bring diverse perspectives to the research.

This chapter argues that citizen science and satellite imagery can be used to effectively map rural off-grid homes for energy system design. To test this, a project

called "Power to the People"[1] (PTTP) was executed to engage citizen scientists to locate homes in rural Kenya, Uganda, and Sierra Leone using VHR satellite imagery. This approach brings the advantages of high temporal data specificity based on the capture time of the imagery, and the near-worldwide coverage of archive satellite imagery, which are particularly useful in off-grid rural regions which can otherwise be data-deserts. It leverages enthusiasm amongst the existing global citizen science community while avoiding exacerbating disproportionate time poverty experienced by those in study regions, which would be worsened by requesting their voluntary labour in mapping efforts [95, 96]. Throughout PTTP, thousands of citizen scientists classified hundreds of thousands of images to locate homes, which are post-processed to create a high-consensus home location dataset.

Success criteria for this work are based on benchmarks for speed, accuracy and cost drawn from the literature. First, citizen scientists must produce home annotations on VHR satellite imagery with accuracy $\geq 62\%$ and speed $\geq 0.25\,\mathrm{km}^2/\mathrm{day}$ [249][2]. Second, the citizen science project should map off-grid areas at a cost[3] $< \$465/\mathrm{km}^2$ [69][4]. Finally, the citizen science project also had to adhere to the core principles of citizen science [85], including the principle of mutual benefit to professional researchers and citizen scientists.

This chapter first outlines the data and materials required to map potential connection points in Section 3.1, and the citizen science based methodology in Section 3.2. This includes details on satellite imagery requirements, citizen science workflow, and metrics used to evaluate performance. Section 3.3 discusses the results of the PTTP citizen science experiment and Section 3.4 elaborates upon insights and implications of the mapping performance as it pertains to electrical system design for rural off-grid areas. Finally, key insights from the chapter are discussed in Section 3.5 as they relate to the broader questions of the thesis.

[1] See: `www.zooniverse.org/projects/alycialeonard/power-to-the-people`
[2] On-the-ground mapping speed facilitated by satellite imagery. Assumes one project employee.
[3] All dollar amounts in the thesis are in United States Dollars (USD or $) unless otherwise specified.
[4] Lowest cost achieved in these mapping case studies. This cost assumes that each land parcel is 1.29 hectare, the average parcel size from [249].

3.1 Data and Materials

To apply citizen science in home-level mapping for rural off-grid areas, VHR satellite EO images are required as input data for annotation, and an online citizen science platform must be selected.

3.1.1 Satellite Imagery

Satellite imagery of currently unelectrified regions is used as input data for citizen science annotation. While other data types, such as aerial or drone imagery, could be used, satellite data is more readily available than these alternatives with more comprehensive coverage. Large satellite imagery providers hold vast sets of recent archive imagery with near-global coverage; the same cannot be said for drone or aerial imagery, particularly in poorer areas. As such, to use these data types would require acquiring equipment to undertake primary data collection, which would come at significant expense. Of course, this could provide an excellent income opportunity for a local drone operation business, if such a business exists; however, given cost constraints, this option was outside the scope of this thesis, and satellite imagery was used.

To enable the accurate mapping of potential connection points (i.e. buildings or dwellings) in off-grid areas, a GSD $\leq 1\,\text{m}$ and minimal cloud cover is required such that homes are perceptible and unobscured. Colour images are needed, as colour can be a useful way to distinguish roofs from the ground in some cases (though not all cases, e.g. brown thatched roof homes surrounded by brown dirt). Given these resolution and colour requirements, both panchromatic (i.e. single color band, often visualised in greyscale, higher quality) and multispectral (i.e. multiple colour bands, lower quality) bands must be acquired to produce pansharpened high-quality full-colour images for annotation. As populations fluctuate and migrate over time, recent imagery should be acquired wherever possible such that grid designs produced using resulting home locations align with the current on-the-ground settlement configuration.

3.1.2 Citizen Science Platform

A web-based citizen science interface is used to facilitate satellite imagery annotation. The following criteria are applied when selecting a platform to increase the likelihood of project success and adhere to the principles of citizen science [85]:

- Compatibility with multiple device types and browsers to reach volunteers with differing levels of technology access;
- Large and diverse existing volunteer base which can join the project;
- Track record of successful citizen science projects;
- Dedicated project landing page and link for online promotion;
- Integrated informational pages to explain research aims;
- Two-way communication interface to engage with the community;
- Cloud storage availability to host large imagery datasets;
- Built-in tools for image annotation, or the ability to build these;
- Facilities to export pixel-level image annotations.

Such an interface can be implemented through a number of online platforms such as Zooniverse[5] or CitSci.org[6]. Alternatively, custom in-house development or development through a specialised agency like Spotteron[7] can be pursued.

3.2 Methods

The general methodology for the work in this chapter is illustrated in Figure 3.1. Satellite imagery is pre-processed to facilitate annotation. Then, it is annotated by citizen scientists to locate potential connection points (i.e. buildings or dwellings). These data are evaluated against a gold standard sample dataset.

[5] https://www.zooniverse.org/
[6] https://citsci.org/projects
[7] https://www.spotteron.net/

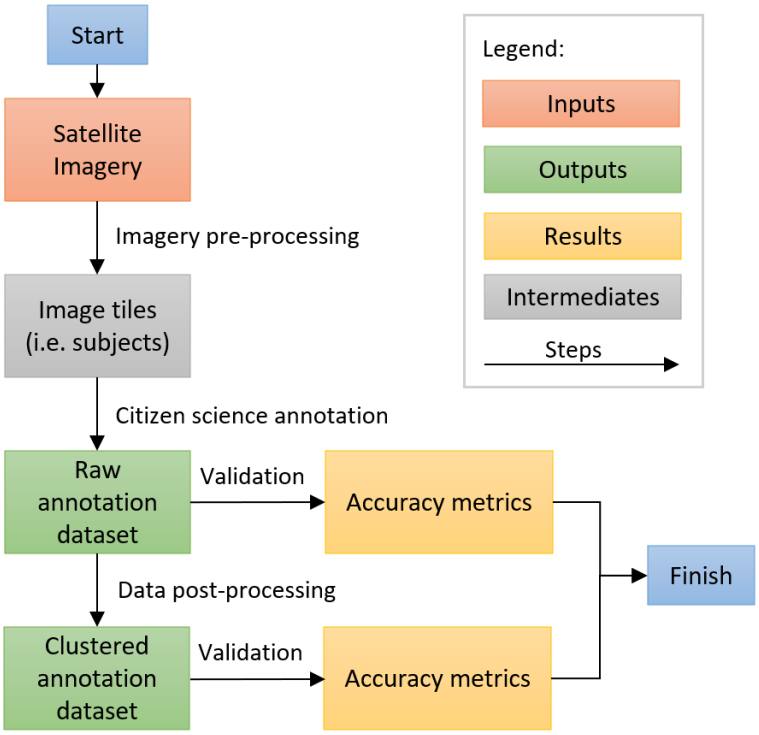


Figure 3.1: Overview of the citizen-science-based mapping methodology for potential connection points in rural off-grid areas.

3.2.1 Satellite Imagery Pre-Processing

Satellite imagery is pre-processed to facilitate online annotation through pansharpening, rendering, tiling, upsampling, and metadata recording.

The original 16-bit geographic tag image file format (GeoTIFF) satellite images are first pansharpened. This process combines the lower-quality multispectral and higher-quality panchromatic images to create a higher-quality colour image. Image histograms are manually colour-corrected to minimize any atmospheric effects and to produce a natural colour appearance for easier interpretation by citizen scientists. Images are then rendered in 8-bit color, reducing their file size at minimal visual expense to enable web upload.

The images are then split into small square tiles for annotation. Three image tiles are generated for each geographic area: one panchromatic tile in greyscale, one multispectral image in colour, and one pansharpened image in colour. These three imagery types are used as each has certain advantages during interpretation.

The panchromatic imagery can make it easier to see edges, while the multispectral imagery has the most accurate colour information, and the pansharpened imagery is a compromise ideally retaining most information from the other two.

Tiles are sized to contain a reasonable quantity of homes for citizen scientists to annotate at any one time and sufficient contextual information to interpret image features. It can be difficult to tell what is included in an overhead image presented at very high zoom, so a compromise must be struck between the risk of making ground features blurry by over-zooming and making them hard to detect by under-zooming. To strike this balance, for imagery with a 0.5 m GSD panchromatic band, 256×256 (i.e. 65,536 pixel or px) panchromatic and pansharpened tiles are generated. Multispectral tiles must be smaller to align geographically with the panchromatic tiles. In this case, for 2 m GSD multispectral imagery, 64×64 (4,096 px) tiles are generated. Tiles on image edges are discarded if too thin to be likely to contain any full homes. Any tiles with cloud cover blocking the view of features on the ground are also discarded. The tiles are then upsampled (i.e. "blown-up") for easier annotation. For the 0.5 m GSD imagery discussed above, panchromatic and pansharpened tiles are upsampled by 200% while the multispectral tiles were upsampled by 800% to obtain 512×512 (262,144 px) tiles for all imagery types[8].

Georeferencing details for each tile – that is, the minimum and maximum geographic coordinates for each image tile – are exported from the GeoTIFF tiles and preserved in a separate comma-separated values (CSV) file to be used in data post-processing. Finally, tiles are converted into a web-ready portable network graphics (PNG) format and uploaded to the citizen science platform.

3.2.2 Citizen Science Workflow

The citizen science annotation workflow is designed to map connection point locations for electrical system planning, with secondary goals of collecting home size and roof material information (e.g. thatch or metal). Roof type and size can be used as a proxy for household income in certain contexts, which is important to

[8]This was found to be easier to label than the original tile size in PTTP beta testing.

know when considering the willingness of inhabitants to pay for electrical service. Additionally, if rooftop PV is desired or is more practical than ground-mounted PV in a particular community context, the size of the roof dictates the capacity of PV which can be installed. These are important parameters to electrical design, particularly in rural developing regions where rooftop PV might be the most economic option, and so are collected during the annotation process.

Given these goals, the annotation workflow is constructed as shown in Figure 3.2. The citizen scientist is presented with a subject, and asked whether they can see any homes[9]. If they indicate that they do not, they are presented with a new subject. If they indicate that they do, they are asked to annotate the home on the image. For each annotation, they are asked to identify the color of the roof, under the assumption that thatched rooftops appear brown and metal rooftops appear light or white. They are also asked to identify roof shape from the options "square or rectangular", "circular or rounded", or "other". When the citizen scientist is finished annotating the image, they are asked whether they have annotated all visible homes, or whether there were too many to annotate them all[10]. Once the citizen scientist answers this question, they can then confirm their annotations and move on to another subject.

Multiple annotation tools can be used to locate homes in satellite imagery, including bounding boxes, circular annotations, point annotations, or free-form polygons. Since an estimation of roof size is desired, point annotations are not appropriate in this application. While free-form polygons can theoretically produce the most accurate data on roof size and shape, these annotations are harder for citizen scientists to create in practice through online interfaces than rectangles or circles; they tend to require higher precision and more steps to create. As the extra

[9]Note that the word "home" was used in PTTP instead of "building". It was hypothesized that the global community of citizen scientists was more likely to understand potentially unfamiliar construction styles (e.g. semi-permanent refugee homes, thatched roof structures) as homes than as buildings, even though they may actually represent businesses, places of worship, etc.

[10]This question was added to PTTP based on urgent feedback from citizen scientists that some images contained too many homes to label (i.e. they were too overwhelming or time-consuming). These images were typically from peri-urban regions with a mix of under-represented home styles and dense areas, or crowded refugee camps.

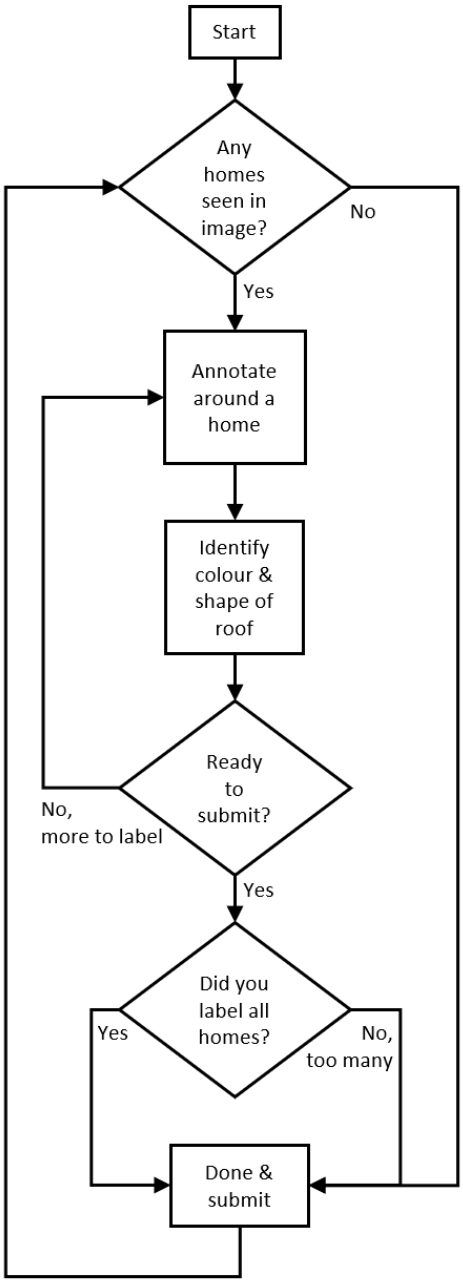


Figure 3.2: Workflow for the online citizen-science-based satellite imagery annotation.

effort and challenge required to produce accurate arbitrary contours might cause frustration amongst citizen scientists, harming data quality and slowing the process of annotation, free-form polygons are not used. Thus, to capture annotations and roof sizes as accurately as possible, a rotatable bounding-box annotation tool is used. This allows citizen scientists to align annotations very accurately with

the rooftops of rectangular buildings to preserve as much detail as possible. An estimate of roof size can still be obtained for other roof shapes using the follow-up questions on roof shape. For instance, if the citizen scientist indicates that the roof is circular, the area of a circle inscribed on the bounding box annotation can be used to estimate the roof area.

Each subject is classified (i.e. viewed and annotated in the above-described workflow) by multiple citizen scientists. If the first five citizen scientists to classify a subject all indicate that there are no visible homes, that subject is retired from the workflow. If any of the first five citizens scientists do see any homes, the subject is classified by ten citizen scientists before retirement. These thresholds are selected based on the previous experience of the Zooniverse team in designing similar task workflows. Given the relative complexity of the rural home annotation task and the diversity of image types in the workflow, cautiously high thresholds are adopted. It is possible that the number of citizen scientists to annotate each image could be reduced without harming data quality. A sensitivity analysis to investigate this would be an interesting area for future work, but is outside the scope of this chapter.

3.2.3 Annotation Post-Processing

Once all subjects are fully annotated, the annotation data are post-processed to identify where contributors agreed that there is a home – that is, where there is a high density of home annotations – using clustering. To achieve this, a density-based clustering algorithm that allows outliers and adapts to variable annotation accuracy and quantity is desired. As citizen scientists may make accidental or incorrect annotations, each annotation should not be forced to be part of a cluster; rather, clusters should only include those annotations in high-density groupings. To this end, the Hierarchical Density-Based Clustering of Applications with Noise (HDBSCAN*) algorithm is applied. HDBSCAN* builds on the Density-Based Clustering of Applications with Noise (DBSCAN) algorithm by producing a hierarchical set of DBSCAN clusters as the ϵ parameter is varied. It optimally "cuts" through the hierarchical tree where the resulting clusters are

most stable and persistent [250]. As the results of HDBSCAN* depend highly on the minimum cluster size (m_{clSize}) hyperparameter selected, a range of m_{clSize} values are applied in clustering experiments, and the best m_{clSize} value is selected based on the clustering results. The accuracy metrics discussed in Section 3.2.4 are used to select the best m_{clSize} value.

The post-processed annotation clusters and original home annotations are subsequently georeferenced for use in mapping and GIS. The annotation image coordinates are mapped to geographic coordinates in each satellite image's original coordinate reference system based on the georeferencing data exported for each image tile during pre-processing.

3.2.4 Data Validation

To validate the data, a set of "gold standard" annotations is generated by the author for comparison with the citizen science annotations. The gold standard annotations are made on a random sample of 188 images which contain HDBSCAN* clusters for $m_{clSize} = 5$ (i.e. which are likely to contain some homes to label). This sample of annotations is used to calculate precision (P), recall (R), and F_1 score (i.e. the harmonic mean of P and R) for citizen science annotations. P, R, and F_1 are defined as:

$$P = \frac{T_P}{T_P + F_P} \tag{3.1}$$

$$R = \frac{T_P}{T_P + F_N} \tag{3.2}$$

$$F_1 = 2\frac{P \times R}{P + R} \tag{3.3}$$

where T_P represents true positives (i.e. positive predictions which agree with ground truth), F_P represents false positives (i.e. positive predictions which disagree with ground truth), and F_N represents false negatives (i.e. negative predictions which disagree with ground truth). Citizen science annotations are treated as predictions and gold standard data are treated as ground truths in these calculations.

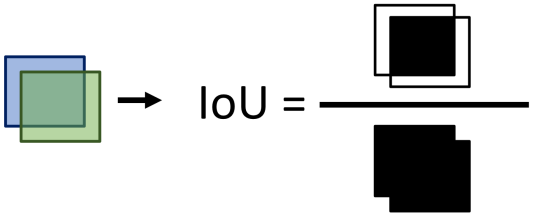


Figure 3.3: Visualisation of the concept of intersection over union (IoU) in the context of satellite imagery annotation, where an annotation is shown in blue and a ground truth is shown in green.

As these image annotations are shapes, to determine whether a ground truth and a prediction are equivalent requires looking at how much they overlap. In practice, a ground truth and a good prediction might not be exactly the same – for instance, the amount of overlap could be 80% or 90% of the union of the area of both shapes. To account for this, a threshold value for the intersection over union (IoU) between a ground truth and a prediction is defined in order to count that prediction as T_P. This IoU value is also used to identify F_N (i.e. ground truths that have an overlap less than IoU with any prediction). IoU can be calculated for a prediction A and ground truth B as:

$$IoU = \frac{A \cap B}{A \cup B} \tag{3.4}$$

This can also be shown graphically as in Figure 3.3. An IoU threshold of 0.5 is used here; this is a commonly used and acceptable threshold. This means that $IoU \geq 0.5$ between a prediction and any ground truth is counted as T_P, $IoU < 0.5$ between a prediction and any ground truth is counted as a F_P, and $IoU < 0.5$ between a ground truth and any prediction is counted as a F_N.

3.2.5 Impact Evaluation

Given the principle of mutual benefit underlying the citizen science methodology, it is important to evaluate not only the quality of the resulting data, but also the quality of the volunteer experience. Specifically, to understand the impacts of citizen science, it is important to evaluate the composition of the citizen science community, their motivations, and any impact that working on the project has had on them.

To this end, an evaluation survey is conducted to evaluate the citizen science experience and impact. A survey is disseminated to gather data about the composition, motivations, and experiences of the citizen science community, centered around the following lines of inquiry:

- Who is contributing to PTTP?
- Why do they choose to engage?
- What are the impacts of engagement for contributors?
- What are the benefits and challenges of contributing?

Here, this survey is promoted via the online citizen science interface and other digital channels to collect insights from existing project contributors. It is disseminated towards the end of the project when contributors have had ample time to experience the project and learn from it. A full list of survey questions is included in Appendix A.

3.3 Results

These citizen science mapping methods are applied in the experimental PTTP citizen science project, whose outputs are then evaluated to determine methodological performance. First, the experimental setup and execution is detailed; then, the resulting data, both in terms of imagery annotation and contributor experience, is evaluated.

3.3.1 Experimental Setup

Satellite imagery used for PTTP were captured by the Superview-1 constellation,the Disaster Monitoring Constellation 3 (DMC-3), and the Korea Multi-Purpose Satellite (KOMPSAT) 3A. Details on each these satellite constellations are presented in Table 3.1. Each has a GSD $\leq 1\,\mathrm{m}$ in the panchromatic band. The most recent available images were used to ensure that the resulting home maps were as up-to-date as possible. These images had capture dates ranging from 2015 to 2019. As acquiring new imagery is more expensive than using archive imagery, archive images were used where possible to minimize costs. New imagery was acquired where no suitable imagery was available with a capture date in 2015 or later.

Table 3.1: Satellite constellations which acquired imagery for the citizen science project, alongside their operators, launch dates, satellite quantities, and ground sample distances (GSD). GSD_P is for panchromatic imagery while GSD_M is for multispectral imagery.

Name	**Operator**	**Launch date(s)**	**# of satellites**	**GSD_P (m)**	**GSD_M (m)**
Superview-1	Beijing Space View Technology	2016-12-28 2018-01-09	4	0.5	2
DMC-3	DMC International Imaging	2015-07-10	3	1	4
KOMPSAT-3A	Korea Aerospace Research Institute	2015-03-25	1	0.55	2.2

It was selected to focus PTTP's mapping effort in Kenya, Sierra Leone, and Uganda. These countries have ongoing rural electrification efforts at different stages, have home location data gaps in OSM, and contain diverse, underrepresented rural home styles which require more accurate mapping (e.g. agricultural, clustered community, refugee, etc.). Specific satellite image sample locations were chosen based on visual inspection and prior knowledge of rural settled areas with underrepresented housing styles. These included thatched roofs, or communities containing both thatched and corrugated roofs; homesteads constituted by multiple structures; homes far from roads or surrounded by vegetation; and homes in refugee settings, where settlements do not follow roads and contain both temporary and permanent structures.

The Zooniverse online citizen science platform was selected to host PTTP, as it meets the requirements outlined in Section 3.1.2. It is free to use and allows rapid and easy project construction, which was critical here, as no budget was available for extended web development. The Zooniverse is the largest and most popular platform for online citizen science, with a contributor base of over 2 million contributors [251]. It offers dedicated project landing pages, built-in project information pages, and a built-in project "Talk" forum for discussion amongst volunteers and researchers. Space is available for up to 10,000 imagery subjects per project, with more space available upon request. The Zooniverse platform has built-in image annotation and data export tools.

Satellite imagery was pre-processed as described in Section 3.2.1. Image tiles were divided into "subjects" for upload to the Zooniverse platform, each of which

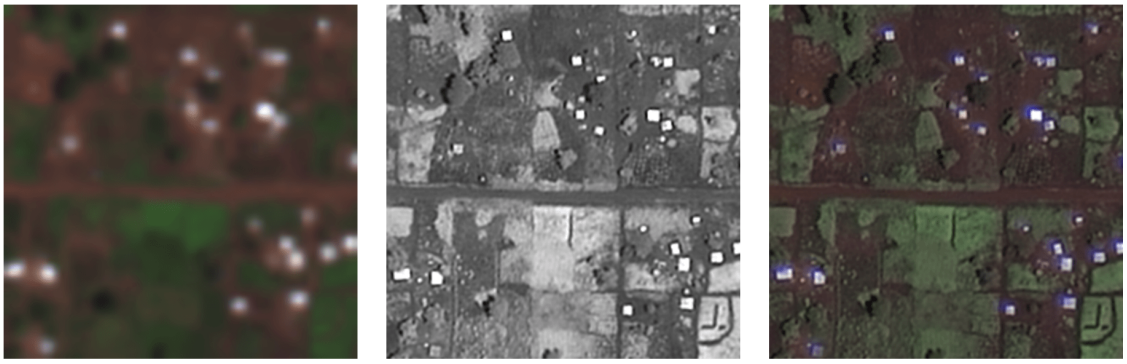


Figure 3.4: Examples of the three satellite imagery types provided to citizen scientists for annotation. Left: Multispectral imagery visualised in natural colour. Center: Panchromatic imagery visualised in greyscale. Right: Pansharpened imagery visualised in natural colour.

included a pansharpened, panchromatic, and multispectral tile of the same area that the citizen scientist could cycle through as they annotated the area. This allowed citizen scientists to view three image types for the area in each tile, which can help them during annotation, as certain information and features are easier to distinguish in each image type. For instance, colour information is easier to see in multi-spectral imagery, while panchromatic imagery can make it easier to perceive edges. An example of the three imagery types is shown in Figure 3.4.

The citizen science workflow described in Section 3.2.2 was implemented in the Zooniverse platform as shown in Figure 3.5. It was tested internally by researchers, and beta tested by citizen scientists, before proceeding to full launch.

3.3.2 Project Execution

PTTP launched on 2nd March and ran until 28th August 2020. A total of 519,420 image classifications including 578,010 home annotations were made on 74,802 subjects throughout project execution. Cumulative classifications over the project duration are shown in Figure 3.6.

Engagement on the project spiked at a number of points during execution based on world events and promotion. First, this project coincidentally started just before the escalation of the Covid-19 pandemic which prompted lockdowns worldwide. Engagement spiked as lockdowns increased in mid-March through early April. The

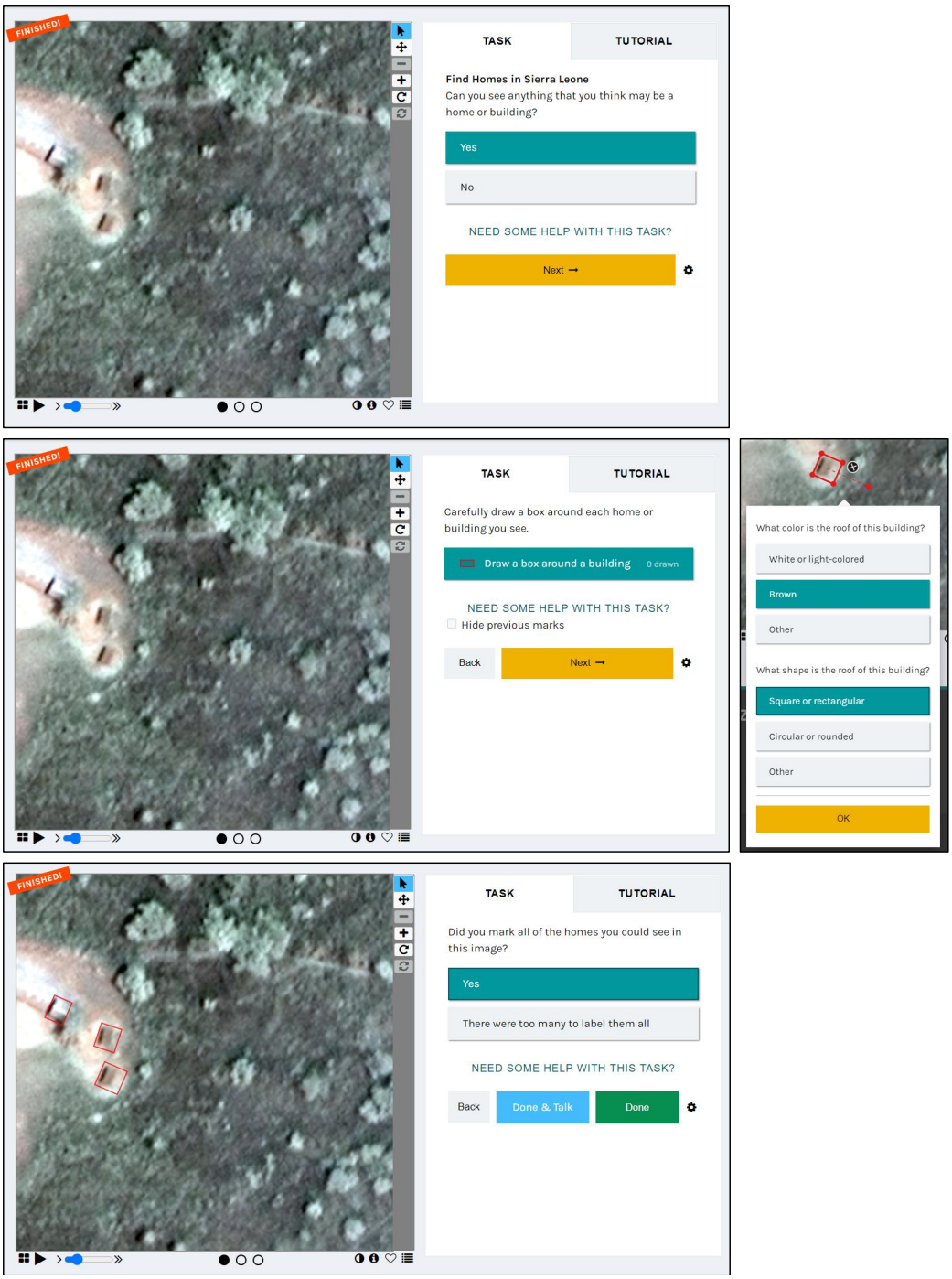


Figure 3.5: Home annotation interface used in the Power to the People citizen science project. Top: The citizen scientist is asked whether they can see any homes. Middle: The citizen scientist is asked to annotate homes (left) and answer follow-up questions for each home annotated (right). Bottom: The citizen scientist is asked to confirm their annotations before proceeding to another image. Please ignore the "finished" banner in the top-left corner of each image.

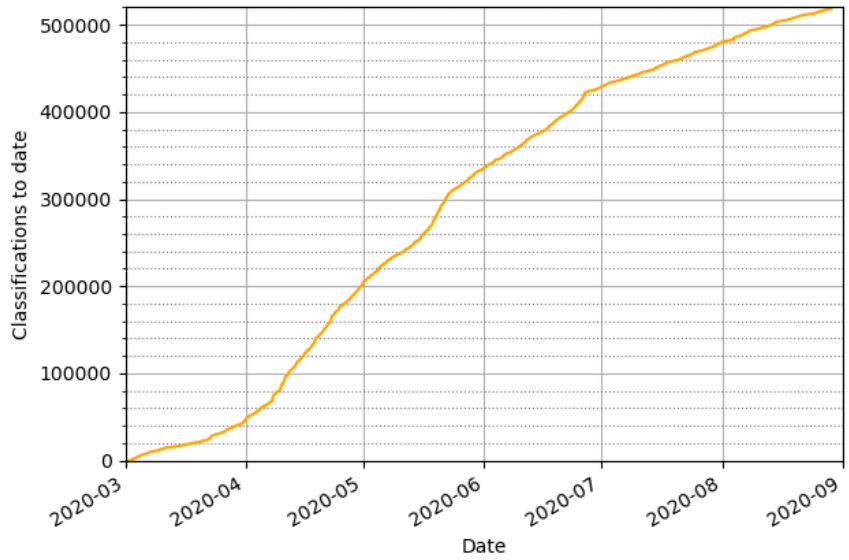


Figure 3.6: Cumulative classifications on satellite imagery subjects over the course of the Power to the People citizen science project.

Zooniverse platform as a whole saw increased use during early lockdowns [252], and this project was no exception. Another spike occurred on Easter weekend (i.e. around 12th April 2020). While there is no way to confirm this, it is likely that with extra time off work and children off school during Easter, contributors were again looking for a way to pass the time. The project was also promoted by various institutions (e.g. Northwestern University [253]) and at virtual events (e.g. the International Science and Engineering Fair [254]), which encouraged engagement as a way to volunteer time towards science and sustainability initiatives.

Throughout the project, 71,251 Superview subjects, 2,915 KOMPSAT 3A subjects, and 636 DMC-3 subjects were mapped, totalling approximately 1,267 km^2 of imagery. Examples of citizen science annotations in various imagery contexts are shown in Figure 3.7. The average imagery retirement speed of satellite imagery tiles was 7.1 km^2/ day. While the Zooniverse platform does not log time spent by citizen scientists on-task, they estimate that for similar annotation tasks on their platform, between 20-75 classifications can be completed per hour [255]. Using an estimate of two minutes per classification, approximately 17,314 hours of volunteer time were contributed throughout the project, or 13.7 volunteer hours per km^2 mapped.

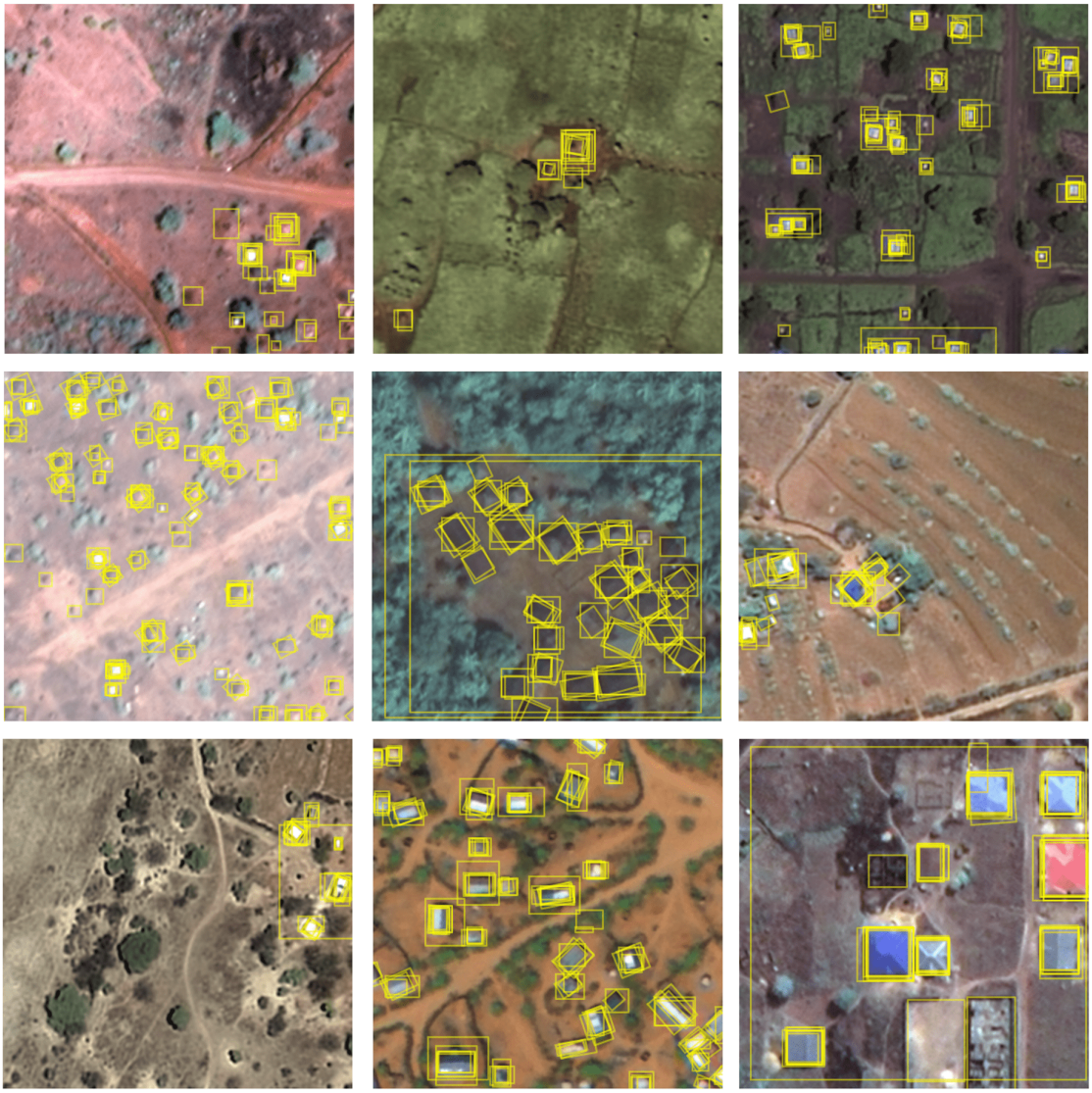


Figure 3.7: Raw citizen science annotations from the Power to the People citizen science project shown as yellow boxes in a number of different geographic contexts.

Following project completion, a record of all annotations was exported from the Zooniverse platform as a CSV with one row per classification containing nested JavaScript object notation (JSON) describing individual annotations. Any classifications made when the project was not "live" (i.e. during platform development, beta testing, or erroneously after project termination) were discarded. The data were then post-processed for validation.

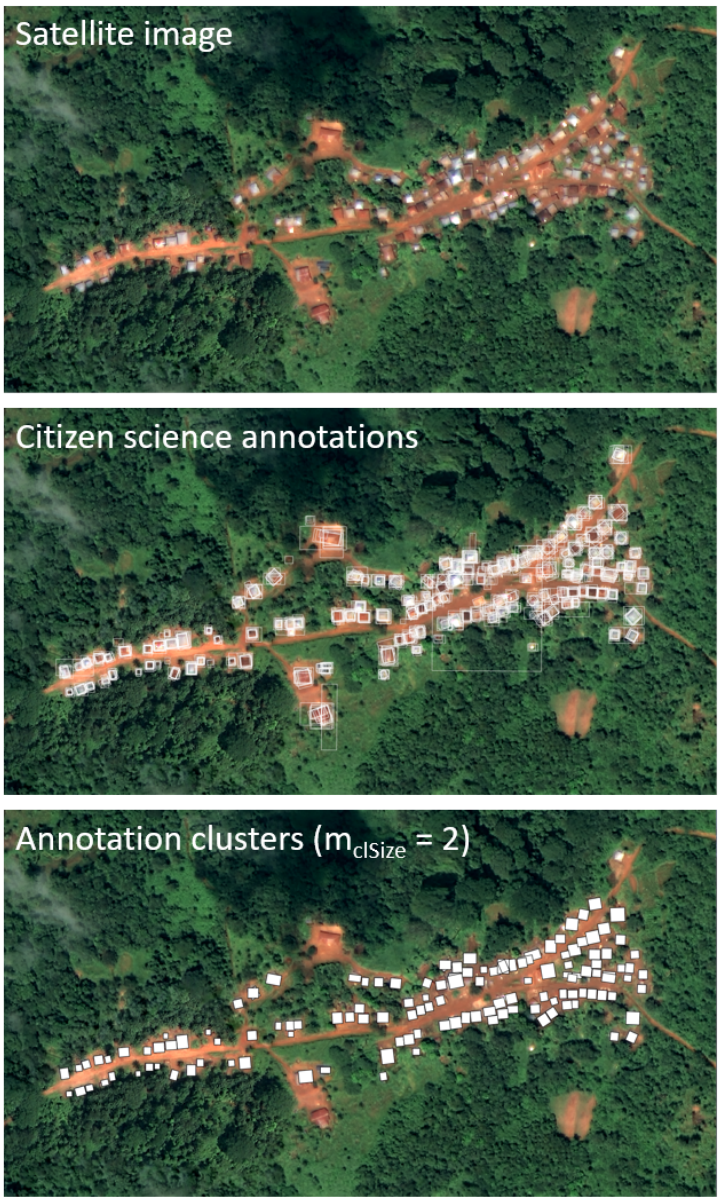


Figure 3.8: The progression from satellite imagery input data through to citizen science annotation and post-processing to identify home footprints. This is illustrated for a sample community area in Sierra Leone (9°32'47.823", -12°10'9.611"). Top: Satellite imagery used as input data. Middle: Raw home annotations produced through the Power to the People citizen science project. Bottom: Clustered annotations.

3.3.3 Data Validation

To understand the accuracy of the generated data, P, R, and F_1 are calculated using raw annotations as predictions and gold standard data as ground truth, as outlined in Section 3.2.4. The resulting values are $P = 0.492$, $R = 0.933$, and $F_1 = 0.644$. The high R value here may indicate that citizen scientists found

most ground truths, but is also influenced by the fact that the raw annotations contain as many as ten sets of annotations per image, and multiple overlapping raw annotations each count as a distinct T_P in this case. The P value of 0.492 indicates that, of all annotations made, 50.8% were F_P – that is, about half of all annotations made did not match an annotation in the gold standard annotation set. This validates the assumption that some method is needed to help find consensus amongst the raw annotations and eliminate the noise.

The data are clustered as outlined in Section 3.2.3 to group overlapping annotations into a singular home footprint wherever possible. An example of the resulting clustered footprints is shown for a case study community in Figure 3.8. P, R, and F_1 accuracy metrics are calculated for annotation clusters produced by HDBSCAN* with different m_{clSize} values to identify the m_{clSize} value which maximised accuracy, as shown in Figure 3.9. The highest precision ($P = 0.689$) is seen for $m_{clSize} = 4$ while the highest recall ($R = 0.487$) and F1 score ($F_1 = 0.568$) are seen at $m_{clSize} = 2$. Clustering improves P compared to the results for raw annotations, but the R for clusters is consistently lower than the R for raw annotations. Again, this is likely due at least in part to the inflation of the T_P count stemming from overlapping annotations.

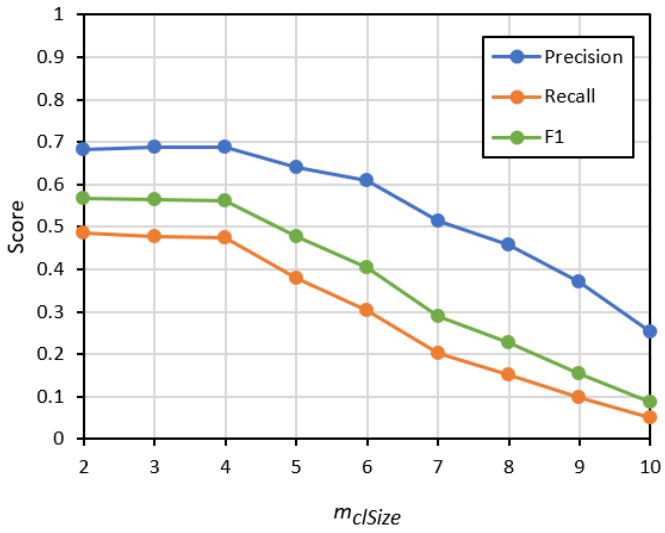


Figure 3.9: Accuracy results for home annotation clusters from the Power to the People citizen science project. Precision, recall, and F1 results are plotted for the clusters generated using HDBSCAN* with m_{clSize} varied between two and ten.

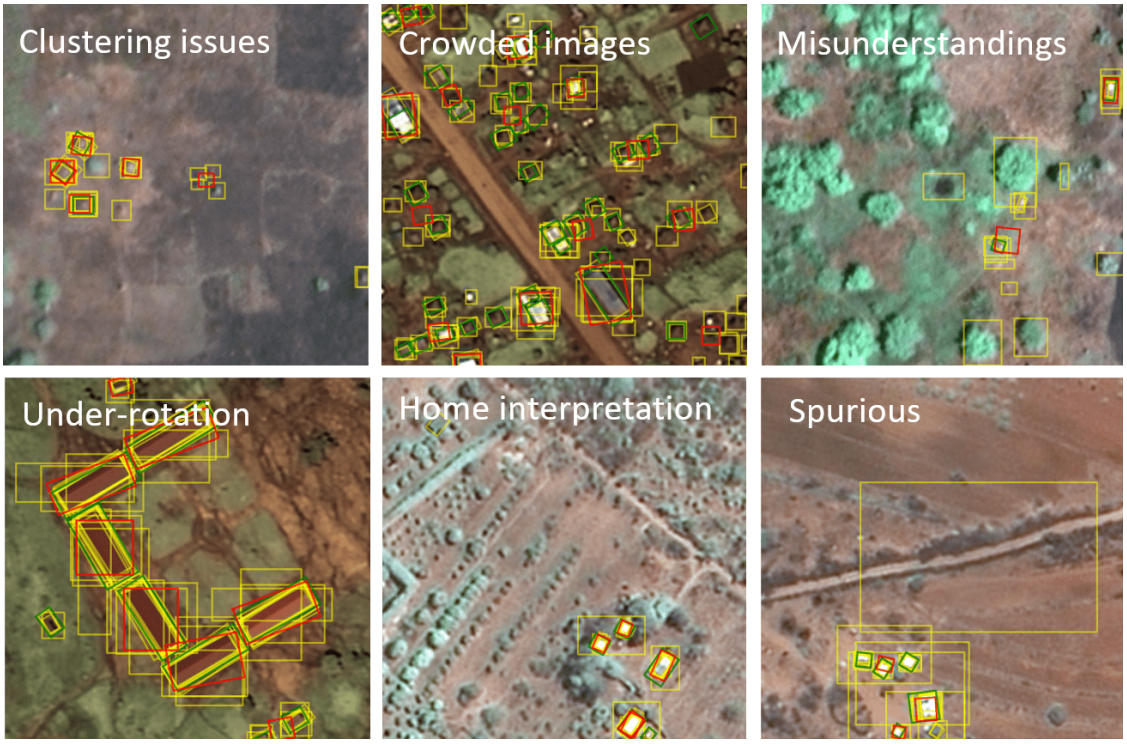


Figure 3.10: Trends observed in satellite image annotations on the Power to the People citizen science project. These include: clustering issues, difficulties with crowded images and under-rotation, different home interpretations, misunderstandings, and spurious annotations. Citizen science annotations are shown in yellow, clusters are shown in red, and gold standard data are shown in green.

A sample of the data are visualised for inspection and interpretation to give a more intuitive understanding of annotation and cluster accuracy. Raw citizen science annotations, clusters for HDBSCAN* with $m_{clSize} = 2$, and gold standard annotations are visualised over pansharpened image tiles. A number of interesting observations arise from comparing the raw data, clusters, and gold standard data, each of which is visualized in Figure 3.10. The following observations appear to have negative impacts on data quality:

- **Clustering issues**: HDBSCAN* occasionally "splits the difference" between annotations, or groups of annotations, which are close together yet too sparse to form distinct clusters. This leads to erroneous cluster locations.

- **Crowding**: Annotation quality is highly variable on crowded images. In these cases, citizen scientists often miss homes. This is unsurprising, given

their communicated inability to complete annotation on many crowded images. Crowded images with rectangular high-contrast homes generally have higher-quality annotations than those with irregular lower-contrast homes. Furthermore, HDBSCAN* does not tolerate the variable density and completeness of annotation in these images well, and the clustering behaviour is somewhat erratic.

- **Rotation**: Some citizen scientists elected not to rotate their annotations to align with roof edges, despite having a tool to do so and being instructed to do so if possible. This makes little difference to the clustering results for homes with circular roofs (as no rotation of a square can best align with a circle) or for homes whose roof edges are already aligned in the image tile with the initial rotation of the annotation tool (i.e. orthogonal to the edges of the image tile) as no rotation was needed in this case anyway. However, for rectangular homes whose roof edges are not orthogonal to the edges of the tile, the lack of rotation on annotations causes subsequent clusters to be under-rotated. This did not pose a problem for smaller square buildings, as their general roof area stays the same and the cluster will be generally placed correctly if at the wrong angle, but for larger and longer rectangular buildings, it can have quite a significant impact.

Meanwhile, the following observations do not appear to dramatically or negatively affect data quality, but are still interesting to note:

- **Home interpretation**: At times, citizen scientists differently interpreted the instruction to "carefully draw a box around each *home* or building you see". In images with groups of small buildings, many citizen scientists annotated each structure as a home, while some interpreted the whole group as one home. During clustering, HDBSCAN* tends to place clusters at the locations with the highest density of annotations and disregards other interpretations. This leads to the clustered dataset generally agreeing with the interpretation of each building as a home. However, this interpretation can be debated; in

many contexts, particularly in rural areas, multiple buildings can constitute one homestead. For electrification design, locating each building is useful, even though each building may not be an individual home, as residents may desire electrical access in each sub-building of their homestead.

- **Misunderstandings**: In certain images, citizen scientists seem to have misunderstood non-home objects (i.e. fields, rectangular yards) to be homes. This is somewhat expected, as there are diverse home styles and settlement types in the input satellite data. It can be challenging to keep track of what a home actually looks like in each particular context when dealing with such diversity. In most instances, misunderstandings are the minority of annotations, and HDBSCAN* eliminated these when clustering.

- **Spurious annotations**: Some images contain annotations which appear to be completely incorrect or accidental. These differ from misunderstandings, in that there are no obvious features which could be misunderstood homes where the annotations are placed. Again, these are a small minority of annotations, and HDBSCAN* typically eliminates them from clusters.

3.3.4 Impact Evaluation

To evaluate the impact of the citizen science project on volunteer contributors, an impact evaluation survey was completed as detailed in Section 3.2.5. The survey was open for responses from 14th August to 14th September 2020 (i.e. approximately two weeks prior to and two weeks after project completion). The survey was advertised via an email to the contributor mailing list, a banner notice on the project homepage, and a post on the project discussion forum. It received 142 responses, which is approximately a 2% completion rate from all contributors. Anonymized quotes from survey responses are provided in italicized block quotes throughout these results as appropriate to illustrate and evidence the findings.

The results of the evaluation survey indicate that this project succeeded in attracting a diverse group of citizen scientists. Survey respondents represent six

continents (Asia, Europe, North America, South America, Africa, and Australia) and 25 countries. As several countries are present in the survey results in small proportions, it seems likely that there would be even more countries represented amongst the full citizen base who contributed to the work in similarly small proportions, but given the quantity of survey responses they are not represented. Web analytics on the PTTP homepage collected by the Zooniverse during the project support this, which found web traffic to the project from 54 countries, including 36 which are not represented in the evaluation survey results. There are high proportions of evaluation survey respondents from the United Kingdom and the United States of America (44% and 27% respectively), which is expected, as the Zooniverse platform has a high pre-existing contributor base in these countries.

Based on survey results, the citizen scientists who contributed to PTTP do not at all fall into the stereotypical bucket of characteristics often attributed to interest in engineering (i.e. middle-aged, well-educated men). In fact, there were more women (59%) than men (36%) represented in survey respondents: the remaining 5% of respondents either identified as non-binary, preferred to self-describe, or preferred not to answer (see Figure 3.11). Over half (54%) of the respondents had no background in science or engineering, and their employment status' varied: 31% were students, 30% were employed full-time, and 19% were retired. The remainder were either employed part-time, in some other employment status (e.g. on disability leave), or preferred not to answer. While 56% of respondents had achieved some higher education degree (i.e. Bachelor's, Master's, or PhD), the remainder were all over the educational spectrum, from currently in school (11%) to having completed high school (20%) or done vocational training (6%). Respondents also varied significantly in age, as shown in Figure 3.12. Evidently, this project succeeded in reaching people who are not the audience typically associated with engineering work.

Evaluation survey respondents reported an overwhelmingly positive experience contributing to the project. 87% qualified their experience as either "good" or "excellent" as shown in Figure 3.13. The most popular reasons reported for engaging in this citizen science effort were a desire to contribute to projects with real-world

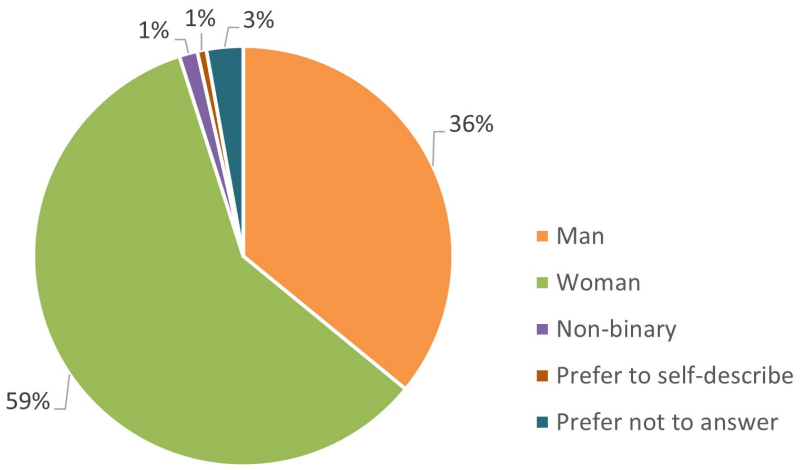


Figure 3.11: Genders of the Power to the People citizen science evaluation survey respondents. Note the higher proportion of women than men represented in the survey.

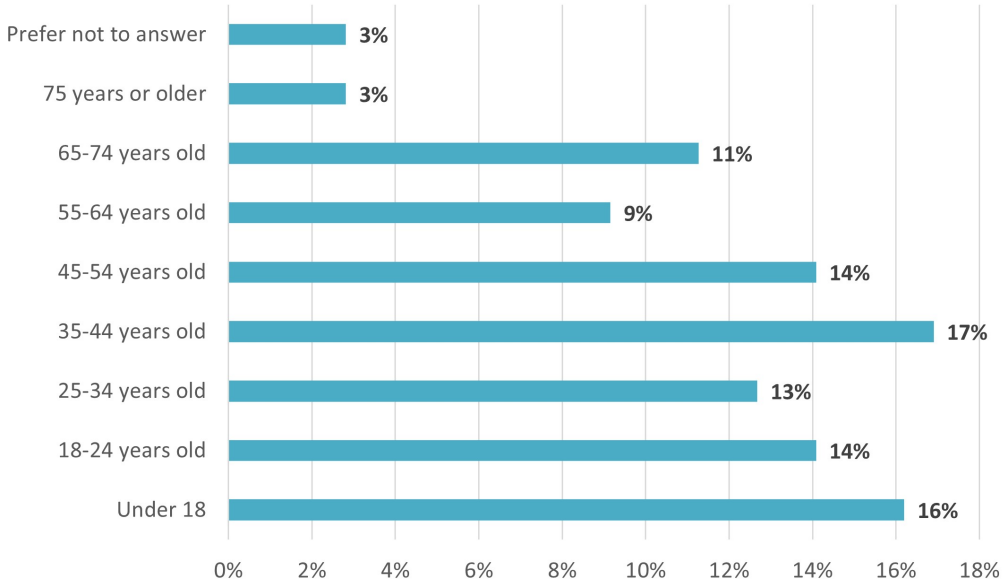


Figure 3.12: Ages of the Power to the People citizen science evaluation survey respondents.

impact (91% reporting) and a desire to contribute to scientific research (79% reporting). When asked to explain any parts of the project they found particularly enjoyable, respondents highlighted the potential for project impact, their sense of discovery (e.g. finding interesting buildings, artifacts, and geographical features), learning about new places, the convenience of volunteering on their own time in their own home, and connecting with researchers (as expressed in the quote below).

> *"I hugely appreciate the time and effort you took to participate in the talk boards with the volunteers. Believe it or not, but that is the thing that will make or break the project."*

Several respondents also indicated that contributing to this work was a great use of their time during the COVID-19 lockdowns, as indicated in the following quotes. This is supported by the spike in engagement seen as lockdowns escalated worldwide.

> *"It has been really great to contribute to this project particularly over the corona-virus lockdown period."*

> *"Has been a useful distraction during the lockdown imposed by Covid, and has been good to know that I have helped in some small way."*

> *"Power To The People has been an easy platform for desktop volunteering for the construction industry teams I work with. Covid has prevented our other community investment activities within the UK in local to our sites, but now that 2 community development projects in Uganda have been cancelled (UK personnel being unable to travel to Africa) it's meant we can still make a contribution while waiting to reorganise the projects."*

In general comments at the end of the survey, some respondents engaged with the ethical and scientific concepts of the research in depth, as evident in the following quotes:

> *"I did have a nagging feeling throughout the time I worked on Power to the People. Sometimes people like the way they live even if it seems backward or difficult to us. Like, who are we to say that our way of life is the right way so we 'force' our way on other people of vastly different cultures. Are we really enriching their lives or just changing it when they're perfectly content living the way they do. I realize that it*

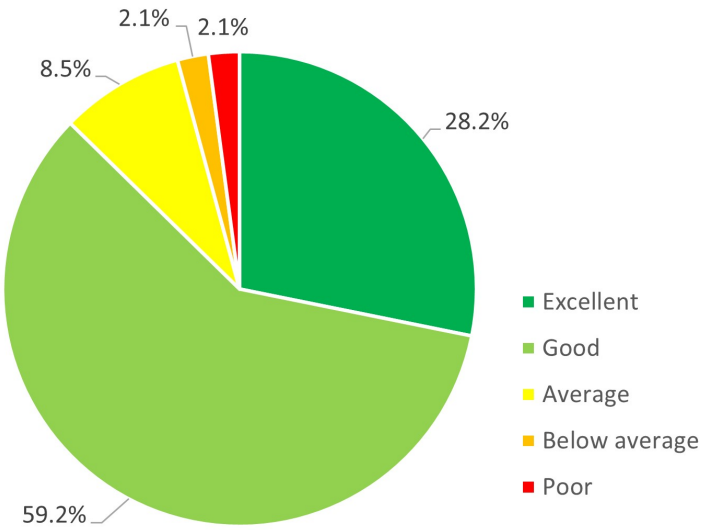


Figure 3.13: Evaluation survey responses for contributor experience on the Power to the People citizen science project. 87% reported that their experience was either "Good" or "Excellent".

> *would be a significant improvement if there was refrigeration for vaccines and things like that. I just felt conflicted while I was working on your project."*

> *"I found it concerning developing more infrastructure around the world as most of the planet is already wrecked on a daily basis. This is very concerning. It is also critical this does not lead to increased demand for fossil fuels or radioactive waste. Although I do believe these people have a right to electricity we should keep a low worldwide population."*

To think about the ethical concerns of this work, in terms of impacts on the cultures and way of life of rural communities as well as the possible environmental damage of additional electrification, indicates a deep, engaged participation. This speaks to the power of citizen science as an engagement tool.

In-project learning was a key impact that this survey sought to evaluate. Results on this are positive: nearly two-thirds (66%) of respondents reported learning something through their experience on this project. Many reported learning about

the geography and settlement patterns in the studied countries through their observation of satellite imagery. For instance, when asked what they learned, responses included:

> *"How to 'read' satellite imagery; something about conditions on the ground in rural Kenya"*
>
> *"I learnt more about the layout and construction of housing in the Uganda."*
>
> *"Housing patterns in rural S. L. [sic]"*
>
> *"The communities in Africa are very numerous and they seem to have great agricultural skills."*
>
> *"I learned more about the landscape of Africa buy [sic] looking at the photos and seeing trees and buildings. It gave me a better understanding of the landscape."*

Others reported learning about the status of electricity access worldwide:

> *"I learned that even in today's world there are still many people living without things that we consider basic necessities like electricity."*
>
> *"That access to electricity is not something to be taken for granted"*
>
> *"I learned that access to reliable and sustainable electricity was part of the UN SDGs and how important it is for livelihoods in Africa to have it."*
>
> *"It hit home what we take for granted. And, it wasn't as easy as I thought to locate the homes."*
>
> *"That there are SO many rural houses that may not have any power, it's sad that the governments in these countries aren't doing more."*

Furthermore, and excitingly, participating in this project also prompted 17.6% of respondents to do their own investigations into concepts relevant to the project. Most of those who sought further information did so via internet search, and the most popular research topic was information about the countries of Kenya, Sierra Leone, and Uganda. This speaks to the broader potential of citizen science to raise awareness and to mobilize curiosity amongst citizen scientists.

3.4 Discussion

These results show citizen science to be a viable method for home-level mapping of rural off-grid areas. The PTTP project was completed successfully and produced high-quality data, albeit with some issues (e.g. crowded images, under-rotated annotations) which require further work. Some lessons identified through the challenges faced in this work are detailed in the following sections, with a focus on how this approach can support better energy access system design in the future.

3.4.1 Annotation Performance

As discussed in Section 3.3.3, the quality of citizen science annotations was reduced in certain circumstances. Three particular cases have been identified where quality could be improved by adjusting the experimental design: addressing image crowding, adjusting expectations regarding rotation, and by improving image clarity through heightened contrast.

Crowded Images

Citizen scientists struggled to annotate images with many homes. As one of the off-grid settlement types mapped in this work was the often-crowded refugee camp context, this difficulty is problematic. Poor mapping of such under-resourced crowded areas could engender poor electrical system design, perpetuating existing inequities. It appears that it was not the relative size of homes in crowded images that contributors found to be challenging, as non-crowded images had homes of

similar sizes, but the quantity of homes per tile. Crowded images seemed to make contributors less likely to annotate all homes in each tile.

In retrospect, the response to the prompt on crowded images in the annotation workflow (i.e. "Did you label every home you have seen in the image?") could have been used to reduce the impact of these struggles on final data quality. For instance, for every contributor who indicated that an image contained too many homes to label, another contributor could be added to the total count of contributors required to label the image. This would help ensure that there would be abundant home annotations. This is an easy improvement for similar projects in the future.

Crowded images could also be presented at a higher zoom level in future projects to reduce the quantity of homes per tile. Gridded population density datasets (e.g. HRSL, GPW) could be used to identify higher-density areas needing higher zoom levels. Depending on the distribution of homes in each grid cell (i.e. the settlement style of the population there), this could result in some empty or sparse areas being presented at higher zoom than needed, but this would be minimally harmful aside from slowing the annotation process. Crowded images could also be excluded from citizen science altogether based on some density threshold. Identifying the density cut-off for higher zoom or exclusion which maximises resulting data accuracy could be an interesting question for future research.

Another option would be to alter the citizen science architecture to divert crowded images out of the pipeline or to iteratively improve crowded images. For instance, a multi-stage citizen science pipeline could allow citizens themselves to identify crowded images requiring increased zoom in a preliminary classification workflow (e.g. through a quantitative approximation of the number of homes contained in the image, or with a qualitative label of "empty", "sparse", or "crowded"). Crowded images could then be reprocessed at a higher zoom level before annotation. Each of these potential mitigation strategies can be applied not only to crowded images in home mapping projects, but to other citizen science projects which involve annotating crowded images in ecology, astronomy, or other fields.

Under-Rotation

Not all citizen scientists rotated their annotations to align with roof edges. This reduces the accuracy of the annotations and leads to inconsistent information about roof size (i.e. home footprint), reducing its viability as a proxy for electrical demand. Contributors may have found rotating annotations to be difficult or time consuming, encountered issues with the rotation interface, or been influenced by their experience with previous projects on the Zooniverse platform with non-rotating bounding box annotation tools. Further follow-up would be required to determine the reason. Regardless, this seems to validate the initial assumption that more difficult annotation tools (i.e. with more steps) may be used less frequently, completely, or accurately, leading to more variable data quality. Given the inconsistency of rotation, perhaps a simple box (i.e. no rotation) or circular annotation tool may create more consistent data in future projects.

Contrast

Finally, images could be pre-processed for higher contrast to help improve annotation accuracy. In this work, images were pre-processed to achieve a "natural" colour profile, which varied depending on the specific satellite, setting, and atmospheric conditions at time of capture. Despite this, on the project "Talk" board, citizen scientists often lamented the image clarity. This could be due to the inherent properties of the image, such as the GSD achieved by the satellites. However, even if resolution cannot be enhanced, pre-processing for high contrast instead of natural colour could help to improve the perceived image clarity. This could in turn improve the accuracy of annotations, particularly in blurry or overexposed images, so long as it does not make homes unrecognizable. This may create a trade-off between the accuracy of locations and roof colour information, which would be acceptable here as location was the priority.

3.4.2 Clustering Performance

There is room for improvement in annotation clustering methods for home mapping. While theoretically well-suited to the task, HDBSCAN* had some issues, such as the tendency to "split the difference" discussed in Section 3.3.3, leading to imperfect results. Of particular note is the low recall (R) achieved by the clusters. This is somewhat expected; mapping diverse home types in rural areas is inherently challenging, causing inconsistencies in annotations. These inconsistencies can translate into cluster misidentification or misplacement. Low R is a known challenge in home mapping: OSM data also has low R [256]. It is likely that low R was caused by contributors not managing to label every home, particularly in crowded images, which is then exacerbated by the clustering algorithm behaviour. Future work may wish to explore other clustering methods, or to create custom variations on HDBSCAN* particularly tailored to home mapping and to the quality of data which can be achieved through citizen science. For instance, it could be explored whether clustering parameters could be tuned to suit each context (e.g. each region, settlement type, or image tile) to produce more accurate clusters.

3.4.3 Cost and Speed

To understand the scalability of this approach, its speed (km^2/day) and cost ($/$km^2$) of mapping are evaluated using the example of the PTTP project. PTTP achieved an average mapping speed of 7 km^2/day, which improves on the benchmark by a factor of 28. As the Zooniverse platform is free to use, the main material cost of this approach is satellite imagery. The highest resolution imagery used in this work is commercially available at $10/$km^2$ [257]. Other costs include researcher or practitioner time and computing facilities, though it is likely that practitioners and researchers already have access to the facilities required to utilize the Zooniverse platform and pre-process data. Given this speed improvement and the low costs of this approach, citizen science seems viable as a competitor to commercial mapping alternatives. However, as will be subsequently discussed, this viability relies on the availability and interest of volunteer contributors.

The costs of this approach can be illustrated using the example of replicating this project[11]. Assuming a cost of \$10/km^2 for satellite imagery, the total imagery cost would be \$12,670. Based on the experience of PTTP, one researcher working half-time on the project should be able to replicate the project at the same precision and mapping speed as achieved here. As such, to hire this researcher over the project duration (26 weeks) at a pay rate of \$28/hour [258] assuming a 37.5 hour work week would cost \$13,650. Therefore, this project could be replicated with a total cost of \$26,410, or \$20.84/km^2. These quantitative results are summarized in Table 3.2.

Table 3.2: Summary of citizen-science-based mapping results compared to benchmarks from the literature[12]. Costs are estimated in Section 3.4.3.

Metric	**Benchmark**	**Citizen science (annotations)**	**Citizen science (clusters)**
Precision (%)	100%	49%	69%
Recall (%)	62%	93%	49%
Speed (km^2/day)	0.25	7	7
Cost (\$/km^2)	\$465	\$20.84	\$20.84

Of course, it must be considered that, despite the validated mutual benefit of this approach to citizen scientists and professional researchers, volunteer time is essentially free labour. In this project, an estimated 13.7 volunteer hours were required per km^2 mapped. If volunteers were instead paid the United States minimum wage of \$7.25/hour, this would add an additional cost of \$99/km^2, which would outweigh other costs. For instance, in the case study of replicating this project, this would add an additional cost of \$125,844, roughly five times the total original approximated project cost. Volunteers are critical to the financial feasibility of this approach. It can be debated whether, ethically, these volunteers ought to be paid; however, given their existing desire to contribute without payment, as detailed in Section 3.3.4, citizen scientists can be taken at their word so to speak and assumed to be happy to contribute without pay for reasons of personal satisfaction and interest.

[11]Costs provided in this example are representative, not exact.
[12]It is assumed that the benchmark would achieve 100% precision (i.e. no F_P).

3.4.4 Implications

Home location data is required as an input to produce detailed grid designs including topology and costing. Consider, for instance the REM [238], an innovative model at the intersection of large-area energy access planning and local topology design. Home locations are required as an input to this approach, and computer vision is used to generate home location data for REM where missing from available sources like the OSM. To train their home mapping algorithm, the REM team used location data hand-annotated by research assistants, as even data annotated by Amazon Mechanical Turk workers were found to be of an inadequate quality [259]. Meanwhile, this chapter demonstrates that citizen science can produce home location data which can be directly used as input to models like REM quickly, at scale, and at far less expense. The higher accuracy of citizen scientists compared to Mechanical Turk workers may be attributed to their familiarity with the annotation interface or the task, their interest in the research question, their enthusiasm, or their emotional investment in ensuring good research outcomes; further study is needed to investigate this discrepancy.

This work also has implications beyond academia. Many governments and utilities worldwide undertake extensive home-level mapping when planning energy access initiatives. Such a project was, for instance, recently undertaken in Kenya for a geospatial electrification planning exercise [260]. This begs the question: could citizen science be leveraged *beyond* academia, in practical and governmental initiatives? Would the same contributor enthusiasm persist in projects driven by public service and not by research advancement? Given the motivations of citizen scientists to participate in this project discussed in Section 3.3.4, including the real-world nature and social benefit of the project, it is certainly plausible that, if communicated correctly, there could be. Perhaps a desire for data control would limit such non-academic citizen science efforts, but as the world adapts to an increasingly open data system in public and private organisations, this may become less relevant. Probing this possibility, and testing the application of citizen

science in practical application for public service or other initiatives, would be an exciting avenue for future study.

3.5 Key Outcomes

The work in this chapter has shown citizen science to be a feasible method to map remote off-grid populations. A large-scale online citizen science project mapped approximately 1,267 km^2 of rural Kenya, Sierra Leone and Uganda at an average rate of 7 km^2/day. Home annotations produced through citizen science achieved a recall of 93% and precision of 49%, which could be increased to 69% through clustering. The estimated cost for this approach is \$20.84/km^2. The dataset of home annotations produced through the PTTP project is made publicly available for future research use – access details are provided in Appendix B.

These results are orders of magnitude faster and cheaper than benchmarks from the literature. As the cost of satellite imagery decreases ever further, this approach is set to become even more accessible over time. Moreover, this approach can bring together a global community of enthusiastic volunteers motivated by the opportunity to contribute to meaningful research. Community members on PTTP learned about rural developing countries and electricity access through their participation; working on the project even mobilized some to continue learning about the work through independent study.

While this approach can produce the home-level location data needed for electrical system design quickly and cheaply while creating positive benefits for the citizen science community, its speed still needs improvement when considering the scale of the global electrification challenge. A speed of 7 km^2/day is great when mapping a town or small county; however, at a global scale, mapping efforts at this rate would take thousands of years. In the next chapter, it is therefore explored whether home mapping can be accelerated even further by leveraging the power of artificial intelligence and computer vision for home-level mapping at a global scale.

4

Mapping at Scale With Computer Vision

Contents

So far in this work, one method – citizen science – has been explored as a means to map rural off-grid homes for electricity access system design. Though this method offers significant cost and speed improvements over conventional site survey community mapping methods, and comes with the added benefits of proven community learning and mobilization, it cannot scale to match the size of the global electricity access gap within a reasonable time frame. To address this, this chapter therefore tackles the question: *Can we map off-grid populations rapidly at a global scale?* It is argued that by using a computer-vision-based mapping approach which builds on robust training data generated through citizen science, mapping speeds and costs can be achieved that match the scale of the electricity access gap.

Given the numerical and spatial enormity of the global electricity access gap, home-level citizen-science-based mapping simply cannot scale to match this urgent challenge. To illustrate, consider the citizen science approach in Chapter 3, which

achieved a mapping speed of 7 km^2/day with an approximate cost of \$20.84/km^2. These results are appropriate and realistic at a regional mapping level. For instance, a county government wishing to map a 500 km^2 rural area at the home-level for electrical design taking this approach would be able to do so at a cost of \$9,000-10,000 in 2-3 months. This is a reasonable time frame and cost in this context, and is likely represent large cost and time savings over an in-person site survey or non-citizen-science satellite imagery labelling endeavour. However, scaling these costs up to the size of the global electricity access gap tells a different story. The world has 112 million km^2 of rural land, 78 million km^2 of which is in LMICs [261]. Mapping the entire rural area of LMICs at home level with citizen science, at a rate of 7 km^2/day and a cost of \$20.84/km^2, would take over 11 million days (i.e. 30,000 years) and cost over \$1.6 billion.

While one could argue that this would be financially worthwhile, the infeasibility of the time constraint is hard to debate. Even if it is assumed that the speed of mapping would scale linearly with the size of the contributor base, and that more citizen scientists could be recruited to accelerate the process, over 200 million volunteers would need to be mobilized to map the entire rural area of LMICs at home level within a year. This seems highly unrealistic; alternative methods must therefore be explored to accelerate rural mapping at scale.

One promising method to enable home mapping at scale is computer vision. Computer vision can automate tasks which humans can accomplish by eye, ranging from the classification of images based on their contents to the detection or segmentation of individual image features, as reviewed in Chapter 2. With the explosion of AI research in recent years and the dramatic increase in graphics processing unit (GPU) efficiency [262], computer vision algorithms have become supercharged with CNNs. These algorithms dramatically outperform classic predecessors in terms of accuracy and speed. Just as citizen scientists learned to identify homes in satellite imagery through observation and practice, computer vision algorithms can learn to do the same. As more and more satellites are launched (see Figure 4.1) with greater data capacities (see Figure 4.2), satellite EO images to enable this approach

should become increasingly available and inexpensive. Additionally, as GPU power continues to increase [262], the computational resources required to enable computer vision should equally become ever more accessible and powerful. Leveraging EO imagery and GPU-driven computation, computer vision could make home-level mapping for rural electricity planning feasible at a global scale.

Of course, satellite images cannot be used to train computer vision algorithms for home mapping in their raw state: they must first be labeled. Unfortunately, the availability of such labeled satellite data in rural LMICs is limited, as identified in **Gap 3** presented in Chapter 2. Most available training datasets for home location in satellite imagery are either only focused on urban areas (e.g. SpaceNet [80], OpenCities AI [143]), only focused on HICs (e.g. the Landcover.ai dataset in Poland [141], the SkyScapes dataset in Munich [142], the Mnih dataset in Massachusetts [128]), or both. The utility of these data is limited when mapping rural off-grid homes in LMICs, whose appearance (e.g. roofing material, size) and context (e.g. proximity to roads, arrangement, surrounding land use) are often very different than urban or HIC homes. Helpfully, however, citizen-science-generated data can be used as training data in this kind of computer vision research. Large

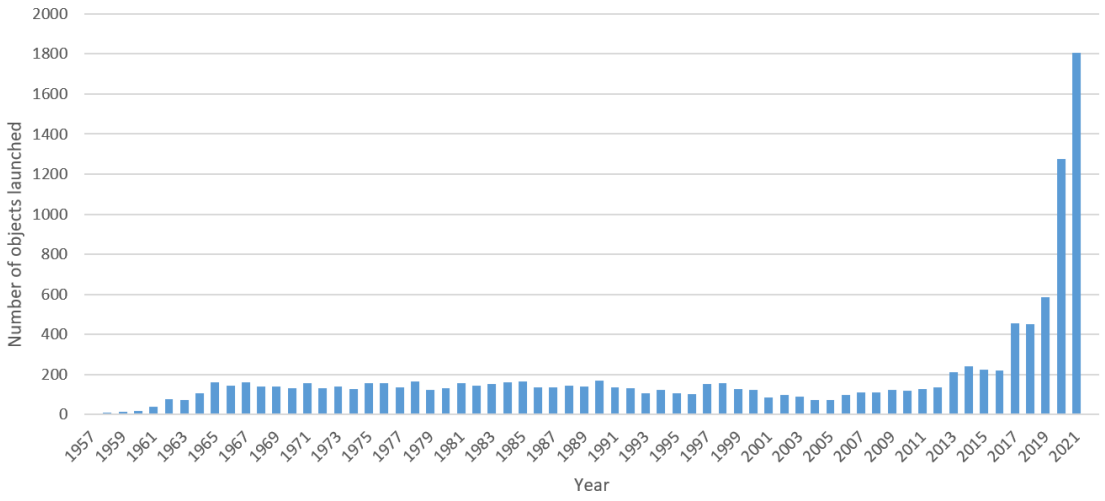


Figure 4.1: Objects launched into space each year as recorded by the United Nations Office for Outer Space Affairs [263]. There is a clear upward trend in launches which is escalating quickly. While not all of these objects are satellites, many are. As more satellites are launched, satellite imagery should become increasingly accessible.

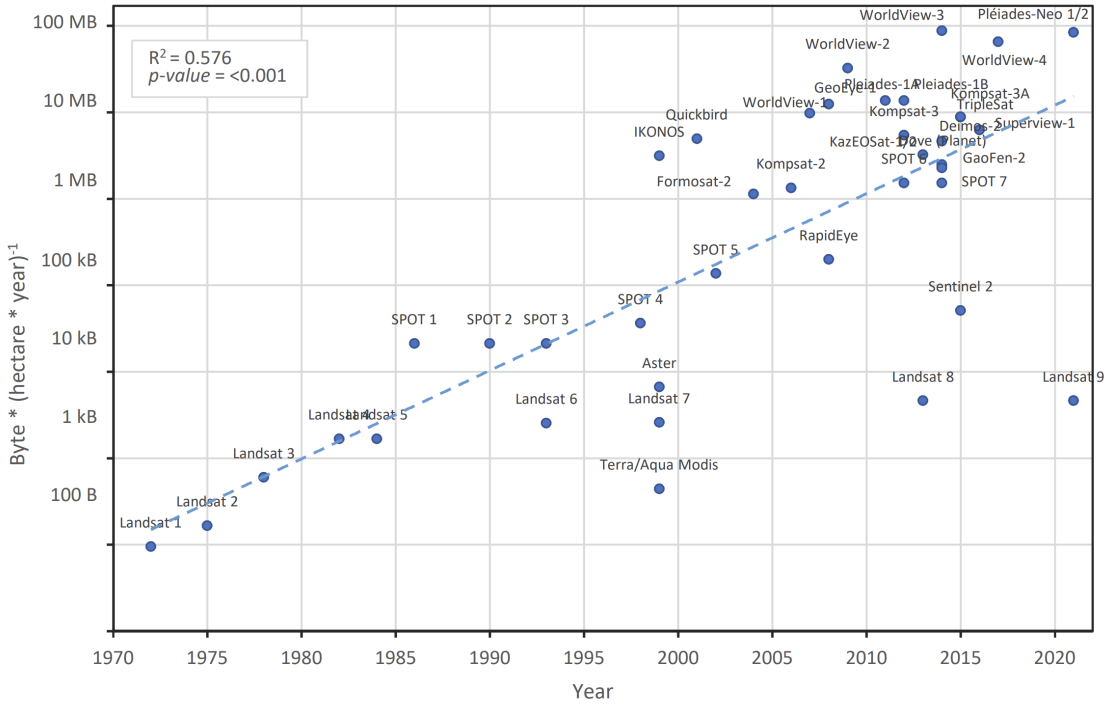


Figure 4.2: Data volume per hectare provided by different satellite constellations over time [264]. The amount of data provided per hectare has increased from 100 B in Landsat 1 to 100 MB in Pleiades-Neo.

annotated image datasets generated through citizen science have been used in computer vision to classify images or detect specific features in previous projects. For instance, the Snapshot Serengeti [265] and Penguin Watch [266] citizen science projects have generated large computer-vision-ready datasets of labelled animal images from camera trap images. The citizen-science-generated home location dataset generated in Chapter 3 will thus be useful here, given the scarcity of other relevant training data.

This chapter argues that computer vision can be used to map off-grid homes at a global scale and at low cost to enable electrification design for universal access. Using VHR satellite imagery and home annotations generated in the PTTP citizen science project, an object detection algorithm is experimentally trained and tested as a means to achieve scalable home-level mapping in Kenya, Uganda, and Sierra Leone. This method retains the advantages of high-temporal specificity and near-worldwide coverage that come with the use of satellite imagery as input data while

promising to cut mapping time and costs significantly compared to on-the-ground benchmarks and previous citizen science results. To be shown to be successful, the computer vision approach must accurately map homes in rural off-grid areas of LMICs at a rate which can scale to global coverage within a reasonable timeframe and cost given the urgency of achieving universal electrification (i.e. less than 5 years, on the order of magnitude of $1 billion).

This chapter proceeds by first outlining the materials and methodologies applied in Section 4.1. This includes the selection of a computer vision algorithm, data pre-processing, configuration and training, and performance evaluation metrics. Section 4.2 outlines the resulting performance of the trained algorithm in the rural home detection task. Then, Section 4.3 elaborates upon insights and implications of the mapping performance as it pertains to electrical system design for rural off-grid areas and discusses scalability. Finally, key insights from the chapter are discussed in Section 4.4.

4.1 Methods and Materials

The methodology used in this chapter is illustrated in Figure 4.3. The annotations produced in Chapter 3 are pre-processed to be used as training data alongside EO satellite imagery. A computer vision algorithm is trained using these data and applied for detection. Its performance is evaluated using cross-validation and is compared to the previous citizen science results.

4.1.1 Computer Vision Approach and Tools

As reviewed in Section 2.1.4 of Chapter 2, computer vision algorithms for classification, object detection, and segmentation can each be applied in different ways to enable satellite-imagery-based mapping. Classification can for instance be used to identify whether each tile is populated (e.g. by assigning a binary "inhabited" or "deserted" label) to create a gridded population dataset similar to HRSL [15] or GPW [248]. However, a gridded dataset is not the goal here;

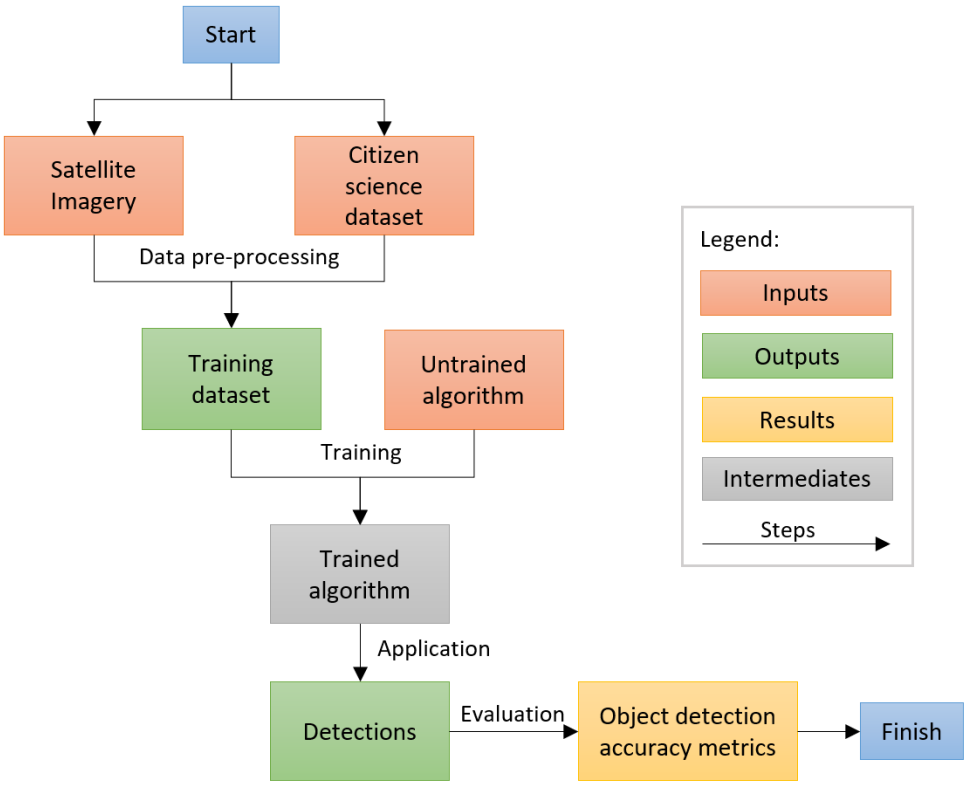


Figure 4.3: Overview of the computer-vision-based method for mapping potential connection points (i.e. homes) in rural off-grid areas.

instead, specific home locations are desired. As such, classification is not used given its unsuitability to the required task.

Object detection and segmentation, however, are both potentially appropriate for the task of mapping homes in rural off-grid areas using satellite imagery. Each can identify individual features in an image, either within a bounding box (object detection) or at the pixel level (segmentation). Note that instance segmentation would be needed in this application, as semantic segmentation would not differentiate between individual homes.

These two algorithm families require different types of training data. Both require labels of individual instances of the feature to be identified; however, instance segmentation requires pixel-level annotations, while object detection requires only bounding boxes. In this case, the available training data produced in Chapter 3 contains bounding box annotations; therefore, object detection is selected as the general computer vision approach.

Computer vision tools are selected to implement and test the feasibility of the object detection based home mapping. Free and open source computer vision tools are preferred, as these are more accessible in resource-constrained contexts (i.e. rural areas of LMICs). Similarly, straightforward and user-friendly tools are preferred, as these minimise the barrier to entry for the use of this method by practitioners and rural development planners. That said, high performance state-of-the-art algorithm implementations and GPU compatibility are also required to match the scale of the problem.

Tensorflow is selected as the computer vision toolkit to meet these needs. Tensorflow has a relatively user-friendly Python-based computer vision application programming interface (API). It is widely supported by an active online community with numerous tutorials and forums available. There is a version of Tensorflow which is compatible with Nvidia GPUs so long as required supporting software (e.g. CUDA, Python dependencies) are installed. The Tensorflow object detection API and model zoo implement many cutting-edge object detection algorithms with minimal barrier to entry for use aside from basic Python literacy. This makes Tensorflow well-suited for practical application in home mapping for grid design.

4.1.2 Training Data Pre-Processing

The home annotation clusters produced in Chapter 3 are used alongside VHR satellite imagery as training data, given the scarcity of rural off-grid homes in alternative training data sources. Specifically, the clusters with the highest R and F_1 compared to the gold standard are used for the main experiment of this chapter, which are those produced using HDBSCAN* with $m_{clSize} = 2$. Unclustered raw annotations are also used in one specific exploration, the results of which are presented in Section 4.2.3; all other results used clustered data.

Several pre-processing steps are required to convert the annotations produced in Chapter 3 to a form which could be used for training in Tensorflow. First, the annotation records are translated to a set of four corner coordinates, based on the five data points recorded in Zooniverse exports which indicate their position: an x and y

coordinate in pixels, an angle of rotation (hereafter θ), and the width (w) and height (h) of the box. While generally highly documented, the Zooniverse documentation is ambiguous about the exact position of the x and y coordinate provided (i.e. whether this point represents the center of the box or any of the corners), and whether θ is in radians or degrees. Through data visualisation alongside satellite imagery and gold standard annotations, it was determined that the x and y coordinate represent the top right corner of the bounding box, measured from the top right corner of the image, and that θ is measured from 0 in degrees *in a negative (i.e. clockwise) direction* about the center of the defined rectangle. This is likely due to the treatment of the upper right corner of an image as $(0, 0)$ in most image processing applications; given this, one could consider the picture in a vertically flipped state, which would make the bottom left corner the conventional $(0, 0)$ for the Cartesian plane and which would make θ actually point in the conventionally positive counter-clockwise direction. However, in this case, the images are not flipped, and instead a $-\theta$ is used when converting annotation records to corner coordinates.

To obtain the x_n and y_n coordinates for each corner of the rotated bounding box for $n = 1 : 4$, the "unrotated" bounding box are first generated by adding w and h to x and y respectively; then, all points are rotated about the center point of this box (i.e. x_c, y_c) by $-\theta$. The "unrotated" annotation corners (x_{n_u}, y_{n_u}) are first calculated as:

$$(x_{1_u}, y_{1_u}) = (x, y) \tag{4.1}$$

$$(x_{2_u}, y_{2_u}) = (x + w, y) \tag{4.2}$$

$$(x_{3_u}, y_{3_u}) = (x + w, y + h) \tag{4.3}$$

$$(x_{4_u}, y_{4_u}) = (x, y + h) \tag{4.4}$$

Then, the box center is obtained as:

$$(x_c, y_c) = (x + \frac{w}{2}, y + \frac{h}{2}) \tag{4.5}$$

Finally, rotated annotation corner coordinates are obtained for each corner n as:

$$x_n = x_c + (x_{n_u} - x_c)\cos(-\theta) + (y_{n_u} - y_c)\sin(-\theta) \tag{4.6}$$

$$y_n = y_c - (x_{n_u} - x_c)\sin(-\theta) + (y_{n_u} - y_c)\cos(-\theta) \tag{4.7}$$

Training data annotations for object detection in Tensorflow are defined as minimum and maximum x and y coordinates of bounding boxes whose edges lie orthogonal to image edges. Such data could be taken directly from the (x_{n_u}, y_{n_u}) pairs, but these annotations are not guaranteed to fully contain each home, especially if annotations were made very close to the roof edges (as was instructed in the original annotation workflow – see details in Chapter 3, Section 3.2.2). Alternatively, satellite images could be clipped around each home annotation such that each image clipping contains one footprint which laid orthogonal to edges. However, if clipped closely to ensure that no additional homes are accidentally included, this might deprive the detector of adequate negative (i.e. non-home) training samples.

Here, to ensure that the entire roof is contained in each annotation and that adequate negative samples were provided, the maximum and minimum x and y amongst all of the rotated (x_n, y_n) pairs are taken as the training labels for Tensorflow. This reduces label accuracy regarding rooftop size, as rooftops on an angle are annotated in a box orthogonal to image edges, and therefore the boxes are oversized compared to the roofs contained therein. Depending on the intended use of the detector output, this error may or may not represent an acceptable trade-off with the speed gained over manual or citizen science mapping methods. In this case, the main aim is to locate connection points at global scale, so connection speed and precise location are the priority, not the exact roof size, making this trade-off acceptable.

The coordinates are then adjusted to reflect the discrete and finite nature of the satellite imagery tiles to be used in training. All annotation coordinates are rounded to integer values to reflect the discrete nature of the satellite image pixels Any annotations which extend outside the bounds of an image tile are "clipped" to the nearest edge. For instance, during annotation, if the annotator's mouse had extended beyond the image edge at $x = 0$, the annotation could have a corner coordinate of $x = -15$ or another negative number. All negative coordinate values are thus set to zero, and all coordinate values exceeding the maximum image width or height are set to that width or height. Finally, any annotations with 0 width or

height are removed. These could result if an annotator had accidentally drawn a vertical line with no width, a horizontal line with no height instead of a bounding box, or the entire width or height of an annotation outside the image area (which was then clipped back to the same value).

The training data are finally processed into appropriate data formats for use in Tensorflow. Annotation coordinates are saved with one annotation per row in a CSV, where each row also contains the name of the file for the pansharpened image tile containing that annotation, the width and height of that tile, and a class value for the annotation. While only one class is used to identify all home annotations in the following experiments (i.e. "home"), these annotations could instead be split into different classes based on roof colour, shape, or size if desired. The resulting CSV is split into training, validation, and testing segments using an 80/10/10 split and converted into a tf.record format required by Tensorflow [267]. Pansharpened image tiles in PNG format are used in training because, based on metadata associated with the exported annotations, the pansharpened tiles are the most frequently annotated by citizen science contributors out of the three imagery types. Additionally, these tiles are likely to contain the most information at the highest resolution of all three imagery types, as they contain the colour information of the multispectral image and the high-resolution edge details of the panchromatic image. Refer back to Figure 3.4 for an illustration of each imagery type for clarification if needed.

4.1.3 Object Detection Algorithm

There are numerous object detection algorithms which could be applied to the home mapping problem, as reviewed in Chapter 2. An algorithm is selected to best suit the specific application of mapping rural homes in LMICs for electrification, principally based on speed and user-friendliness. Mapping at scale for rural electrification does not need to be accomplished real-time, as home locations are not in continuous movement, but should not slow down the electrical design process. As such, training times should be on the order of days at a maximum, and detection times on the order of minutes to hours, over a community-scale region requiring electrification.

As discussed in Section 4.1.1, a user-friendly and open-source approach accessible to academics, engineers, and practitioners is preferred, particularly in under-resourced LMIC contexts where the rural electrification gap is most prominent and pressing.

An algorithm from the R-CNN family is chosen for the user-friendliness of their open-source implementations compared to alternatives like those in the YOLO family. A judicious choice is made to use Faster R-CNN [268], a high performing R-CNN which uses a region proposal network (RPN) to predict where there may be an object in an image and then predicts object class for these regions using a single pass through a CNN. Its RPN shares convolutional features with the detection network, minimising the region proposal bottleneck found in previous R-CNN versions [268]. Specifically, a Faster R-CNN implementation from the Tensorflow 1 model zoo [269] with an Inception v2 RPN [270] is used. Using an implementation from the model zoo makes this approach much more user-friendly to practitioners than having to implement the architecture from scratch. This implementation has been timed at 58 ms per 600×600 (360,000 px) image using an Nvidia GeForce GTX Titan X GPU [269]. As each tile of the highest-resolution imagery used in this work represents approximately 0.0164 km^2, approximately 24,400 km^2/day could be mapped per day with the specified GPU using this implementation according to the benchmarked speed. For context, this translates to mapping the entirety of London in about an hour and a half, or all of Kenya in about 24 days. Note that since this speed was benchmarked on a single Nvidia GeForce GTX Titan X GPU, a somewhat dated model first released in 2015, it is not expected to exactly match the speeds achieved in the the following experiments, which used two more up-to-date Nvidia GeForce RTX 2080 GPUs.

4.1.4 Hyperparameters and Training

A Faster R-CNN implementation is pre-trained with the Common Objects in Context (COCO) dataset and fine-tuned with the prepared training data for home detection. This approach makes use of transfer learning, a common method in object detection which leverages low-level features (e.g. corners, edges, etc.) learned

from generalized object detection (i.e. via pre-training with COCO in this case) to jump-start a specific detection task.

Training hyperparameters are adapted from published hyperparameters for the selected Faster R-CNN implementation [271] initially configured for transfer learning with the Oxford-IIIT Pets Dataset [272]. Data augmentation options are selected to realistically reflect the variance which can be expected in satellite imagery. Satellite imagery tiles may vary in brightness, contrast, saturation, and hue based on atmospheric conditions, time of capture, and other effects. Additionally, they may contain more or less context surrounding each home annotation due to different tile sizes, zoom levels, or image resolutions. The following data augmentation options are therefore implemented: random horizontal flipping, brightness adjustments, contrast adjustments, hue adjustments, saturation adjustments, colour distortions, and image cropping. As these variations are realistic to the task at hand, they make the training data and subsequently the trained detector more robust.

To select a learning rate, rates from 10^{-2} to 10^{-9} are tested for 10^5 steps, descending by one order of magnitude with each testing iteration. The performance of the model trained at each learning rate (in terms of the $AP^{IoU=0.5:.05:.95}$ metric described in further detail in Section 4.1.5) is recorded and visualised to find the learning rate resulting in peak performance. Learning rate values on either side of the initially identified peak are also tested to confirm the peak. The learning rate which produces the best results (i.e. the highest $AP^{IoU=0.5:.05:.95}$) taking this approach is 10^{-3}, as shown in Figure 4.4, so this learning rate is used in object detection experiments.

4.1.5 Performance Evaluation

The performance of the object detection algorithm is first evaluated using average precision (AP) and average recall (AR) metrics. The standard set of AP and AR metrics employed in the COCO competition are used, which evaluate detection accuracy over a number of IoU thresholds, maximum detection counts, and object sizes. These metrics build on the basic definitions of P (Equation 4.8) and R

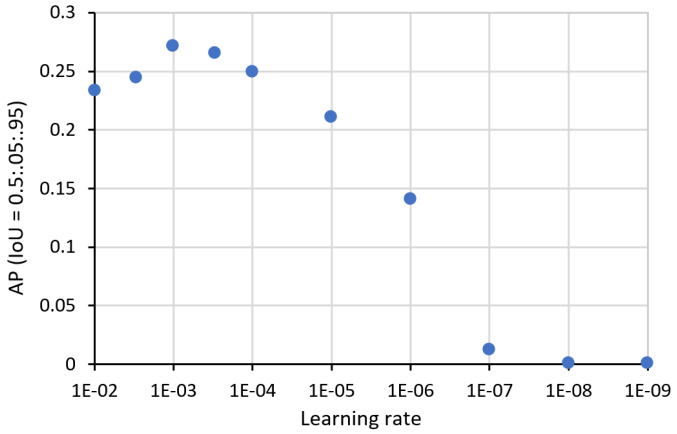


Figure 4.4: Experimental results used to select the best learning rate for subsequent object detection experiments. The best performance (i.e. the highest $\text{AP}^{\text{IoU = 0.5:.05:.95}}$ value) can be seen at a learning rate of 10^{-3}.

(Equation 3.2) provided in Chapter 3. AP is the precision averaged across all recall levels. It describes how many predictions were T_P over n varying R values (i.e. as the confidence in these predictions varies) and is defined as the area under the interpolated P-R curve for any object class being detected:

$$AP = \sum_{k=1}^{n-1} (R_{k+1} - R_k) P_{interp}(R_{k+1}) \tag{4.8}$$

Note that as the P-R curve tends to "wiggle", an interpolated P-R curve P_{interp} generated using number n of R-values is used in this calculation, defined as:

$$P_{interp} = \max P(R'), \; R' \geq R \tag{4.9}$$

For multi-class object detection, the mean average precision (mAP) is the mean of the AP values across all classes. Here, since only one class is being detected, AP is sufficient. AR on the other hand describes how many T_P were "recalled" (i.e. predicted) over various IoU values – that is, for various thresholds dictating how much overlap is needed between a prediction and a detection to count as T_P, as defined in Chapter 3. It is calculated as the mean R over IoU values between 0.5 and 1.0, or twice the area under the R-IoU curve, defined as:

$$AR = 2 \int_{0.5}^{1} R(x) dx \tag{4.10}$$

where x represents the range of IoU values.

In the standard COCO metrics, AP and AR are calculated for different data subsets and constraints to evaluate detection accuracy in different circumstances. These include different IoU thresholds, different maximum quantities of detections per image, and different detection sizes. IoU thresholds of 0.5 and 0.75 are used for AP, as well as the average for thresholds from 0.5 to 0.95 in 0.05 steps. The maximum number of detections considered per image is varied for AR: 1, 10, and 100 detections are considered. Different sizes of detected objects (i.e. "small" with area $\leq 32\times32$, "medium" with $32\times32 <$ area $< 96\times96$, "large" with area $\geq 96\times96$) are used for both AP and AR [273]. These metrics are listed in Table 4.1.

Ten-fold cross-validation is used in this evaluation. This involves training the selected algorithm ten times with different random data subsets to ensure that the results are robust and not artificially inflated by overfitting or by use of a biased partition between training and testing data. The pansharpened image tiles are divided into ten random partitions and Faster R-CNN is trained ten times; during each training round, a different partition of images (and associated annotations) is set aside for testing. After training, each model is tested, and its AP and AR performance metrics are recorded. The final performance metrics for this model are taken as the average of the performance metrics across all ten models.

Note that cross-validation evaluates the model performance against the training data itself, not against gold standard data. This does not give much indication about how the model actually achieves the desired task, but rather it indicates how the model succeeds in replicating training data. Since imperfect training data is used, as described in Section 3.3, model performance is compared to an imperfect standard.

The performance of the trained model is therefore also evaluated against gold standard data to evaluate its ability to accomplish the actual desired task outcome. The model is trained one more time while intentionally withholding the 188 images for which gold standard labels are produced in Chapter 3. This allows a direct comparison between detections, citizen science annotations, and gold standard researcher-produced data. P, R, and F_1 are calculated using the detections on these images as predictions, as described in Chapter 3, to evaluate detection accuracy.

4.2 Results

Faster R-CNN was trained using satellite imagery tiles and citizen science annotation clusters to detect homes using the methods described in Section 4.1.3. After training, models were exported and their performance was evaluated using the metrics outlined in Section 4.1.5.

4.2.1 Cross-Validation

Faster R-CNN was trained ten times for cross-validation with a different data subset withheld for testing each time. The results of the object detection algorithm performance across the ten-fold cross validation are provided in Figure 4.5 and Table 4.1. Note that the columns of Table 4.1 simply show the results over each iteration of cross-validation, to illustrate that the variation between iterations was not significant. Given this, it is the average values in the final column which are the primary interest, and which are plotted by theme in Figure 4.5. Across all metrics, as both AP and AR values range between 0 and 1 with 1 being optimal, it is evident that the performance of the object detection model could be improved. Nevertheless, these results show promise. AP values for state-of-the-art models can range from 0.5 to 0.8 depending on the training data and task at hand. The AP results for $IoU = 0.5$ fall within this range, as shown in Figure 4.5a. The results also provide some indication of detector performance in different types of images with different sizes and quantities of homes. As shown in Figure 4.5b and 4.5d, medium and large objects (i.e. with area $\geq 32\times32$) are detected more accurately than small objects. This indicates that larger homes are more easily detected than smaller homes. It can be theorized that these homes were better annotated in the training data by citizen scientists, generating higher performance. Additionally, performance improves as more homes are allowed to be detected per image, as shown in Figure 4.5c. This makes sense, as some images (e.g. those representing village centers or refugee camps) do contain many homes, and if only one or ten detections are allowed, many homes will be missed.

Table 4.1: Average precision (AP) and average recall (AR) metrics calculated through 10-fold cross-validation of the trained detector.

	Cross-validation iteration										
Metric	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Average**
$AP^{IoU=0.5:.05:.95}$	0.250	0.255	0.277	0.273	0.261	0.250	0.284	0.272	0.260	0.260	**0.264**
$AP^{IoU=0.5}$	0.556	0.546	0.572	0.570	0.553	0.558	0.583	0.570	0.560	0.558	**0.563**
$AP^{IoU=0.75}$	0.193	0.202	0.240	0.230	0.214	0.191	0.243	0.227	0.204	0.204	**0.215**
AP^{small}	0.175	0.182	0.213	0.210	0.196	0.204	0.213	0.204	0.198	0.213	**0.201**
AP^{medium}	0.344	0.329	0.346	0.345	0.346	0.311	0.357	0.346	0.331	0.317	**0.337**
AP^{large}	0.341	0.301	0.369	0.370	0.350	0.395	0.412	0.307	0.348	0.405	**0.360**
$AR^{max=1}$	0.078	0.084	0.083	0.080	0.081	0.078	0.089	0.082	0.081	0.087	**0.082**
$AR^{max=10}$	0.313	0.326	0.326	0.328	0.320	0.310	0.344	0.332	0.316	0.333	**0.325**
$AR^{max=100}$	0.431	0.435	0.443	0.452	0.436	0.427	0.456	0.448	0.435	0.443	**0.441**
AR^{small}	0.371	0.373	0.387	0.407	0.377	0.372	0.395	0.396	0.377	0.386	**0.384**
AR^{medium}	0.510	0.504	0.509	0.511	0.520	0.497	0.523	0.513	0.506	0.512	**0.511**
AR^{large}	0.501	0.505	0.532	0.494	0.492	0.573	0.546	0.455	0.529	0.564	**0.519**

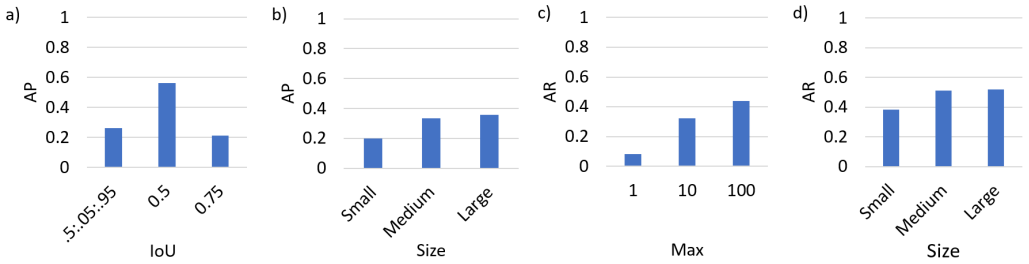


Figure 4.5: Average cross-validation results, including average precision (AP) for different (a) intersection over union thresholds, and (b) detection sizes; and average recall (AR) for different (c) maximum numbers of detections; and (d) detection sizes.

4.2.2 Comparison With Gold Standard

For a more intuitive understanding of performance, Faster R-CNN was also trained with the 188 images with gold standard annotations purposefully withheld. The trained detector was then applied to these images, and P, R, and F_1 were calculated using the gold standard annotations as ground truth. The results were $P = 0.67$, $R = 0.36$, and $F_1 = 0.47$. This P result is is similar to the P of the citizen science annotations used in training (i.e. $P = 0.68$ for the the annotations used with $m_{clSize} = 2$). This indicates that a similar proportion of the detections made by Faster R-CNN were correct as compared to the training annotations; that is, the model was able to "learn" to label with the same precision as the input training data. However, the detector achieved lower R when compared to the citizen-science-generated training data (which achieved $R = 0.487$). This indicates that the trained detector missed more ground truths than the clustered citizen science annotations did. The detector thus seemed to suffer from some accuracy issues that the citizen-science-generated training data did not. This is further explored in Section 4.2.4.

4.2.3 Training With Raw Data

Faster R-CNN was trained one additional time with raw citizen science annotations instead of clustered annotations to see if this would improve the performance compared to the gold standard. Aside from the usage of raw annotations instead of annotation clusters, the data pre-processing and experimental setup was held constant, and the 188 images with gold standard annotations were purposefully withheld from training. The trained detector was then applied on the images containing gold standard annotations, and P, R, and F_1 were calculated using the gold standard annotations as ground truth. The results were $P = 0.53$, $R = 0.57$, and $F_1 = 0.55$. This is a lower P value than achieved in the clustered annotations ($0.53 < 0.67$), but a much higher R value ($0.57 > 0.36$) and an improvement in the F_1 score ($0.55 > 0.47$). These improvements are likely due to the differing data input: the raw annotation data has overlapping annotations on high-consensus homes but also contains more noise than the clustered training set. The raw annotation dataset

is more likely to contain at least one annotation on all homes than the clustered dataset, which purposefully discarded annotations with low consensus, eliminating both a lot of noise and some correct home locations which were more difficult to detect. This was the likely cause of the higher R (i.e. more ground truth homes were included in the training set, meaning the algorithm managed to learn to detect more of them) and the lower P (i.e. more noise annotations are also included in the training set, making the algorithm learn to misdetect more noise as homes). It must also be considered that there are more training data within raw annotations than clusters (i.e. 577,366 raw annotations which could be used in training versus 63,106 clusters), and this has an impact on the accuracy which can be achieved in training.

4.2.4 Visualisations

To gain a better intuitive understanding of detector performance, sample detections were superimposed on satellite imagery tiles and visualized for inspection. Examples from various contexts are visualised in Figure 4.6. Generally speaking, the model appears to perform well and detect diverse home styles in varying contexts.

That said, it is perhaps more useful to understand difficulties the detector faced than to laud its successes. The detector replicated several of the issues present in the citizen-science-generated training data, including struggling to annotate crowded images containing many homes. It makes sense that the detector would replicate this problem inherent to the training data. Furthermore, the detector struggled with lower resolution and lower contrast images, as well as images blurred due to atmospheric effects. These factors made citizen scientists more likely to generate inaccurate annotations, which were then used for training, and therefore these types of inaccurate annotations were replicated by the detector.

The detector however also struggled in certain ways that the citizen scientists did not. Given the diversity of contexts present in the training dataset, the detector appeared in some cases to have learned features which, when cross-applied in different contexts, led to false home detections. For instance, the detector seemed to have learned that a high contrast bounded edge in an image is likely to indicate

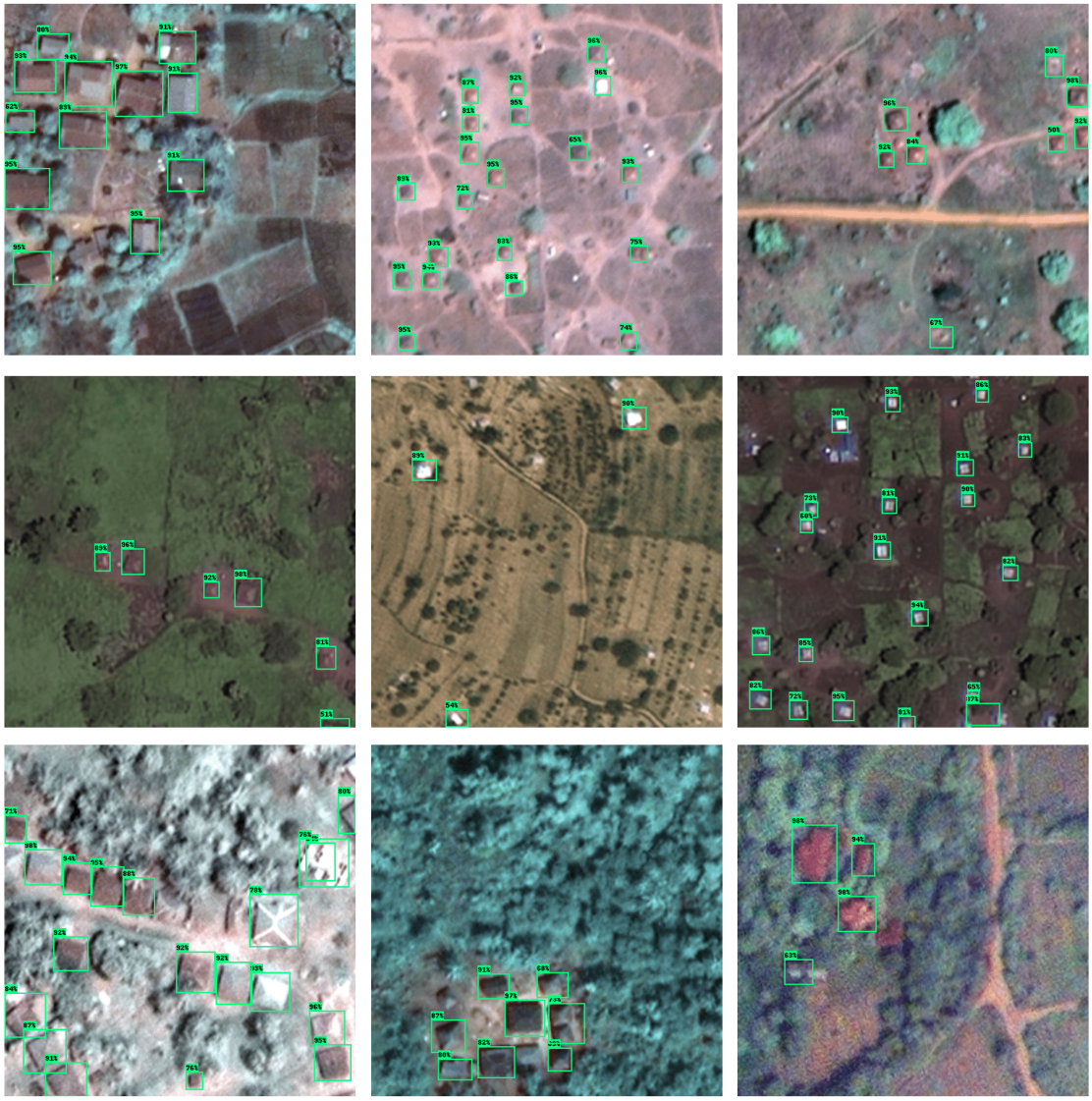


Figure 4.6: Home detection examples for Faster R-CNN trained with the clustered citizen science data in different geographic contexts, superimposed on satellite imagery. Detections with a probability of 50% or higher are visualized as green boxes.

a roof, prompting a detection. However, in certain contexts in the training dataset, dark fences or hedges created a high-contrast feature and caused misdetections (e.g. fenced yards were occasionally detected as homes). While it is easy for a human to tell that a hedge or a fence is not a roof, the detector does not have this contextual knowledge, and so it mistakenly predicts homes in some of these cases. The detector also performed poorly in certain, but not all, agricultural contexts. It is unclear based on intuitive observation what differentiated these contexts and caused such good performance in some circumstances and poor performance in

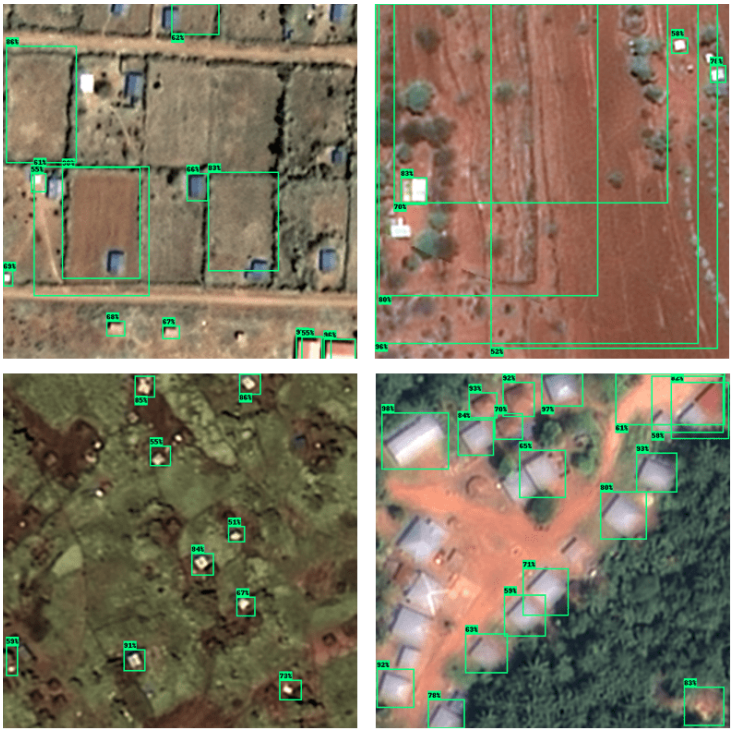


Figure 4.7: Issues observed in the trained detector. It struggled with crowded (bottom left) and blurred images (bottom right), misdetected some fenced homesteads as homes (top right), and performed poorly in certain, but not all, agricultural settings (top left).

others. Further study is needed on this point. Issues in the detection results are illustrated in the examples provided in Figure 4.7.

4.2.5 Training and Application Speed

The model was timed during both training and application. On two Nvidia GeForce RTX 2080 Ti GPUs, training proceeded at approximately 86 ms per step, and detections took an average of 33 ms per detection. To put this in more intuitive terms, this means that 100,000 steps of training would take approximately 2.4 hours, and approximately 30 images could be put through the detector each second for detection. This is an improvement over the benchmark speed provided for this algorithm implementation, likely due to the use of more advanced GPUs and slightly smaller image files.

Table 4.2: Summary of computer-vision-based mapping results compared to citizen science results. Costs for citizen science were previously estimated in Chapter 3. Costs for the computer vision approach are estimated in Section 4.3.2.

Metric	Citizen science	Object detection
Precision (%)	69%	67%
Recall (%)	49%	36%
Speed (km^2/day)	7	42,938
Cost ($/$km^2$)	$20.84	$11.40

4.3 Discussion

The results of the computer vision mapping approach are summarized and compared to the citizen science results from Chapter 3 in Table 4.2. These results show object detection to be a promising scalable mapping method for rural off-grid areas.

4.3.1 Algorithm Performance

The results of the object detection experiments in this chapter confirm the intuition that CNN-driven computer vision algorithms can only work as well as the data you train them with. To train a high-performing object detection algorithm, accurate training data are needed, with P and R values approaching 1. The clustered citizen science annotations used to train the object detection algorithms in this chapter do not achieve this, and so the resulting trained detector performs imperfectly. As highlighted in Section 3.4.1, a number of citizen science process improvements could improve the quality of the training data. This would be a key piece of future work to improve the detector performance. For instance, crowded images could be processed by more citizen scientists, or excluded from training to improve training data quality. It would be interesting to see whether computer vision algorithms trained only on annotations in non-crowded images could still accurately detect homes in crowded images. This is an interesting avenue for future study.

Despite these issues with training data quality, the visual detector results (e.g. in Figure 4.6) indicate that even this imperfect algorithm could make a useful first pass home-level mapping tool. It could at least act as a filtering step, which could be used to simplify the work required from humans to annotate satellite

imagery. For instance, detections could be fed back into a citizen science pipeline for classification (e.g. using this prompt "Does this annotation contain a home roof?"), a task which can be achieved much more quickly through citizen science than annotating from scratch. It could alternatively be used to accelerate and improve the accuracy of a citizen-science generated dataset in a tandem process. For instance, citizens could be integrated in an iterative verification feedback loop with an object detection algorithm to improve annotation quality on crowded images. The algorithm could be periodically retrained throughout the citizen science project, and its detections could be fed back into a citizen science workflow for verification (e.g. as "correct" or "incorect"). These verifications could be used to refine training data on challenging or crowded images over time.

Pre-processing methods could be also tested to improve the quality of the training data and thus the performance of the object detection algorithm. For instance, images could be clipped closely around the annotations used for training prior to training the algorithm to minimize the number of non-labelled homes the algorithm was exposed to as "false" samples. As previously discussed however, this could also create class imbalance issues between positives and negatives; a certain amount of negative samples would need to be preserved. Striking an appropriate balance here is also an interesting line of inquiry for future work.

As highlighted in Section 4.2.4, the overall P and R values do not tell the whole story, as the detector performance is far better in some contexts than others. For instance, the detector struggled in certain agricultural areas, and replicated mistakes made by citizen scientists in crowded areas (see Section 3.4.1). A quantitative study of these discrepancies in performance would add to the qualitative and visualisation-based study presented here. For instance, the trained detector could be run over particular regional data subsets to quantify differences in performance in different contexts. Additionally, it could be investigated whether training multiple separate detectors with different data subsets would improve performance. There is necessarily a limit to the regional or contextual specificity of any detector – to ensure that there is adequate training data available to train a detector, and

for it to be a useful tool which is easy and efficient to apply over large areas, tradeoffs between specificity and performance must be made. However, increasing specificity slightly (e.g. to country level, or for specific settlement styles, such as agricultural and refugee) may offer improved performance with little extra effort. This is an interesting avenue for future study.

4.3.2 Cost and Speed

To validate the presented argument, this approach must improve upon the citizen science approach in terms of speed (km^2/day) and cost ($/$km^2$) of mapping, showing it to be appropriate to the scale of the electricity access gap. The citizen science project which produced the initial training data achieved an average mapping speed of 7 km^2/day, which was already much faster than the on-the-ground or manual benchmark. Meanwhile, the Faster R-CNN implementation used in this experimental setup could map at 42,938 km^2/day using two Nvidia GeForce RTX 2080 Ti GPUs. This is a speed improvement of multiple orders of magnitude.

Once a detector is trained using these data, each additional km^2 mapped tends towards a cost of $10 (i.e. the cost of imagery). For instance, if another 10,000 km^2 of imagery were mapped using the trained detector produced using PTTP annotations, the cost per km^2 mapped would drop to $11.40, assuming a $2,000 investment cost in a GPU to run the detector[1]. The only other cost besides the imagery and the graphics card is researcher time – which, given the speed of the detector trained here, is negligible (i.e. 10,000 km^2 can be mapped in less than one day).

Global Cost and Speed

To illustrate the scalability of this approach, the example of mapping the entire rural land area of LMICs globally is considered. First, to map rural areas of LMICs at home level using this approach, adequate representative training data for all LMICs would be required. These data could be acquired through a series of citizen science projects similar to that presented in Chapter 3, with very small imagery

[1]Nvidia GPUs compatible with Tensorflow are available in the $300-$2,000 range – see `https://shop.nvidia.com/en-gb/geforce/store`.

samples compared to the scale of the entire area to be mapped. The costs of such a project can be estimated as follows:

- **Cost of training set imagery**: A common rule of thumb is that at least 1,000 annotations are needed per object class to train object detection algorithms successfully. Given the diversity of rural settlement styles and home types in LMICs, 5,000 annotations per country could be targeted, stemming from settled regions with varied styles. As a conservative estimate, 10,000 image tiles similar to those created in Chapter 3 could be selected for annotation per LMIC being mapped to ensure that 5,000 annotations can be achieved. Regions could be selected based on prior knowledge of regions with differing rural settlement styles, and tiles could be pre-filtered using larger-scale population density datasets such as HRSL or GPW. As there are 137 LMICs recognized by the World Bank [274], 685,000 tiles would need to be annotated to generate this training dataset; with 0.0164 km^2 tiles as used in Chapter 3, this is 11,234 km^2 of imagery. This would cost $111,234 at $10/$km^2$ [257].

- **Citizen science duration and cost**: At the speed of 7 km^2/day achieved in Chapter 3, annotating these 11,234 km^2 of imagery would take approximately 4.4 years to complete. However, it is worth considering that the citizen science project in Chapter 3 was run with no promotion budget. With an investment in promotion and additional staff time on the order of hundreds of thousands of dollars, perhaps additional citizen scientists could be mobilized to double or even quadruple the contributor base (i.e. approximately 12,000 or 24,000 contributors respectively), resulting in a proportional decrease in execution time to 1-2 years. This is not an unreasonable increase; popular citizen science projects have contributor bases in this range (e.g. 30,000+ for Snapshot Serengeti[2], 77,000+ for Galaxy Zoo[3]).

[2] `https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti`
[3] `https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/`

- **Researcher time for citizen science execution**: The cost of a researcher to run the basic mechanics of the project would be $115,500, assuming 4.4 years of half-time work at 18.75 hours/week and 50 weeks/year, paid at a rate of $28/hour [258]. Should the project be accelerated as described above, this may be better allocated to a full-time researcher over two years.

Subsequently, to train a detector using these data and apply it at global scale, the following costs would be incurred for computing resources, staff time, and imagery:

- **Computing resources for detector training and application:** These will be dominated by either the purchase of GPUs or a subscription to a cloud GPU computing resource such as Google Colab. In the case of purchasing GPUs, Nvidia models compatible with Tensorflow are commercially available in the $300-$2,000 range[4]. At a speed of 42,938 km^2/day, as achieved using two GPUs by the trained detector in this work, the entire rural LMIC area globally could be mapped at home level in 1,817 days, or 5 years. By increasing the quantity of GPUs used and splitting the detection task amongst them, this could be reduced to a year or less (i.e. using 10 GPUs instead of two). These ten GPUs would incur a maximum $20,000 one-time start up cost.

- **Researcher time for computer vision:** Staff time will be required to run training and detection. Over one year, this should run $52,500 for one staff member, assuming 37.5 hours/week and 50 weeks/year paid at a rate of $28/hour [258] as above. One could invest in further GPUs to reduce the staff time required (as more GPUs would accelerate the entire process). There is a trade-off between staff time requirements and GPU investment; this can be cost-optimised depending on the GPUs to be purchased and the hourly rate of researchers to be hired.

- **Satellite imagery for detection**: Unsurprisingly, at the scale discussed, the satellite imagery required as input data for the area to be mapped is the

[4] `https://shop.nvidia.com/en-gb/geforce/store`.

largest expense. For instance, to map the entire rural area of LMICs with imagery costing \$10/km^2, the imagery cost would be over \$780 million dollars. This is undoubtedly expensive. However, considering the huge global benefit and possible reuse of this data in other development initiatives, it is not at all absurd, particularly if large organizations or country governments are willing to contribute to the cost. Additionally, one could reduce this cost by pre-filtering the rural area of LMICs with a population density dataset, and only acquiring imagery in the areas which are anticipated to be inhabited. As population density datasets are not perfect, this could mean that some households, particularly in low-density areas, are missed. However, it would eliminate a lot of unpopulated land, which would be likely to result in significant cost savings. This tradeoff would need to be negotiated by the project team.

In sum, mapping the entire rural land area of LMICs globally using this approach would cost \$299,234 for training imagery, researcher time, and computing facilities, and up to \$780 million for imagery upon which to apply the detector, though this can be reduced dramatically through pre-filtering. These costs are summarized in Table 4.3.

Table 4.3: Summary of costs for computer-vision-based mapping over the entire rural area of low- and middle-income countries globally.

Research Stage	Item	Cost
Citizen science	Imagery	\$111,234
	Researcher time	\$115,500
Computer vision	Computing (GPUs)	\$20,000
	Researcher time	\$52,500
	Imagery	Up to \$780,000,000

Country-Level Cost and Speed

Moving down one step in scale from global to country level, consider as a case study the cost and timeline of using this approach to map the entire rural area of Kenya (i.e. 576,334 km^2 [261]). To generate a training dataset with thousands of rural Kenyan homes at the speed achieved in Chapter 3 would take less than three months.

Assuming a 500 km^2 sample for training data, imagery would cost approximately $5,000. Researcher time would cost $6,825, assuming a half-time researcher for 13 weeks paid at $28/hour [258]. Then, by using two Nvidia GeForce RTX 2080 Ti GPUs to train and apply the Tensorflow Faster R-CNN implementation used here, the area could be mapped in under two weeks. Budgeting $10/$km^2$ for imagery, $5,000 in start-up computing costs to acquire two GPUs, and an additional month of full-time staffing ($4,515) to supervise the detector training and application, the whole area could be mapped for just under $5.8 million. While $5.8 million is a hefty bill, it is well within the range of governmental budget items for infrastructure or development projects. Again, this is dominated by the imagery cost, which will decrease as satellite EO imagery costs decrease. Imagery costs could also be reduced dramatically through pre-filtering imagery purchases with a population density dataset, at the risk of missing homes in sparsely populated areas.

Cost and Speed Implications

As illustrated in these examples, the larger the area mapped using the same training data set and computing resources, the closer the cost per km^2 mapped gets to the cost of the imagery itself. This is illustrated in Figure 4.8 for the case of mapping all of rural Kenya. With all costs and time constraints considered, this approach allows home level mapping to be easily scalable at the country level and feasibly scalable at a global level within a realistic timeline and cost.

4.4 Key Outcomes

The work in this chapter has shown computer vision to be a an effective method to map remote off-grid populations at global scale. Given the scarcity of other useful training data in a rural context, the citizen-science-generated VHR satellite imagery annotations generated in Chapter 3 are used to train a Faster R-CNN model to map rural homes in Kenya, Uganda, and Sierra Leone. The detector achieved a precision of 67% and recall of 36% when trained on the clustered citizen science annotations

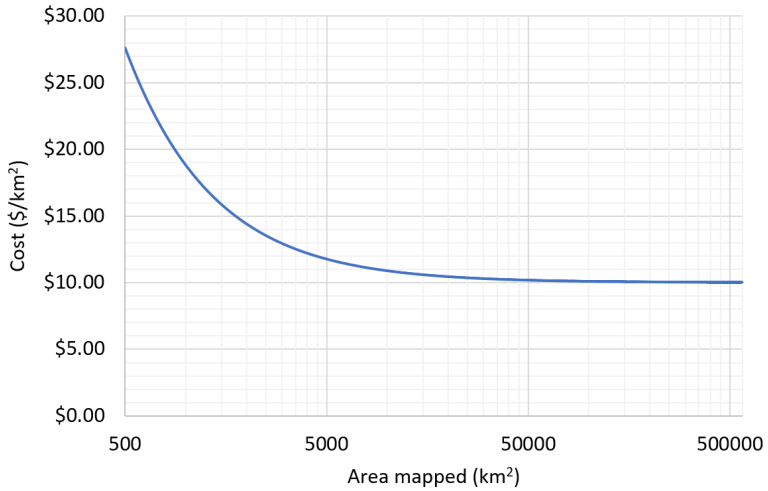


Figure 4.8: Costs of mapping with the object detection approach as the mapped area increases, for the case study of mapping rural Kenya. As more area is mapped, the cost of mapping approaches the cost of imagery. Assumptions are detailed in Section 4.3.2

.

and could map at a rate of 42,938 km^2/day. The recall could be improved to 57% by training on raw annotations instead of annotation clusters.

This work shows that once adequate training data is acquired, the costs per km^2 for computer-vision-based mapping are marginal beyond the cost of the satellite imagery data input. This approach is proven to be scalable for home-level mapping of rural area of LMICs worldwide to enable electrification design. Furthermore, these methods will become increasingly accessible as VHR satellite imagery becomes more widely available and cheaper, and computational power continues to grow.

With the location data problem for potential connection points solved at a cost and speed acceptable to the scale of the rural electrification challenge, the demand data gap is next considered.

5

Understanding Spatially Specific Demands

Contents

To design electricity access systems for rural off-grid people in LMICs, two key data gaps must be filled: the locations of the potential consumers, and their anticipated electrical demands. Having developed methods in Chapters 3 and 4 to tackle the location data gap, this chapter explores the issue of demand data by addressing the question: *Can we estimate the diverse and spatially specific energy needs of off-grid populations?* It is argued that by leveraging existing empirical socioeconomic datasets, spatially specific estimates can be generated for off-grid communities quickly and at low cost to enable electricity access system design.

As discussed in Chapter 2, the most appropriate demand estimation method in rural off-grid LMIC contexts is the stochastic bottom-up engineering-based approach, which models demands based on appliance power ratings and physical behaviour and aggregates these at household and community levels. This approach allows for a high degree of spatial specificity by utilizing site-specific input data

to configure appliance usage and ownership assumptions. This is useful when modelling anticipated demands in off-grid regions of LMICs with high spatial cultural variance and varying energy usage norms.

To estimate demand using bottom-up engineering methods, energy usage must first be understood at the household level in terms of devices and appliances owned, and how these are used. As identified in Chapter 2, household energy use can be generally divided into four categories: (1) lighting, (2) cooking, (3) thermal comfort, and (4) other appliances. The importance of each of these categories, and the preferred energy vector to accomplish each, varies based on context.

The proportion of households expected to use electricity to fulfill each of these needs must be integrated in demand estimation. As discussed in Chapter 2, while one might safely assume that electric lighting will be near universal amongst electrified households, the prevalence of electric thermal comfort and other appliances can vary significantly. Additionally, the expected timing and frequency of usage of each electrical technology will vary. As such, data is needed to understand how the expected ownership and usage of electrical appliances will vary across households in the context under study.

Regrettably and unsurprisingly, appliance use datasets which are typically used as inputs for demand estimation in HICs [155–160, 275] are lacking in LMICs, as identified in **Gap 4** from Chapter 2. Recent MTF surveys conducted by the World Bank and ESMAP [276] are beginning to collect this type of information in LMICs, but do not approach the precise georeferencing and minute-level temporal resolution in HIC datasets. Furthermore, these data are necessarily unavailable in communities without electricity access, as electrical appliance usage cannot be recorded in an unelectrified community. Cross-applying time-of-use data collected in HICs for LMIC demand estimation is not realistic given the cultural and economic differences impacting energy usage.

Demand estimation methods built for rural LMICs [168, 170, 174] therefore instead rely on local surveys to understand energy needs and aspirations. However, such surveys are expensive, time-consuming, and prone to inaccuracies. Off-grid

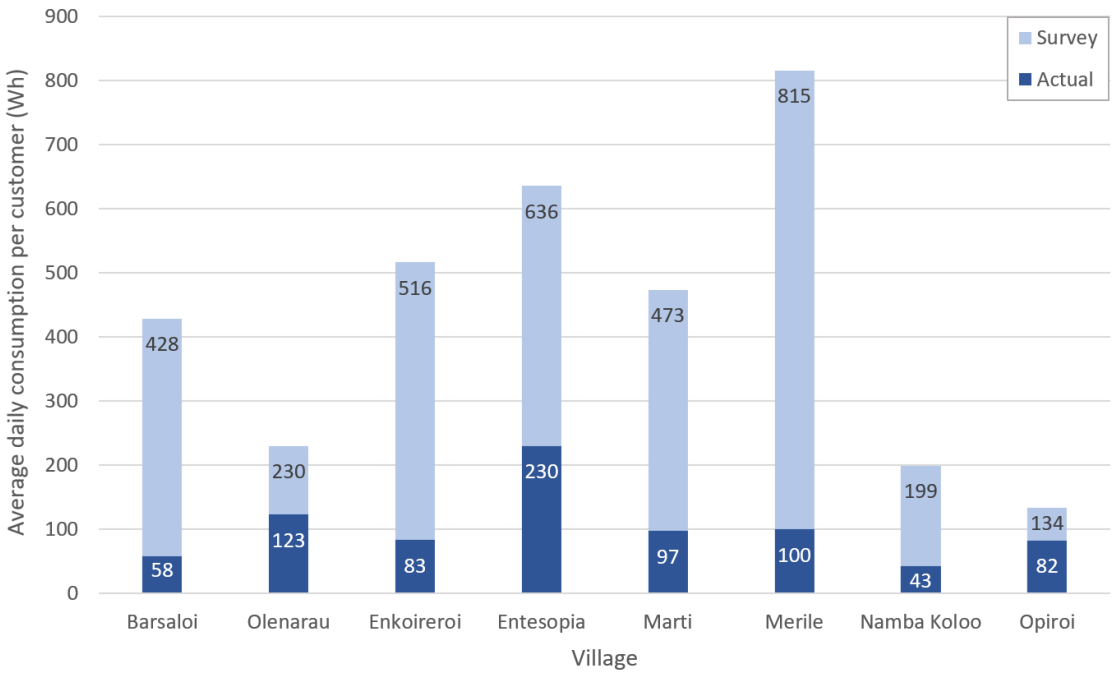


Figure 5.1: Surveyed and actual average daily demand per customer in eight Kenyan villages [178]. Load profiles constructed from local surveys on aspirations in off-grid communities can be prone to significant error, resulting in oversized and therefore unaffordable systems.

communities tend to often overestimate their future consumption given their limited experience with electricity [177], as shown in Figure 5.1. This can unnecessarily increase system size and costs [178]. While a certain amount of oversizing can beneficially allow for load growth and add resilience, dramatic oversizing will simply render energy unaffordable, limiting the resulting access. Innovative approaches such as service-based [180] or value-based [181] surveying can collect more detailed and nuanced data on energy and underlying developmental needs than traditional surveys. However, these approaches entail a higher time investment and cost, and require strong community trust difficult to achieve at scale.

Given these issues, it can be tempting to use crude demand estimation heuristics [165] or to "design for" a specific tier of access as defined by the MTF [189] (e.g. as in OnSSET [233]). However, such approaches disregard the local specificity and diversity which must be taken into account in the design best-fit systems. Therefore, as identified in **Gap 4** from Chapter 2, spatially specific, scalable, and accurate

demand estimation approaches for off-grid LMIC regions are needed. These should account for local context and diversity while producing realistic estimates quickly and cheaply. Such methods can serve as a "sanity check" to ensure that systems are not unnecessarily and dramatically oversized based on unrealistic community aspirations.

In this chapter, a scalable and spatially specific demand estimation approach is proposed to estimate electricity demands in rural off-grid communities at low-cost using existing empirical data. MICS data produced and made freely available by UNICEF are used to understand the combinations of appliances owned amongst different spatial subsets of a country and their relative prevalence. These results are then used as inputs to an LMIC-tailored stochastic engineering demand estimation approach to generate spatially specific demand estimates which account for intra-community diversity. This approach is applied in a case study in Sierra Leone to explore the spatial variance in resulting electrical demand estimates and the discrepancies between these estimates and tier-based alternatives. Finally, the implications of this approach are discussed in terms of electrical design and energy affordability. To be shown to be successful, this approach must (1) produce spatially specific demand estimates which (2) show significant geographic variance, and (3) differ from tier-based and non-spatially specific load profiles in ways that could influence electrical design, while (4) remaining scalable across rural off-grid contexts in LMICs quickly and at low cost.

This chapter first outlines the methods and materials used in the work in Section 5.1, including the MICS survey data, the appliance ownership analysis method, and the stochastic bottom-up demand estimation approach. Then, Section 5.2 presents the results of a case study application in Sierra Leone: first, the spatially specific appliance ownership trends resulting from analysis of MICS data, and then load profiles estimated using these data as inputs to a stochastic bottom-up method for synthetic communities for each region. The results are compared with an MTF tier-based demand estimation approach. The implications of these results in electrical design and the core limitations of the work are elaborated in Section 5.3, and the key takeaways are outlined in Section 5.4.

5.1 Methods

The general methodology for this chapter is illustrated in Figure 5.2. First, appliance ownership is analysed using existing MICS data. Then, load profiles are estimated using appliance ownership data as inputs. These are finally compared with tier-based load profile estimates.

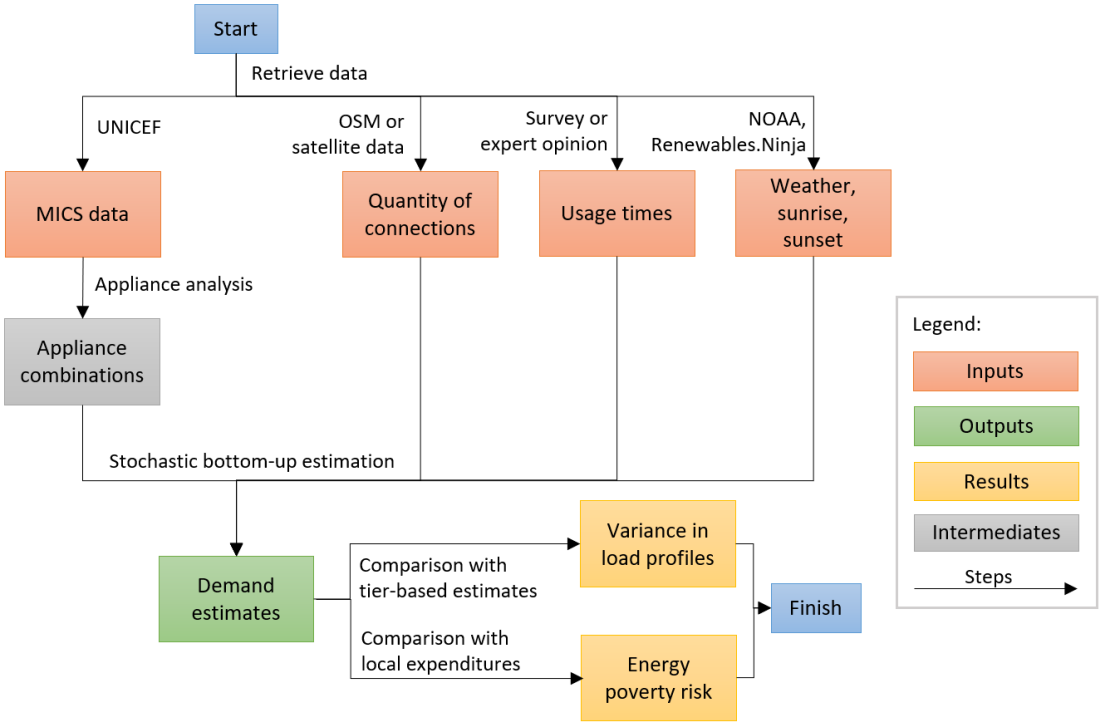


Figure 5.2: Overview of the methodology applied for spatially specific stochastic demand estimation in rural off-grid areas.

5.1.1 Appliance Analysis

Spatially specific appliance ownership is studied in LMICs using MICS data. The MICS programme surveys key indicators of women's and children's health; it is currently undertaking its seventh survey round since its inception in 1993. MICS rounds three through six document ownership of a selection of household appliances and cover 96 countries, 89 of which are LMICs, as shown in Figure 5.3. As these data are drawn from an existing UNICEF-funded programme, there are no data collection costs associated with re-using them for demand estimation.

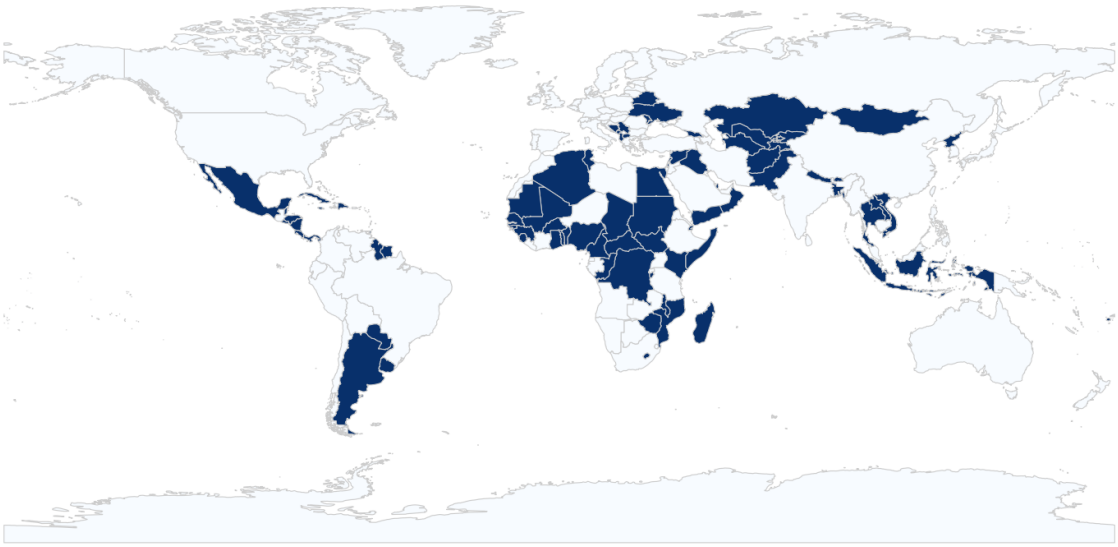


Figure 5.3: Geographic coverage of the Multiple Indicator Cluster Survey rounds 3, 4, 5 and 6 datasets collectively, shown in dark blue. These datasets document appliance ownership in 89 low- and middle-income countries.

Table 5.1: Multiple Indicator Cluster Survey (MICS) questions documenting home location, electricity access, and appliance ownership. Note that "Area" refers to urban or rural designation.

MICS round	Area	Region	Electricity	Appliances
MICS 3	HH6	HH7	HC9	HC9
MICS 4	HH6	HH7	HC8	HC8
MICS 5	HH6	HH7	HC8	HC8, HC9
MICS 6	HH6	HH7	HC8	HC7, HC8, HC11, HC12

The MICS survey tools collect data on locations of households, their electricity status, and the appliances they own. Relevant questions for each survey round are listed in Table 5.1. New appliance options have been added over time: while MICS rounds three and four record mobile phones, radios, televisions, refrigerators, and fixed telephones, MICS round five adds the option for country-specific appliances, and MICS round six additionally studies computer ownership. It is assumed that country-specific appliances are added based on their importance and prevalence in the country being surveyed. Note that the ownership of electric lighting is not tracked in MICS, but is assumed to be present in all newly electrified households.

The electrification status of homes collected in MICS is used to pre-filter the data; appliance ownership is only studied amongst those who report existing access.

By studying electrified households, a realistic appreciation of what households may own once they have electrical access is obtained. This filtering additionally reduces appliance ownership "noise" from any homes without electricity access who happen to own appliances. However, it is important to note that electrified households in countries with low overall electricity access rates are likely to be relatively wealthy. This increases their ability to purchase appliances and electricity, potentially inflating their demands compared to an average or low-income household, as will be later discussed in Section 5.3.5.

The relative prevalence of appliance *combinations* are studied across households (i.e. not simply the relative prevalence of each appliance across households in a community). It is hypothesized that certain combinations of appliances are unlikely to occur in practice, even if their constituent appliances are owned in the community in varying quantities. So, a combination-based approach is taken to preserve the option to generate realistic and representative household-level profiles.

All possible combinations of the n appliances studied in the MICS under consideration are generated. Combinations of size $r = 1 : n$ are considered, as a household may own any number of the appliances. Additionally, one combination of size $r = 0$ must be considered, representing a household that owns none of the studied appliances but is electrified (i.e. only owning lighting). The total number of possible appliance combinations (n_{combs}) can therefore be calculated as:

$$n_{combs} = (\sum_{r=1}^{n} {}_nC_r) + 1 = (\sum_{r=1}^{n} \frac{n!}{r!(n-r)!}) + 1$$

The number of households who own each combination of appliances is then calculated over different spatial subsets of the MICS dataset, including along regional, urban, and rural divides. Data subsets with higher spatial resolution can be pursued if available and feasible in the MICS under study. The relative proportions of households owning each appliance combination are studied to understand the diversity of loads in that subset. The top ten appliance combinations are compared across subsets to identify similarities and differences in appliance ownership trends.

Popular appliance combinations are also ranked in terms of their electricity access level (i.e. as per the MTF) to understand the relationship between variance in appliance combinations owned and access rates. This is accomplished by first ranking all appliances studied in terms of access level based on the MTF tier at which each appliance is introduced [189]. Where two appliances are introduced in the same tier, they are ordered by the appliance's minimum energy consumption for that tier. Then, the combinations of appliances are ranked by access level based on the number of appliances included in the combination and the access ranking of those appliances. The prevalence of appliance combinations in data subsets is visualised in access level order to understand the variance in access levels across households.

5.1.2 Demand Estimation

The results of the appliance ownership analysis are used to configure stochastic bottom-up engineering demand estimation. Load profiles are built up from the appliance level, based on appliance combination prevalence, to the household level, and then aggregated at the community level. The demand estimation framework adopted in the RAMP model [174] is used, given its user-friendly open-source Python-based approach accessible to practitioners or policymakers in under-resourced LMIC contexts. However, the appliance ownership results could be similarly applied in other stochastic bottom-up engineering demand estimation frameworks.

The demand estimation approach can be summarized as shown in Algorithm 1. Three modelling layers are employed: usertypes, users, and appliances. The relationship between these attributes is illustrated in Figure 5.4. To estimate demand in a community of n households, a set of i different usertypes is defined, each containing a certain number of households (j) which each have a given set of appliances (k). The quantity j of the n households belonging to each usertype i is defined based on the prevalence $frac_i$ of the relevant appliance combination in the MICS data being considered (i.e. $j = n \times frac_i$). Appliance types belonging to each usertype are defined in terms of their power consumption (P), total time of use in a day (t_{tot}), minimum time kept on after switch-on (t_{min}), time frames

Algorithm 1 Stochastic bottom-up algorithm used to estimate community demands.

```
Get randomized t_peak within theoretical peak at maximum usage.
frame_peak ← [t_peak − t, t_peak + t]
P_sum ← [0, 0, 0...0]                          ▷ Placeholder for daily profile
for i in Usertypes:
    for j in Users:
        for k in Appliances:
            Check f_ijk for appliance usage.
            if Appliance_ijk is used:
                Get randomized t_tot
                Get randomized frames
                t_running ← 0
                while t_running < t_tot:
                    Compute random switch-on time and duration t_on
                    Ensure t_on in frames & t_on > t_min_ijk
                    if P of Appliance_ijk can vary:
                        Compute randomized P during switch-on event
                    Compute randomized m from [0...m_ijk]
                    Compute P_ijk for duration of switch-on event
                    P_sum ← P_sum + P_ijk
                    t_running ← t_running + t_on
```

in which a switch-on can occur ($frames$), duty cycle ($cycle$), weekly frequency of use (f), quantity of occurrences in a household (m), and a constraint to define whether switch-on for all instances of this appliance in a household are always simultaneous ($fixed$). A percentage random variability (δ) is attributed to t_{tot}, $frames$, and optionally P (i.e. for thermal appliances or those with variable power). It is suggested to assign δ to one of three values based on the likelihood of random variability: $\delta = 0$ for no variability, $\delta = 20\%$ for a small amount of variability, and $\delta = 35\%$ for more substantial variability. However, these parameters can be configured to specifically suit the study context.

To account for the higher likelihood of intensive appliance use during a peak window ($frame_{peak}$) wherein household members are more likely to switch on more than one of the same appliance (e.g. multiple indoor lights), a peak time (t_{peak}) is calculated. This is randomly selected with uniform distribution from the window of time in which appliance usage would peak under the fictitious assumption of

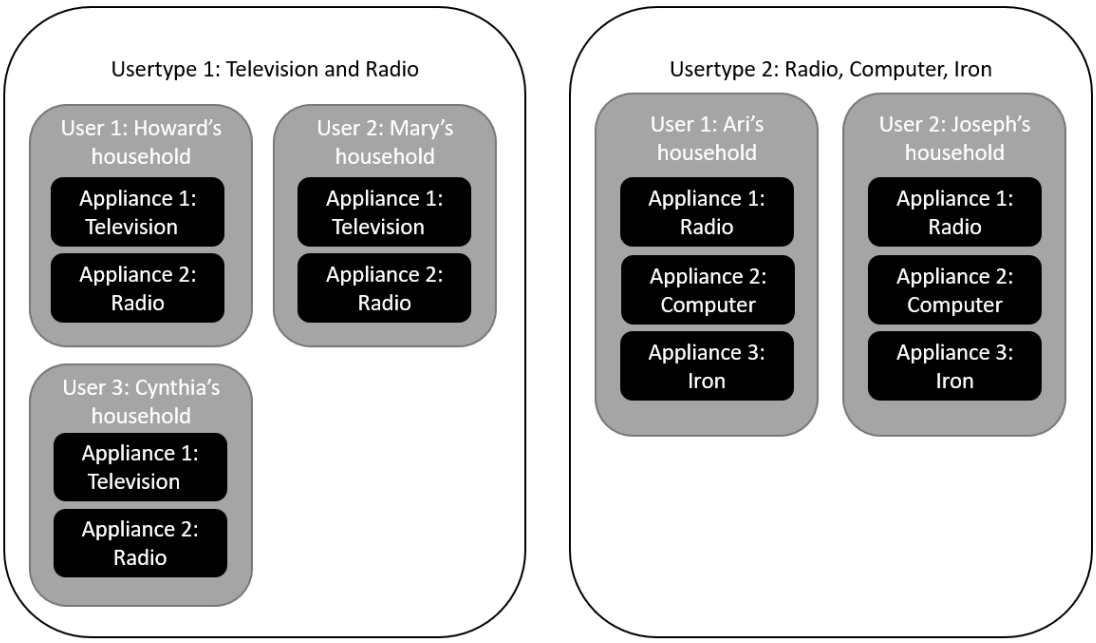


Figure 5.4: Illustration of the nested modelling layers employed in stochastic bottom-up demand estimation. In this example, there are two usertypes; the first has three users, and the second has two users, for a total of five users in the community. Each user in usertype 1 has two appliance types, while each user in usertype 2 has three appliance types. These appliances and their stochastic behaviour are defined and aggregated to estimate demands.

all appliance instances in the household being used at full power. Then, a peak time window of usage is computed as $frame_{peak} = [t_{peak} - t, t_{peak} + t]$ where t is the product of a random sampling with normal distribution around t_{peak} and standard deviation equal to $t_{peak} \cdot \delta_{peak}$ and $\delta_{peak} = 0.15 \cdot t_{peak}$ [173]. The number of appliances turned on ($m_{on_{ijk}}$) is selected randomly based on a uniform distribution ranging $[0, m_{ijk}]$ outside of $frame_{peak}$, whereas it is selected randomly based on a Gaussian distribution with a mean value that is the average of the range $[0, m_{ijk}]$ and a standard deviation at the extremes of $[0, m_{ijk}]$ during $frame_{peak}$.

To configure this approach, the results of the previous appliance analysis are used to define usertypes, by creating one usertype per appliance combination represented in the relevant data subset. The quantity of households to be modelled for any particular community is obtained from either OSM (where accurate and complete), or through the methods in Chapters 3 and 4. It is assumed that all households

are to be modeled as gaining electrical access and using their connection; however, this could be modified if desired.

Note that in the case where there are many possible appliance combinations and a small number of households, rounding error can affect the quantities j and n. For instance, in a case where four different appliance combinations each account for 0.5% of the population, if a community of 50 households is modelled, the number of households represented by each of these combinations will be 0.25 and will thus each round down to 0, effectively erasing one household from the total community. To counteract this, if rounding error makes the total community population allocated to usertypes less than n, additional homes are added to usertypes to reach n by either: (a) adding households to the lowest-access combination represented in the data subset, or (b) randomly allocating households to usertypes represented in the data subset. Here, the first approach is taken. Similarly, if more than n homes are allocated to usertypes, households are removed to reach a count of n.

Other input parameters are retrieved from openly-available datasets where available and sourced from expert opinion where unavailable. Appliance P values are taken from the MTF [189]. Sunrise and sunset times are retrieved at the specific locations of study from the National Oceanic and Atmospheric Administration (NOAA) [277] to time the possible $frames$ of use for lights and appliances. Weather data are retrieved at specific study locations from Renewables.Ninja [54] – these are used to define the frequency of use of thermal comfort appliances and the duty cycling of appliances affected by outdoor temperature (e.g. refrigerators). Fan use is varied based on temperature fluctuation throughout the year; during the hotter, dry season, fan use is increased compared to the cooler, wet season. Refrigerators are assigned a lighter duty cycle during cooler night-time hours and a heavier duty cycle during hotter day-time hours. Additional input data on the timing of appliance use (e.g. $frames$, t_{tot}, and t_{min}) are sourced from country-specific datasets where available, or expert opinion where unavailable. This varies based on the context considered. Further details on the sourcing of these input data specific to the case study will be discussed in Section 5.2.

5.1.3 Affordability Check

Resulting demand estimates are analysed alongside household expenditure data to understand whether the modelled demand would place communities at risk of energy poverty. Expenditure data can be found through different data sources depending on the country studied, including through a census, an LSMS, or an integrated household survey (IHS).

Using these data, a maximum monthly energy budget before energy poverty risk is calculated as a percentage of total expenditure. Thresholds for this vary in the literature; given the propensity for fuel stacking in LMICs, a threshold of 5-10% of household expenditure on electricity can be used, and 10% is used here. The average electricity usage in kWh per household per month is then obtained from the demand estimate, and the energy budget is divided by that number of kWh. This sets a price cap for electricity per kWh before energy poverty risks kick in, assuming a price-per-kWh tariff structure.

These rates are compared with existing grid electricity tariffs to determine if they represent politically coherent energy prices. They can also be compared with the capital cost of proposed infrastructure to meet estimated demands to determine payback period and any required subsidy. If no subsidy is available, it can be determined whether the system is financially feasible at this size. If subsidy is available, the difference between the system costs and the sum of the affordable payments over the system lifetime can be allocated to subsidize the system.

5.2 Case Study Application and Results

This approach is applied within the context of Sierra Leone. The most popular appliance combinations are first determined over each spatial data subset. Community load profiles representing a synthetic village in each region are then generated using stochastic bottom-up methods and appliance combinations as inputs. Results are compared to tier-based and non-spatially specific demand

Table 5.2: Questions relevant to electrical appliances or household location from the Sierra Leone 2017 Multiple Indicator Cluster Survey.

Code	Topic
HH6	Urban/rural
HH7	Region
HH7a	District
HC7b	Radio ownership
HC8	Electricity access
HC9a	Television ownership
HC9b	Refrigerator/freezer ownership
HC9c	Iron ownership
HC9d	Fan ownership
HC11	Computer ownership
HC12	Mobile telephone ownership

Table 5.3: Ranking of appliances in the 2017 Multiple Indicator Cluster Survey for Sierra Leone from lowest to highest access level[1].

Rank	Appliance	Tier	E_{min} (kWh/year)	P (W)
1	Mobile telephone (charging)	1	1.5	2
2	Radio	1	1.5	4
3	Television	2	14.6	40
4	Fan	2	29.2	40
5	Computer	2	31	50
6	Iron	4	120.5	1100
7	Refrigerator/Freezer	4	657.0	300

estimation alternatives. Finally, affordability checks are undertaken to understand the potential for energy poverty.

5.2.1 Appliance Analysis Results

Data from Sierra Leone's 2017 MICS survey are analyzed to understand spatial variance in the combinations of appliances owned by households. Specific questions in this survey on the locations of households, their electricity status, and their appliances are listed in Table 5.2. As landlines were rarely owned based in the MICS data (i.e. only 1.5% of households with electricity access owned a landline), reflecting the tendency of LMICs like Sierra Leone tend to technologically "leap-frog" directly to mobile phones [278], these are excluded from the analysis.

[1]Tiers, power ratings and minimum energy consumption values from Table 6.13 in [189].

These appliances are ranked in access level order as shown in Table 5.3. As the computer is not considered in the MTF appliance definition [189], it is placed in Tier 2 following Mullen [279]. While MICS does not specify whether computer ownership refers to laptops or desktops, it is assumed here that all computers owned by households for personal use are laptops. As such, their energy consumption is taken as 31 kWh/year from [280] and their rated power is taken at 50 W [281].

Each possible combination of these appliances is generated. With $n = 7$ and $r = 1 : 7$, the total number of combinations is:

$$n_{combs} = (\sum_{r=1}^{7} {}_7C_r) + 1 = (\sum_{r=1}^{7} \frac{7!}{r!(7-r)!}) + 1 = 128$$

The number of electrified households owning each combination of appliances is calculated over regional, district, urban and rural data subsets. Only 79 of the possible appliance combinations are found to be represented across the whole dataset. Looking at rural and urban subsets of the data, only 42 different appliance combinations are represented in the rural subset while 76 are represented in the urban subset. Urban households have a higher diversity of appliance combinations, perhaps representing better access to appliances or differing priorities.

The ten most prevalent combinations are identified in each spatial data subset, and are shown in Table 5.4. Additionally, the combinations are ranked by their frequency within the top ten combinations of the regional and district subsets. For instance, if a combination is within the top ten combinations for 12 of 14 districts, it is assigned a frequency score of 12 for districts, and then ranked compared to the frequency score of other combinations. Observing the results, the same top eight combinations appear across rural, urban, ranked district, and ranked regions lists in different orders, namely:

- Mobile phone
- Mobile phone, radio
- Mobile phone, television
- Mobile phone, radio, television

- Mobile phone, radio, television, fan
- Mobile phone, radio, television, fan, refrigerator/freezer
- Mobile phone, radio, television, fan, iron, refrigerator/freezer
- Mobile phone, radio, television, fan, computer, iron, refrigerator/freezer

These eight combinations also represent either six or seven of the top eight combinations across each regional data subset as well as in the rural subset. However, these lists also contain the following three combinations amongst their top eight combinations:

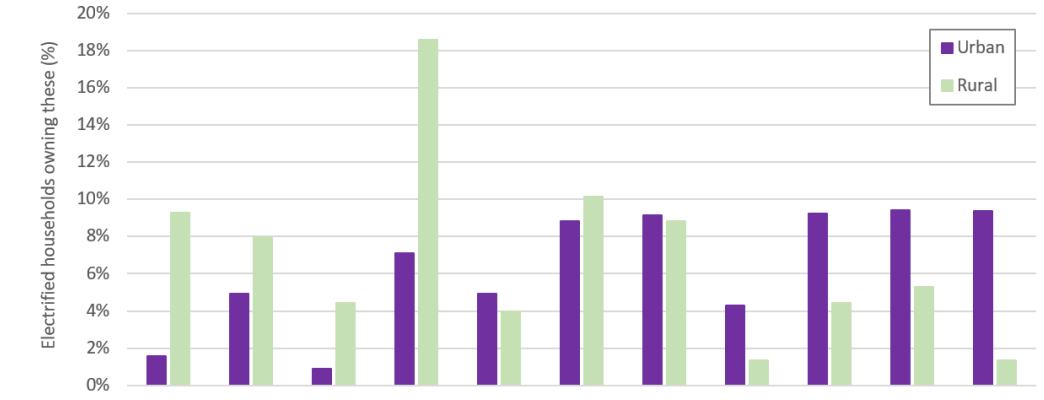
- None (no appliances)
- Radio
- Mobile phone, radio, television, refrigerator/freezer

As there are more urban than rural electrified households in the Sierra Leone 2017 MICS data, it is unsurprising that the rural combinations deviate from the other subsets. However, as this thesis tackles system design for rural electrification, demands important to rural users must be prioritized. These three combinations are therefore added to the previous list, resulting in a list of 11 popular combinations which are studied in further detail and used to compare appliance ownership in different areas. These 11 combinations represent 70% of all electrified households in the dataset. The remaining 30% of households have other appliance combinations which are less popular or only present in much smaller quantities across the dataset.

These combinations are ordered by access level, and their relative prevalence is visualized across urban and rural (Figure 5.5) and regional (Figure 5.6) data subsets. Each subset is also visualized on an individual graph in Figure 5.7.

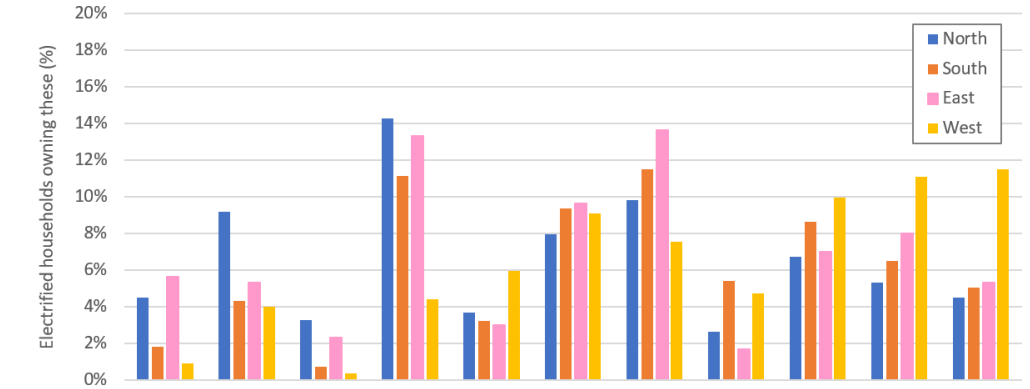
Urban and Rural Combinations

There is a higher prevalence of "None" (i.e. electrified homes owning none of the appliances tracked in MICS) amongst rural households than urban households. One could theorize that this is a result of rural households gaining a connection but



Mobile		X		X	X	X	X	X	X	X	X
Radio			X	X		X	X	X	X	X	X
Television					X	X	X	X	X	X	X
Fan							X		X	X	X
Computer											X
Iron										X	X
Fridge								X	X	X	X

Figure 5.5: Prevalence of ownership of appliance combinations in urban and rural electrified homes in Sierra Leone. Combinations are arranged in increasing access order.



Mobile		X		X	X	X	X	X	X	X	X
Radio			X	X		X	X	X	X	X	X
Television					X	X	X	X	X	X	X
Fan							X		X	X	X
Computer											X
Iron										X	X
Fridge								X	X	X	X

Figure 5.6: Prevalence of ownership of appliance combinations in electrified homes in each region of Sierra Leone. Combinations are arranged in increasing access order.

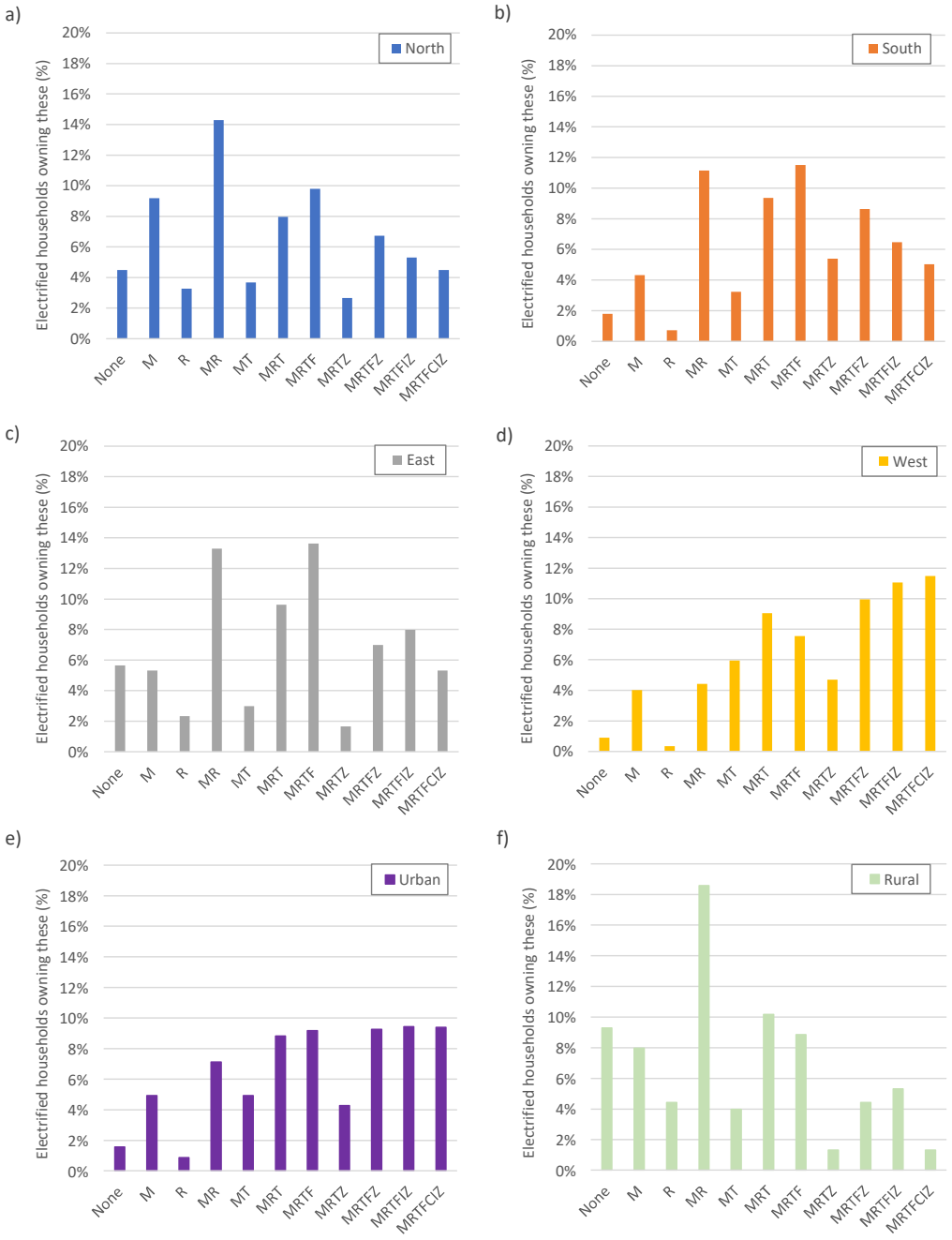


Figure 5.7: Prevalence of ownership of appliance combinations in electrified homes for the following data subsets for Sierra Leone: a) North, b) South, c) East, d) West, e) Urban, and f) Rural. Combinations are arranged in increasing access order. Appliances are represented by their first letter, aside from Refrigerator/Freezer, represented by "Z".

Table 5.4: Top ten most popular appliance combinations across different data subsets, and ranked in terms of frequency across provincial and district top-ten lists. Appliances are represented by their first letter, aside from Refrigerator/Freezer, represented by "Z".

Rank	All	Urban	Rural	North	South	East	West	Frequency (Provinces)	Frequency (Districts)
1	MRTF	MRTFIZ	MR	MR	MRTF	MRTF	MRTFCIZ	MR	MR
2	MRTFIZ	MRTFCIZ	MRT	MRTF	MR	MR	MRTFIZ	MRTF	MRTF
3	MRT	MRTFZ	None	M	MRT	MRT	MRTFZ	M	M
4	MRTFZ	MRTF	MRTF	MRT	MRTFZ	MRTFIZ	MRT	MRT	MRT
5	MRTFCIZ	MRT	M	MRTFZ	MRTFIZ	MRTFZ	MRTF	MRTFZ	MRTFZ
6	MR	MR	MRTFIZ	MRTFIZ	MRTZ	None	MT	MRTFCIZ	MRTFCIZ
7	M	MT	MRTFZ	None	MRTFCIZ	M	MRTZ	MT	MT
8	MT	M	R	MRTFCIZ	M	MRTFCIZ	MR	MRTFIZ	MRTFIZ
9	MRTZ	MRTZ	MT	MT	MRF	MRF	M	MRF	MRTFCZ
10	MTF	MTF	MRF	R	MRTFCZ*	MR	MTF	MRTZ*	MRF

*These combinations were tied for tenth place. Where there was a tie, the higher-access-level combination is listed.

then being unable to purchase appliances due to low availability or affordability. It has previously been seen in rural electrification programmes that low energy usage can follow after household connection (e.g. in Kenya [282]). It should be investigated whether affordability, accessibility, cultural norms, or some other factor are encouraging rural Sierra Leoneans not to acquire appliances following connection.

Radio ownership is much more prevalent in rural areas than urban areas. This can be seen in the higher prevalence of the "Radio" and "Mobile, radio" appliance combinations in rural than urban households from Figure 5.5. This may speak to the rural population gaining more of their information and entertainment from the radio than from other sources (e.g. television, internet). It may also reflect lower literacy amongst rural populations, who may find it harder to access written information and news via mobile phone or online and who instead gain it through radio. It may also indicate that radios are sometimes used by farmers in the fields; perhaps the higher radio prevalence is related to the higher agricultural population in rural areas who listen to the radio as they work.

Appliance combinations that indicate higher levels of access are more prevalent amongst urban households than rural households. This may reflect the wealth discrepancy between urban and rural homes; urban households are more likely to have the means to purchase more expensive, higher-access-level appliances. They are also more likely to have the means to pay for more energy and actually make use of these appliances. It may also reflect differences in electrical capacity and quality amongst urban and rural energy access systems. Smaller systems like SHS often cannot sustain power-hungry appliances (e.g. refrigerators, irons), but grid connection can, and is much more common in urban than in rural areas of Sierra Leone.

The relationship between rurality, poverty, and grid reliability must also be considered. Richer areas are reported to have fewer power cuts and less load shedding than poorer areas in many LMICs (e.g. as shown in Accra, Ghana [283]). It can be hard for those experiencing more frequent cuts to climb the ladder to

higher-access appliances, which can become damaged by voltage inconsistencies and other electricity quality issues.

Regional Combinations

From Figures 5.7d and 5.7e, it is evident that the Western region appliance ownership pattern closely reflects the urban appliance ownership pattern. This makes sense, as a large proportion of the urban population in Sierra Leone lives in the Western region, specifically in Freetown and the surrounding area.

Given this, one might consider consider whether other regional patterns truly represent geographic variation, or simply a weighted combination of the urban and rural appliance patterns for that region. This was investigated with MICS data and proven to be untrue. The percentage of electrified urban and rural homes in each region were used to create weighted averages of the prevalence of each appliance combination. These did not match the regional prevalences, showing that there is more than a weighted combination of urban and rural dynamics at play. An example showing the difference for the Eastern region is provided in Figure 5.8

Looking at the North, East, and South regions in Figures 5.7a-c, there are other non-negligible variations in the popularity of different appliance combinations. For instance, households in the North are much more likely than those in other regions to own just a radio. These differences are considered during demand estimation.

5.2.2 Demand Estimation Setup

To test how these spatially varying appliance data translate into load profiles, demand estimates were generated for a synthetic communities representing different data subsets and approaches. A community size of of $n = 200$ dwellings was used to represent a synthetic rural off-grid village in each region. Demand was estimated over a month (i.e. 31 days) at minute-level resolution, using the month of January as a case study. The demand estimation approach described above based on spatially specific MICS data estimation is compared to demand estimates based on achieving a defined MTF tier of access.

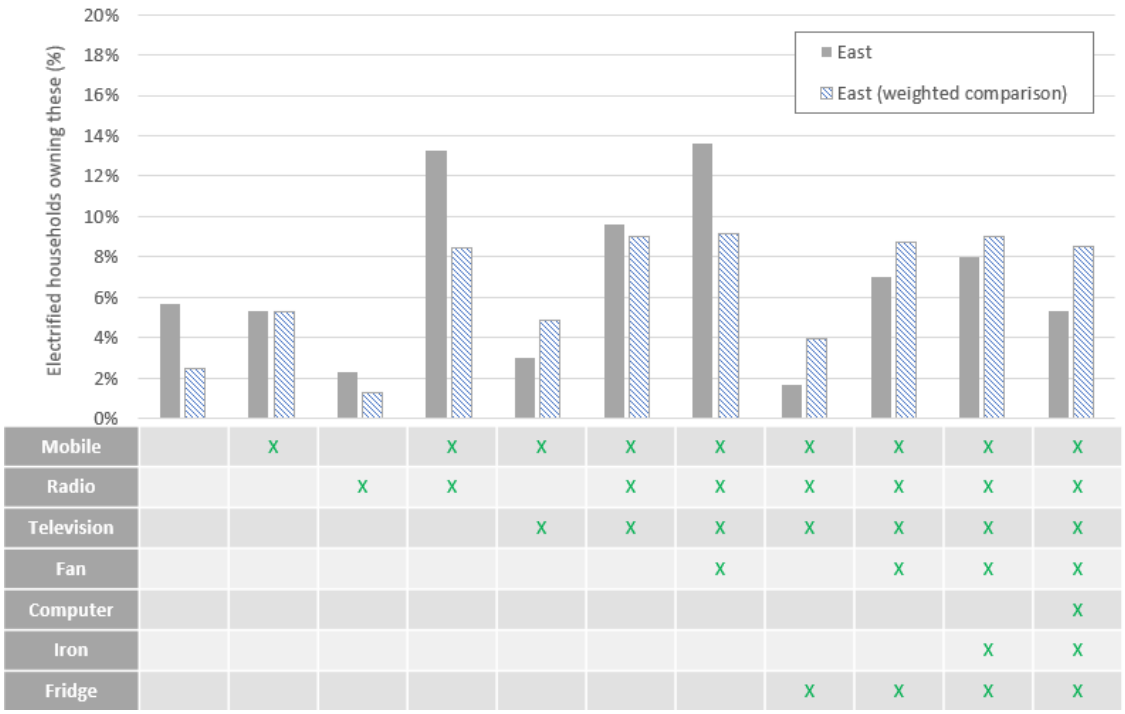


Figure 5.8: Prevalence of ownership of appliance combinations in electrified homes for the Eastern region of Sierra Leone compared with a weighted average of Urban and Rural prevalences for this region. The regional data do not reflect a mere weighted average of rural and urban trends, indicating more nuanced geographic trends.

MICS-Based Spatially Specific Demand Estimation

Using the MICS appliance ownership data, demand was estimated for a synthetic village in the North, South, East, and West regions of Sierra Leone. Additionally, a synthetic village was modelled after the rural and urban MICS data subsets, as well as based on the entire country-wide dataset.

For other spatially specific data inputs (e.g. sunrise, sunset, weather), specific locations in each region were selected as data retrieval landmarks: Freetown for the West, Kenema for the East, Kabala for the North, and Pujehun for the South[2]. Sunrise and sunset times from NOAA [277] and weather data from Renewables.Ninja [54] were used to configure appliance usage windows and refrigerator demand cycles, as shown in Tables 5.5 and 5.6 respectively. Appliance P values were taken from

[2]Note that the synthetic communities are not intended to represent these actual towns, which vary in size and relative urbanity.

the MTF[3] [189]. For input parameters where no data was available (i.e. $frames$, t_{tot}, t_{min}, m lights per household), estimates were sourced from a Sierra Leonean expert with significant experience in the energy sector and held constant through all demand estimation experiments. Note that for all appliances besides lights, one occurrence of each appliance was modelled per household. Specific input parameters for each appliance are explained in detail in Appendix C and shown in Table 5.7.

Tier-Based Estimation

Estimates made using MICS data inputs were compared with estimates based on MTF tier definitions [189]. As the appliances considered in the MICS profiles represented Tiers 1, 2, and 4 (as per Table 5.3), load profiles for these MTF tiers were created. The appliances included were modelled based on the power levels and usage times defined in the MTF – see Table 6.13 in [189]. As the weekly frequency of use for each appliance is not specified in the MTF, all are assumed to be used each day. While the refrigerator is defined as used at a rated power of 300 W for 6 hours [189], this is interpreted as a duty cycle at 300 W power, 25% of the time (i.e. 300 W for 5 minutes, 0 W for 15 minutes) all day. The RAMP framework was still used for demand estimation based on these inputs, and unless specified, all remaining parameters were held constant. Inputs for each tier are shown in Tables 5.8, 5.9, and 5.10.

5.2.3 Demand Estimation Results

The resulting average load profile and range is shown for each synthetic community in Figures 5.9 and 5.10. All of these load profiles follow roughly the same pattern: a peak in the evening from lighting and entertainment devices, steady usage overnight from fans and outdoor lights, and lower use during the day. There is a small bump at mid-day resulting from fan usage on weekends; the difference between weekday and weekend average profiles is shown in Figure 5.11. This also shows a higher peak in the evenings on weekends due to the impact of iron usage on weekend

[3]Where two possible values were provided, the higher was used to ensure that the resulting demand profile would err towards overestimation.

Table 5.5: Average sunrise and sunset times for the month of January in Freetown, Kenema, Pujehun, and Kabala [277].

Town	**Sunrise**	**Sunset**
Freetown (West)	7:12	18:53
Kenema (East)	7:02	18:46
Pujehun (South)	7:04	18:49
Kabala (North)	7:06	18:45

Table 5.6: Refrigerator duty cycle details used in the case study application to account for daily temperature variation.

Cycle	$\boldsymbol{frame}$	$\boldsymbol{P_1}$	t_1	$\boldsymbol{P_2}$	$\boldsymbol{t_2}$
Heavy	Sunrise-Sunset	300	20	5	10
Light	Sunset-Sunrise	300	10	5	20

Table 5.7: Appliance-level input parameters and data for bottom-up stochastic demand profile generation in the case study application. WD = Weekday, WE = Weekend, occ. = occasional; unless specified, assume all usage is both on weekday and weekends.

Appliance	$\boldsymbol{n}$	$\boldsymbol{P(W)}$	$\boldsymbol{\delta_P}$	$\boldsymbol{cycles}$	$\boldsymbol{frames}$	$\boldsymbol{\delta_f}$	$\boldsymbol{t_{tot}}$	$\boldsymbol{\delta_t}$	$\boldsymbol{t_{min}}$	$\boldsymbol{fixed}$	$\boldsymbol{occ.}$
Indoor light	5	12	0	0	Sunset-Midnight	0.35	120	0.2	10	no	100%
Outdoor light	4	12	0	0	Sunset-Sunrise	0.2	540	0.2	30	yes	100%
Mobile charger	1	2	0	0	10PM-2AM	0.35	120	0.2	20	no	100%
Radio	1	4	0	0	Sunset-Midnight	0.35	90	0.2	30	no	80%
Television	1	40	0	0	Sunset-Midnight	0.35	90	0.2	30	no	80%
Fan	1	40	0	0	6PM-Sunrise & 12-2PM (WE)	0.2	720 (WD), 840 (WE)	0.2	60	no	100%
Computer	1	50	0	0	Sunset-Midnight	0.35	90	0.2	30	no	80%
Iron	1	1100	0.3	0	Sunset-Midnight (WE)	0.35	30	0.2	5	no	80%
Refrigerator	1	300	0	2	Constant	0	1440	0	1440	yes	100%

Table 5.8: Tier 1 input parameters and data for tier-based demand estimation in the case study application [189].

Appliance	n	$P(W)$	δ_P	$cycles$	$frames$	δ_f	t_{tot}	δ_t	t_{min}	$fixed$	$occ.$
Task light	4	1	0	0	Sunset-Midnight	0.35	60	0.2	10	no	100%
Mobile charger	1	2	0	0	10PM-2AM	0.35	120	0.2	20	no	100%
Radio	1	2	0	0	Sunset-Midnight	0.35	120	0.2	30	no	100%

Table 5.9: Tier 2 input parameters and data for tier-based demand estimation in the case study application [189].

Appliance	n	$P(W)$	δ_P	$cycles$	$frames$	δ_f	t_{tot}	δ_t	t_{min}	$fixed$	$occ.$
Task light	4	2	0	0	Sunset-Midnight	0.35	60	0.2	10	no	100%
General light	4	12	0	0	Sunset-Sunrise	0.2	60	0.2	30	no	100%
Mobile charger	1	2	0	0	10PM-4AM	0.35	240	0.2	20	no	100%
Radio	1	4	0	0	Sunset-Midnight	0.35	240	0.2	30	no	100%
Television	1	20	0	0	Sunset-Midnight	0.35	120	0.2	30	no	100%
Fan	1	20	0	0	6PM-Sunrise	0.2	240	0.2	60	no	100%
Computer	1	50	0	0	Sunset-Midnight	0.35	90	0.2	30	no	100%

Table 5.10: Tier 4 input parameters and data for tier-based demand estimation in the case study application [189].

Appliance	n	$P(W)$	δ_P	$cycles$	$frames$	δ_f	t_{tot}	δ_t	t_{min}	$fixed$	$occ.$
Task light	4	2	0	0	Sunset-Midnight	0.35	120	0.2	10	no	100%
General light	4	12	0	0	Sunset-Sunrise	0.2	120	0.2	30	no	100%
Mobile charger	1	2	0	0	10PM-4AM	0.35	240	0.2	20	no	100%
Radio	1	4	0	0	Sunset-Midnight	0.35	240	0.2	30	no	100%
Television	1	40	0	0	Sunset-Midnight	0.35	120	0.2	30	no	100%
Fan	1	40	0	0	6PM-Sunrise	0.2	720	0.2	60	no	100%
Computer	1	50	0	0	Sunset-Midnight	0.35	90	0.2	30	no	100%
Iron	1	1100	0.3	0	Sunset-Midnight	0.35	20	0.2	5	no	100%
Refrigerator	1	300	0	2	Constant	0	1440	0	1440	yes	100%

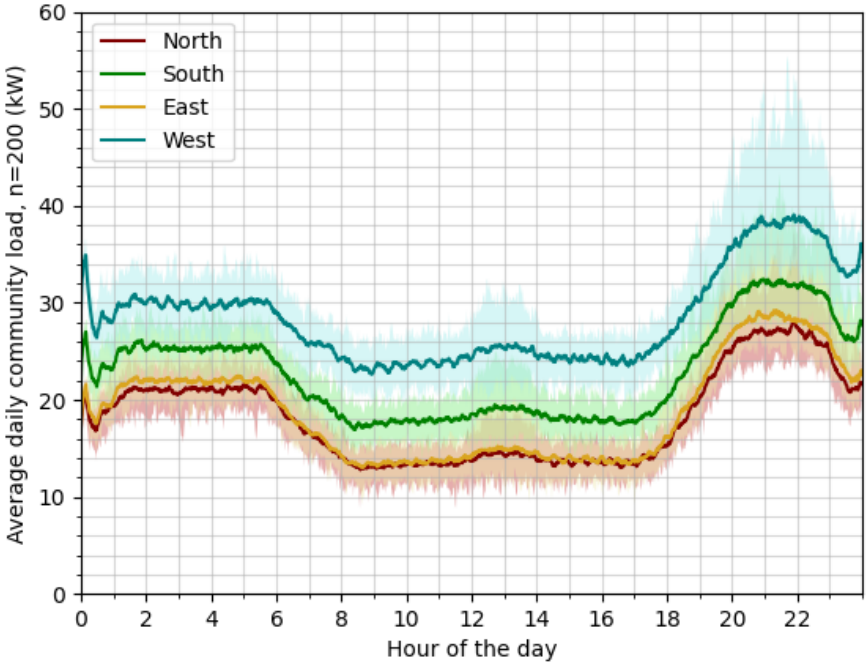


Figure 5.9: Average daily load profiles for the synthetic communities of 200 households generated using appliance combinations for each region of Sierra Leone from MICS data. The full range of data for each case is shown in the lighter semi-transparent band.

evenings. The similarity between the load profile shapes is expected, given the same possible windows of operation and expected durations of operation modelled across all experiments. The differences expected are indeed not in terms of general load shape, but in terms of the amplitude. Note that there are some artifacts near midnight at the start and end of the profile, as the stochastic framework simulates each day individually and does not account for continuity between days.

Looking across regions, the load profile for the Western community is the highest consistently throughout the day and has the highest peak power. This is expected, as the Western province is highly urban and had more households with high-access and high-powered appliance combinations. The load profiles for the Northern and Eastern communities are very similar; the Southern community load profile lies between these and the Western community. There is an average gap of 10.3 kW between the average Northern and Western community load profiles throughout the day, ranging between 7.9 and 14.3 kW depending on the time. The average

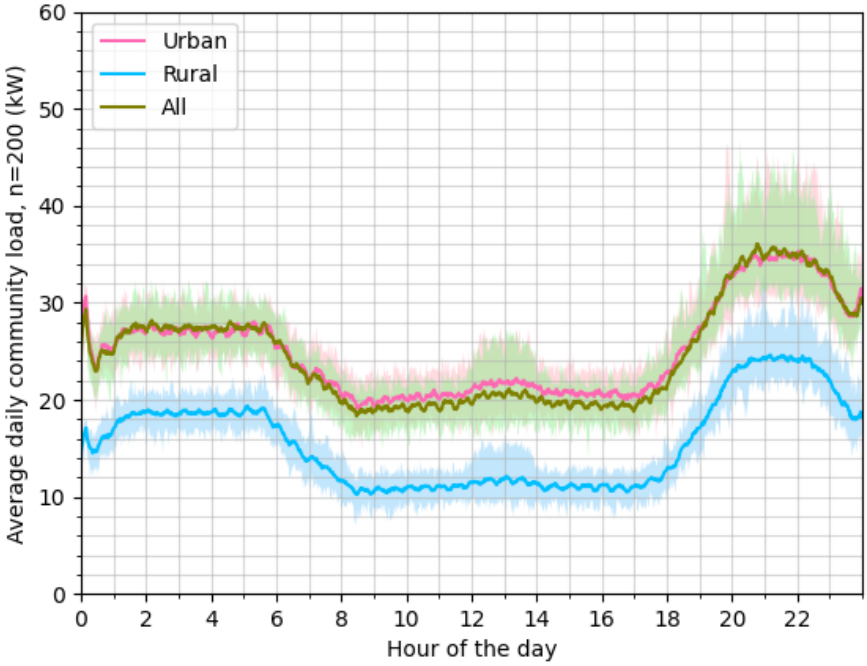


Figure 5.10: Average daily load profile for the synthetic communities of 200 households generated using appliance combinations from urban and rural subsets of Multiple Indicator Cluster Survey data for Sierra Leone. The full range of data for each case is shown in the lighter semi-transparent band.

Southern community profile lies 5.8 kW on average below the average Western community profile, ranging between 3.9 and 8.4 kW below depending on the time.

Considering the urban and rural community load profiles in Figure 5.10, it is immediately clear that the demand profile generated from the entire dataset ("All") is dominated by urban respondents. This makes sense; there were far more urban households than rural households with electricity access in the MICS dataset, as the grid in Sierra Leone does not extend to many rural communities. It is also evident that the load profile for the rural synthetic community is far lower than the urban-only profile, with an average 9.5 kW discrepancy throughout the day, ranging between 7.4 kW and 13.6 kW. This result aligns with the intuitive expectation that poorer rural populations have lower power requirements than more affluent urban populations with better access to higher-powered appliances and reliable grid connections.

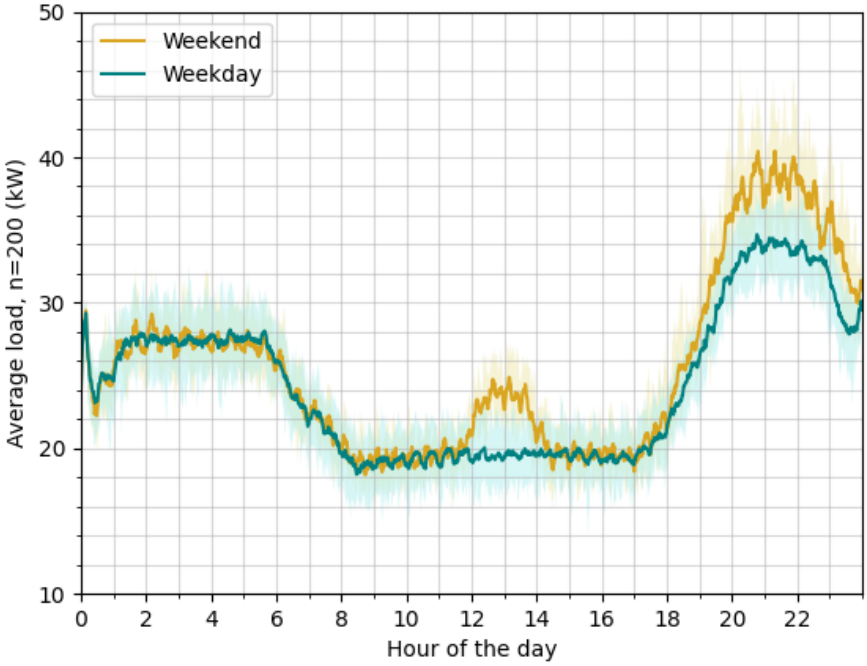


Figure 5.11: Average daily load profile on weekends and weekdays for a synthetic community of 200 households based on appliance combinations across the entire Multiple Indicator Cluster Survey dataset. The full range of data for each case is shown in the lighter semi-transparent band.

Daily energy consumption is calculated across the synthetic communities for each region. Results are shown in ascending energy order in Figure 5.12. The rural community uses the least energy each day, followed by North and East communities, then South, Urban, and West. The energy results are also shown in Table 5.11, alongside their standard deviation (σ) and the coefficient of variance (CV). This table also includes the average energy usage at day and at night, calculated using the sunrise and sunset times for each region, and using the average of these for the urban, rural, overall, and tier-based villages. More energy is used at night than during the day in all cases. This of course has implications for the design of solar stand-alone or autonomous mini-grid systems, two common electricity access technologies in Sierra Leone. As these cannot produce energy at night, when the majority of the energy use occurs according to these results, storage or back-up (i.e. diesel) generation would be necessary to ensure energy needs can be

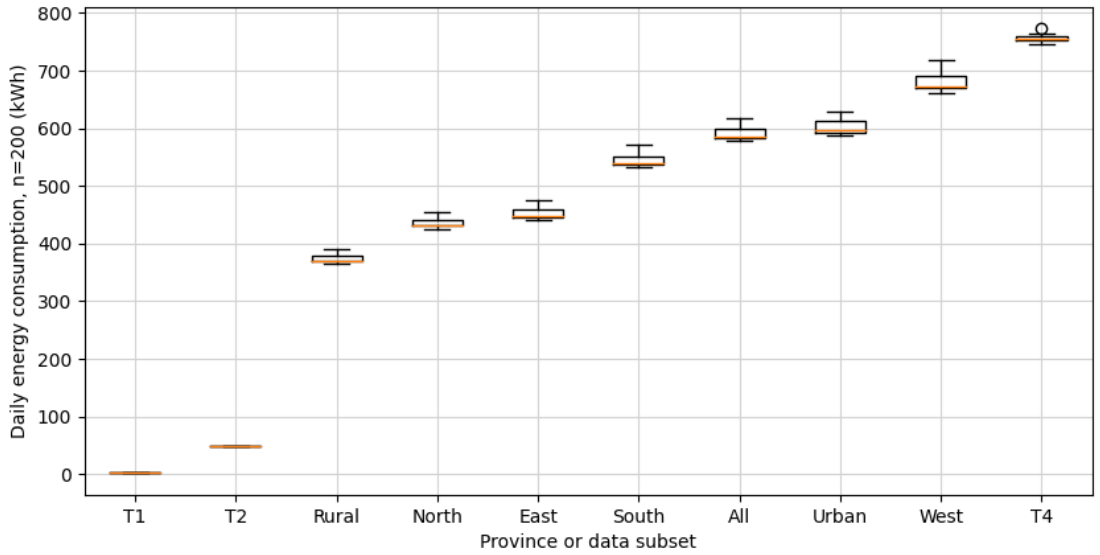


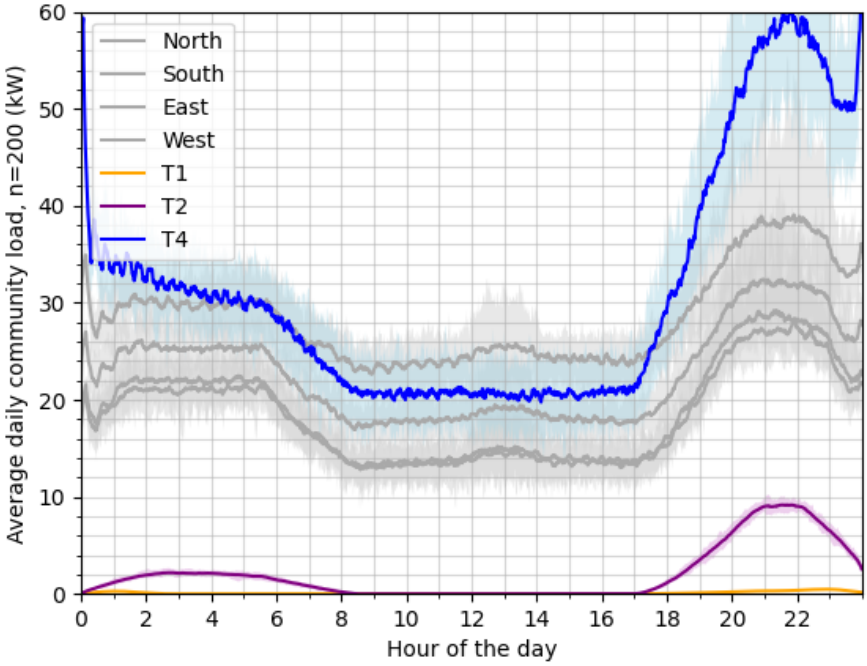
Figure 5.12: Daily energy consumption for each synthetic community of 200 households representing different Multiple Indicator Cluster Survey data subsets in Sierra Leone and different energy access tiers, in ascending energy use order.

Table 5.11: Daily energy consumption results in each synthetic village for the case study application in Sierra Leone. From top to bottom: average energy consumption; average energy used during day and night hours; minimum, median, and maximum daily energy consumption; standard deviation (σ); and coefficient of variance (CV).

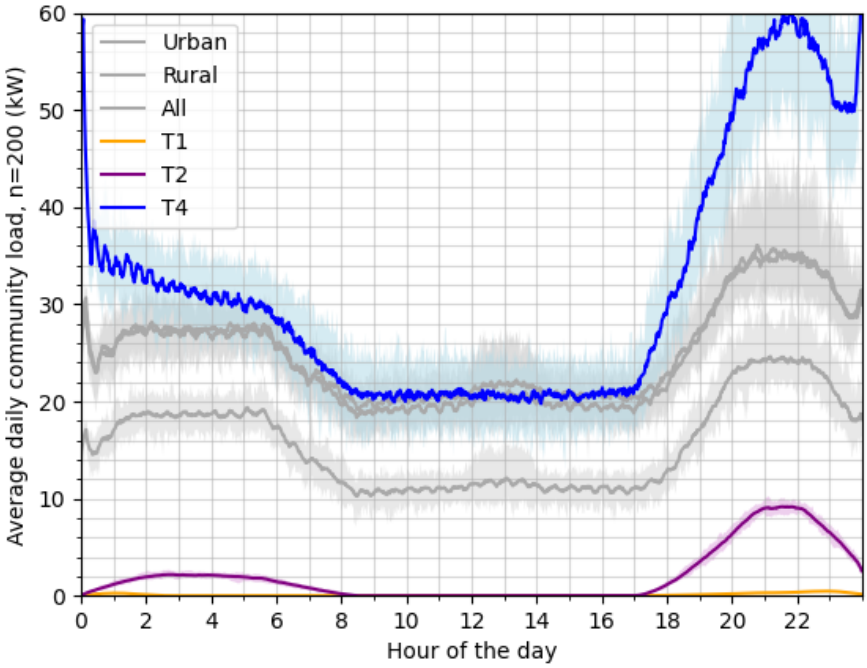
	N	**S**	**E**	**W**	**Urb**	**Rur**	**All**	**T1**	**T2**	**T4**
E_{avg}	437	545	453	683	603	374	592	2.00	49.1	756
E_{day}	164	218	169	288	247	135	235	0.07	2.46	258
E_{night}	272	327	284	395	356	240	357	1.93	46.6	498
E_{min}	426	533	442	660	588	366	579	1.98	48.8	747
E_{med}	432	540	447	674	597	370	586	2.00	49.1	756
E_{max}	455	572	475	718	630	392	618	2.01	49.4	773
σ	9.26	11.93	11.91	18.43	14.53	7.91	13.67	0.01	0.16	5.60
CV	0.021	0.022	0.026	0.027	0.024	0.021	0.023	0.005	0.003	0.007

met. This speaks to the marriage of solar and storage which is often necessary in green off-grid electrification to avoid night-time fossil fuel consumption. Finally, note that the CV values in Table 5.11 for tier-based profiles are smaller than for the MICS-based profiles, reflecting the increased intra-community variance when different appliance combinations are considered.

Comparing the Urban and Western communities, the Western profile actually has a greater amplitude throughout the day. This is interesting, since as previously discussed, most people Western residents are urban. One can hypothesize that urban



(a) Regions



(b) Urban and rural

Figure 5.13: Comparison of regional and urban/rural synthetic community load profiles for Sierra Leone with tier-based community load profiles. The full range of data for each case is shown in the lighter semi-transparent band.

residents in the Western province have better access to higher-powered appliances and/or higher income than the average urban resident in Sierra Leone. This could increase their usage compared to other urban residents. People in Western province are also likely to have better grid access, which may encourage them to buy more high powered appliances than urban households in other areas with SHS power or mini-grids which may offer lower power ratings.

These load profiles are compared to profiles generated based on tier definitions in Figure 5.13. Most load profiles fall between Tier 2 and Tier 4. The Urban and Western profiles are near or above the Tier 4 profile at times, and below it at others.

5.2.4 Affordability Check

Household expenditure data from the Sierra Leone IHS [284] is used to check demand estimates against the risks of energy poverty. This income and expenditure survey, which serves as Sierra Leone's LSMS, is openly available online through Statistics Sierra Leone [285]. Average household consumption expenditure in each region of Sierra Leone is shown in Table 5.12, alongside a maximum budget that can be allocated to energy each month before energy poverty risk (i.e. 10% of total expenditure) and a maximum tariff per kWh to meet estimated demands without energy poverty risk. Costs are taken in Leones (Le) from the 2018 IHS, so the 2018 average conversion rate from Leones to USD (i.e. 7,712 Le/USD [286]) is used in Table 5.12.

Table 5.12: Average monthly nominal consumption expenditure and estimated energy consumption per household, as well as the electricity tariff required per kWh for energy poverty prevention (i.e. such that energy does not exceed 10% of total expenditure).

Region	Monthly expenditure (Le)	Energy budget (Le)	Electricity use (kWh)	Tariff required (Le/kWh)	Tariff required ($/kWh)
North	334,001	33,400	67.7	493	0.06
South	325,231	32,523	84.4	385	0.05
East	356,487	35,649	70.2	508	0.07
West	781,640	78,164	105.9	738	0.10

The resulting required tariffs are on a reasonable order of magnitude, but generally speaking rather low. For instance, in the United States, average residential electricity prices in 2020 were \$0.13/kWh [287]; all of these tariffs fall below that. To understand how they compare to actual tariffs in-country, these results were compared to existing electricity tariffs in Sierra Leone. These vary amongst consumer groups and infrastructure types. For on-grid electrification, residential customers in Sierra Leone pay either a "social" tariff of 560 Le/kWh (i.e. \$0.07/kWh) if they use less than 25 kWh/month, or a "normal" tariff of 1,600 Le/kWh (i.e. \$0.21/kWh) if they use more than that [288]. Residential tariffs on mini-grids are much more variable [289], ranging from 1,400 Le/kWh (i.e. \$0.18/kWh) for basic residential service on the Segbwema grid to 6,715 Le/kWh (i.e. \$0.87/kWh) for residential service from Power Leone WP 2. The maximum tariffs identified to avoid energy poverty risk based on demand estimates fall below the on-grid social tariff for North, South, and East, and between the social and normal tariffs for the West. Were the normal on-grid tariff to apply across all regions – as it would, based on the average estimated demand in each being above 25 kWh/month – the resulting average electricity charge per month would range between 22% and 42% of average income depending on the region considered. The mini-grid tariffs equally appear unrealistic to the level of usage estimated here; for instance, in the Northern region, satisfying estimated demands at a tariff of 6,715 Le/kWh (i.e. \$0.87/kWh) would represent 33% of total household expenditure. One can assume mini-grids are designed for lower tiers of access, with a capacity limit close to Tier 1 or 2 use, and not designed to meet eventual demands experienced by wealthier grid-connected customers. A principal reason for the high proportion of average income which would be required to meet these demands is, as previously discussed, the fact that most people in Sierra Leone with existing electricity access are comparatively wealthy.

5.3 Discussion

These results prompt interesting observations not only on the spatial variance of demands and the utility of the proposed methodology, but on patterns within

appliance acquisition and a number of areas for future study.

5.3.1 Spatial Variance

The appliance combinations and demand estimates produced using the MICS data varied regionally and across rural and urban divides, proving spatial demand variance. This has implications for the design of electricity access systems. Consider the case of designing a mini-grid based on solar and battery storage in a rural Northern community of $n = 200$ households. If demand was estimated to enable system sizing based on appliance ownership across the whole MICS dataset, the resulting system design would have oversized generation and storage. For instance, the worst-case estimated peak power was simulated at 33 kW for such a community with $n = 200$ based on the rural MICS subset, compared 47 kW if the whole MICS dataset is used. If designing for 100% reliability, then the system will be designed to provide a peak power which overshoots the rural context by at least 40%. This oversizing could provide useful room to grow demands or could trigger a "failure to launch" if it rendered the energy produced by the system completely unaffordable.

Storage oversizing is particularly important to consider in this example, as the peak load occurs in the evening after sun-down when solar PV cannot generate energy. To quantify this, consider the daily energy which must be used after nightfall (i.e. which must be satisfied by battery capacity alone), as presented in Table 5.11. In the synthetic community modelled based on rural MICS data, the average daily energy consumed at night was 240 kWh. Meanwhile, the synthetic community modelled after the entire dataset in aggregate consumed 357 kWh on average overnight. Comparing 240 kWh to 357 kWh, designing based on the overall case would create a 49% overshoot in energy storage sizing. Again, this may not be a bad thing in terms of technical operation, simply allowing for more days of battery-only autonomy in the case of poor weather. Indeed, this could improve the resilience of the system and allow for future load growth. However, it could render electricity prohibitively expensive, causing a failure to launch (i.e. low system

uptake). The tradeoff between reasonable system cost, ability to meet future load, and resilience will be further explored in Chapter 6.

5.3.2 Comparison With MTF

By using MICS appliance ownership data to configure stochastic modelling, demand estimates better integrate intra-community diversity than the tier-based approach. Designing for a certain tier of access across *all* households erases the diversity of needs, life-stages, ability-to-pay, and other factors impacting electrical demand *amongst* households. This is evident in the results, as stochastic estimates based on a uniform tier of access across a community shows lower variance compared to the developed approach using MICS data (see CV in Table 5.11).

The demand estimates configured with MICS data approximate a Tier 3 community, as shown in Figure 5.13, despite the fact that no Tier 3 appliances (i.e. washing machines, food processors) are considered in the analysis. Indeed, no Tier 3 appliances are studied at all in the 2017 MICS in Sierra Leone. This shows one issue with the MTF definitions in practice; when applied at the community level, a distribution of households at different access tiers is more likely to be present than a single energy usage mode. While a constellation of access levels can emulate a certain tier of behaviour on average, it includes variance that a uniform tier assumption does not. It seems that the general definition of Tier 3 in the MTF may not be useful in the context of Sierra Leone. This prompts the question of whether this or other tiers are not useful in other particular country contexts. Perhaps country-specific definitions appliance definitions for each tier are needed.

This analysis highlights issues with the expectation that everyone should reach Tier 5 access. In practice, even amongst the wealthiest people in Sierra Leone, there are a diversity of appliance ownership patterns. Not everyone owns high-powered, high-tier appliances like irons and refrigerators. If it is assumed that households all will want and acquire high-powered white goods such as freezers when estimating demands for electrification, the shape of the resulting load profiles changes dramatically and becomes unrealistic. As discussed in Chapter 2, the expectation of

eventual uniform ownership of "white goods" (e.g. dishwashers, washing machines) is a primarily white and western phenomenon; appliance ownership aspirations actually vary based on local and familial wealth, class, race, and religion [150]. Expecting everyone in Sierra Leone or any LMIC to eventually attain Tier 5 might indeed be a neo-colonial or assimilationist goal which is not actually desired at the community level.

One additional peculiarity of the tier-based estimation approach is the definition of the refrigerator provided in the MTF [189]: that is, a usage of 6 hours a day at 300 W without mention of a duty cycle. It is up to the reader to interpret that this is likely to indicate a duty cycle with a 25% "on" window. Put in the hands of someone with less electrical literacy, this could be misinterpreted as six consecutive hours of usage, which would grossly misrepresent the effect of a refrigerator on household and community load profiles.

Give these issues, it can be questioned whether designing an electrical system such that all households attain a given tier of access is actually useful. This approach, where applied, should be reviewed in terms of its suitability to purpose.‘

5.3.3 Appliance Acquisition Pathway

Observing the 11 most popular appliance combinations in the MICS dataset for Sierra Leone as outlined in Section 5.2.1, one can observe a potential trend in appliance acquisition. These combinations, presented in access-level-order at both appliance and combination levels, are:

- None (lights only)
- Mobile phone
- Radio
- Mobile phone, radio
- Mobile phone, television
- Mobile phone, radio, television
- Mobile phone, radio, television, fan

- Mobile phone, radio, television, refrigerator/freezer
- Mobile phone, radio, television, fan, refrigerator/freezer
- Mobile phone, radio, television, fan, iron, refrigerator/freezer
- Mobile phone, radio, television, fan, computer, iron, refrigerator/freezer

Ordered this way, a potential appliance acquisition pathway becomes apparent. Once a household obtains an electrical connection and lighting, they may next acquire a mobile phone or a radio, eventually getting both. They then add a television, and subsequently a fan or refrigerator/freezer, eventually getting both. Next an iron is purchased, and finally a computer is acquired.

This is not necessarily the case in all households, nor is it even confirmed here. Perhaps all households in these cases acquired all of their appliances at once, and this progression is incorrect. It is just interesting that when ordered this way, a potential appliance "build-up" becomes apparent.

When considered in this order, it also becomes interesting that a computer can be considered a Tier 2 appliance based on its power consumption when it is seemingly the last appliance added. One can hypothesize that this is because it is the least affordable appliance of the set, which is the least attainable as a household moves up the access ladder. It is the most likely to be seen as a luxury despite its lower energy consumption than something like a refrigerator, which may justify its placement in a higher tier (e.g. Tier 5). Conversely, looking at energy consumption and social benefit instead of purchase cost, it can instead be argued that the social mobility provided by access to information and education via a computer is incredibly high per kWh consumed, assuming that an internet connection is available. This would make a computer incredibly efficient in terms of social benefit, perhaps far more efficient than other Tier 5 appliances. This raises the question of whether appliances should be assigned to tiers based primarily on power consumption, or whether their relative affordability, efficiency of social benefit, or other factors should also factor into their ranking.

5.3.4 Extension: Appliance Quantities

The MICS data used in the appliance combination analysis does not contain information on the quantity of each appliance owned by households. As such, this factor is omitted from the previous methods to maintain a generalized approach which can be cross-applied in other contexts with MICS data available. However, depending on the country under consideration, there may be other data sources to fill this gap. For instance, in Sierra Leone, the 2018 IHS survey can be used to analyze appliance quantities. The IHS captures appliance quantities in Questions 1 and 2 of Section L (durable goods). As in the MICS survey, IHS data are labelled by province, district, and urban or rural classification, allowing for spatially specific analysis.

To explore the importance of varying appliance quantities, the proportion of households in each region of Sierra Leone found to own *any* quantity of each appliance, which owned m quantity of the appliance for $m = 1 : 10$, is investigated over each spatial IHS subset. The results of this exploration are shown in Figure 5.14. Across all data subsets, 72% or more of households only own one of each appliance for all appliances except for mobile phones. Looking at the urban and rural subsets, it is far more common for households in urban areas to own multiple cell phones compared to urban areas (i.e. 67% versus 36% owning multiples).

Owning more than one of a given appliance does not necessarily increase overall household demands. For instance, while two refrigerators in one household are likely to consume twice as much energy as one refrigerator – as both refrigerators are likely to be always on – two televisions in one household do not necessarily mean that energy consumption by televisions is doubled. A household may own multiples of certain appliances (i.e. televisions, lights, fans) such that as household members move between rooms, they need not carry the appliance with them, but instead can turn one on and the other off. Conversely, appliances like mobile phones and computers are more likely to be used independently by individual household members, making it more likely that each additional appliance in the household will linearly scale the load from that appliance type. So, if appliance quantities were to be considered in demand estimation for Sierra Leone, only multiples of mobile phones

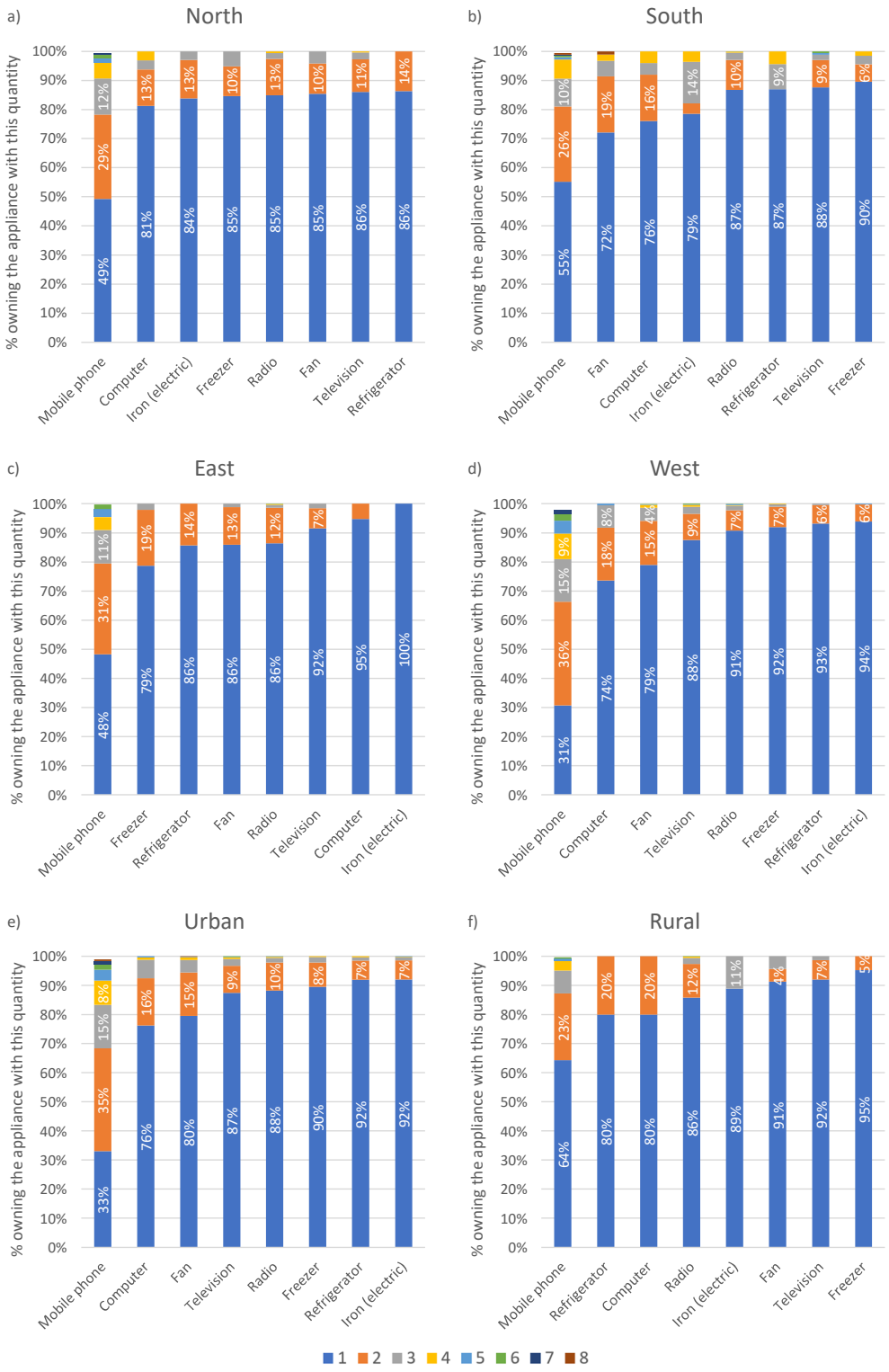


Figure 5.14: Quantities of appliances owned per household from subsets of the Sierra Leone Integrated Household Survey dataset: a) North, b) South, c) East, d) West, e) Urban, and f) Rural. Each bar represents 100% of households which own the appliance.

and refrigerators/freezers might be considered to linearly scale energy consumption in the household. Note that these appliance usage assumptions are culturally-dependent, and would need to be reviewed and adjusted depending on the context.

Additionally, the impact of owning appliance multiples on demand is proportional to the power consumption of the appliance and its expected duration of use. Appliances with higher rated power and more frequent use will have a greater impact on demand even if infrequently owned in multiples. For instance, while the proportion of households owning refrigerators which own more than one only falls within the range of 7-20% depending on the data subset considered, these are higher-powered appliances which are always on. This makes their impact different from something like irons which, while also high-powered with a small proportion of homes owning multiples, are used very occasionally for a short duration of time. While refrigerators will create a sustained amplitude increase, irons are therefore more likely to create demand peaks. The intersection of appliance quantity, rated power, and usage pattern can be considered to more accurately determine resulting demand.

5.3.5 Extension: Demand Scaling

The methodology developed in this chapter leverages the appliance ownership patterns of those with existing electricity access to project the demands of the newly electrified. However, it must be considered that those who currently have electricity access in a country like Sierra Leone with low access rates and grid coverage are usually among the relatively wealthy. Indeed, as shown in Figure 5.15, those in Sierra Leone with electricity access fall exclusively in the top three wealth quintiles. These households have more income to spend on appliances and electricity than average. As such, resulting demand estimates can be seen as inherently aspirational when applied in an off-grid community spanning lower wealth quintiles. They therefore may more realistically represent a high-end bound on anticipated electrical demand, or may represent electrical demands a number of years after connection, rather than being interpreted as average demand immediately following connection.

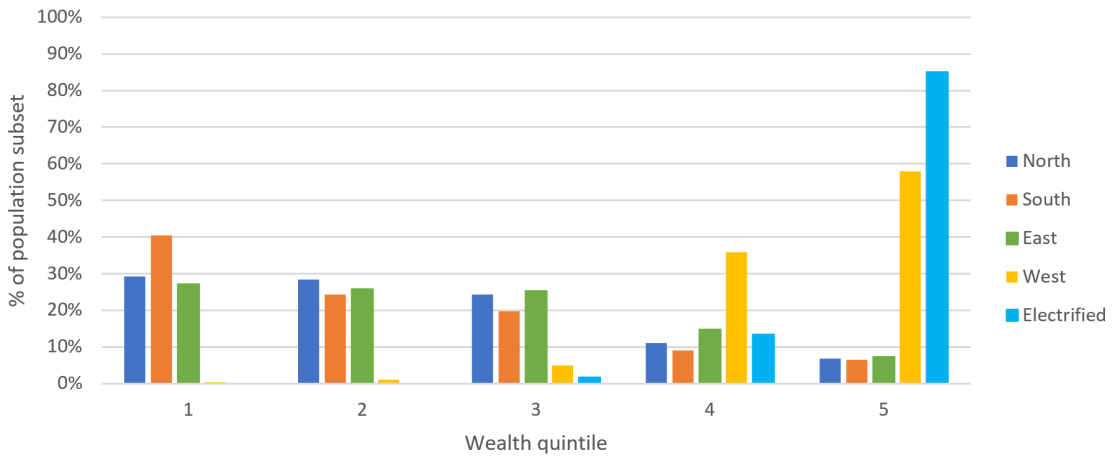


Figure 5.15: Proportion of those in each region of Sierra Leone and those with electricity access in each wealth quintile as recorded in the 2017 Multiple Indicator Cluster Survey. No one with electricity access fell in the lowest two quintiles; most are in the richest wealth quintile. Wealth across regions is uneven, with the West being the richest.

To account for this, the methodology presented in this chapter can be extended to downscale demand expectations towards a lower-end estimate. This can be accomplished by assigning appliance ownership proportionally to the current distribution of households amongst wealth quintiles. To achieve this, wealth quintile data can be retrieved from MICS for the region under consideration, and the proportion of all households falling within each wealth quintile can be calculated. Then, appliance ownership in each quintile can be evaluated at the location under study.

Appliance ownership in a community can be estimated based on a weighted sum of appliance ownership in each quintile. The number n of households over which demand is to be estimated can be subdivided into groups $n_1 : n_5$ proportional to the prevalence of wealth quintiles one (poorest) through five (richest). Appliances can then be allocated to the households based on appliance ownership in their quintile and region.

Cognizant that lower wealth quintiles in poorly-electrified LMICs are unlikely to have appliance ownership data available, as few such households are electrified, the following heuristics are recommended to estimate the lower-wealth appliance ownership patterns for quintiles one, two, and three where data may be unavailable:

- Quintile 1: All households use task lighting, and 50% of the households additionally charge one mobile phone.
- Quintile 2: All households use task lighting and general indoor lighting. They may also own a mobile phone, radio, or both. It is modelled that 25% have neither, 25% have a mobile phone, 25% have a radio, and 25% have both.
- Quintile 3: All households use task lighting, general indoor lighting, and outdoor lighting. These households may also own a mobile phone, radio, television, and fan, or some combination of these. This creates a total of 16 possible appliance combinations. The relative proportions of these combinations are taken from the *fourth* wealth quintile. If the total of these proportions does not add up to 100% (i.e. because there are other combinations included in the fourth wealth quintile besides these 16), the difference is distributed evenly amongst all 16 possible combinations.

Appliance ownership can then be aggregated across households of all wealth quintiles, and community demand can be simulated using a bottom-up stochastic approach. This creates a lower-end demand estimate for generation and storage sizing.

A similar approach can be taken to downscale to a "mid-level" demand based on anticipated (or desired) wealth growth in a region following electrification. Demand estimates can be generated based on the projected or desired proportions of people in each wealth quintile following electrification. For instance, in areas where the population is currently skewed towards lower wealth quintiles, an eventual movement towards an even distribution amongst wealth quintiles (i.e. $n_1 : n_5$ are equal) can be used to project mid-level demands. While not tested here, these methodological alterations are explored in Chapter 6.

5.3.6 Limitations and Areas for Future Study

This demand estimation approach assumes 100% reliability, which is not realistic in Sierra Leone at present. The grid in Sierra Leone experiences frequent and lengthy disruptions, with a system average interruption duration index (SAIDI) of 62.1

minutes and system average interruption frequency index (SAIFI) of 24.8 disruptions in 2020[4] [290]. These outages would impact expected demand first by depressing overall average load profiles; when electricity can be used on fewer days of a month, the monthly average will decrease despite the same usage on the days it is available.

Lower reliability can also engender reluctance to purchase expensive high-powered appliances. Though these appliances might be accessible to the relatively-wealthier people who currently have electricity and may be able to afford the risk of these appliances being damaged by a service disruption, this risk could be too great for lower-income households to bear. Even if they can afford initial appliance purchase, perhaps they would not be able to afford the replacement cost in the case of damage, causing reluctance to connect. This is the type of distrust that could spread socially; for instance, once a neighbour has had their refrigerator fried by a service disruption, it could be difficult to trust the grid enough to buy one yourself. The effect of poor reliability on appliance purchase needs further psychological and anthropological study. In the absence of reliable grid energy, perhaps robust appliances which can accommodate frequent disruptions are a solution.

Additionally, this approach does not account for the possibility of different versions of the same appliance. Each instance of each appliance was modelled identically for each usertype; no variance was included to account for different versions of the same appliance (e.g. with different power ratings). While such variation is likely in practice, this is beyond the scope of this work and left for future study.

Finally, no sensitivity analysis has been undertaken in this work. This approach is sensitive to the input data used; the level of uncertainty in this data will have affect the accuracy of the load profiles produced. It is assumed that the MICS dataset on appliance ownership is highly accurate, given the rigorous standards of this international UNICEF-driven effort. A greater source of uncertainty is the expert-sourced appliance usage windows. An interesting piece of future work would be to collect empirical data to confirm these windows, or to interview multiple

[4]For disruptions greater than 5 minutes.

experts to see if their anticipated usage windows agree. A sensitivity analysis could then be undertaken for the range of usage windows sourced through these methods. Stochastic parameters such as the δ values of 20% and 35% represent a final source of uncertainty. These are selected based on the original RAMP model literature [174] which undertakes a sensitivity analysis and finds them to be robust in the case study presented therein. However, as the application of this approach in Sierra Leone takes different data inputs and makes different assumptions than the case study presented in the literature, the sensitivity to these values could be reevaluated in this context. This is left to future study.

5.3.7 Scalability

This method's scalability stems from its reuse of data produced in existing, well-funded programmes. Collecting MICS or similar data is undoubtedly expensive, and requires a huge time investment, but this cost is already allocated and absorbed by the organisation undertaking the survey (i.e. UNICEF). The data is freely available online, and is being continuously collected in incrementing rounds in LMICs all over the world.

Once data is acquired, the analysis and demand estimation approach described herein can be completed swiftly. For instance, producing a synthetic community profile ($n = 200$) over a month (31 days) was timed at 613 seconds on a laptop with a Intel(R) Core(TM) i7-7500U CPU and 8 GB RAM with no graphics in this work. That is, it takes about 10 minutes to estimate demand over a month at village level on quite an average laptop computer. For an electrical engineer working on a grid design for a rural area, this should not represent a huge time impediment.

5.4 Key Outcomes

This chapter has proposed a scalable method which uses existing data to estimate spatially specific household demands in off-grid regions of LMICs. MICS data is proposed as an input data source for stochastic bottom-up demand estimation. Appliance ownership methodologies are proposed to analyze the most popular

appliances in a given spatial context and implications in terms of access level. Through a case study application of this approach in Sierra Leone, which leverages the 2017 MICS dataset and the RAMP stochastic demand estimation framework, the underlying premise of spatial variance of demand amplitude is shown to be true in this context. Urban and Western demand estimates are of greater amplitude than other regions and rural contexts. Furthermore, regional differences cannot simply be attributed to urban and rural dynamics. The load profiles generated using this approach are compared to a tier-based demand estimation approach that does not account for intra-community diversity (i.e. by designing "for" a certain tier such that all households have the same level of access). They are found to approximate a Tier 3 load, despite a total lack of Tier 3 appliances. This leads to a critique of the definition of the tiers themselves which, defined at the household level with a given set of appliances, may not be suitable in all contexts or when applied across communities. By leveraging existing appliance ownership data to generate spatially specific stochastic bottom-up demand estimates at scale, data collection cost and time are eliminated, making accelerated spatially specific demand estimation possible.

Having proposed feasible solutions to the location and demand data problems, these data can finally be applied in the spatial feasibility design of electricity access systems.

6

Spatial Design of Appropriate Electrification

Contents

With scalable methods to generate connection point location data and spatially specific demand estimates in hand, these data can finally be applied in the design of electricity access systems. This chapter therefore addresses the question: *Can we design appropriate least-cost electrical systems to match community spatial context?* It is argued that, using a scalable spatial framework, feasibility-stage electrical design can be achieved with home-level specificity at low cost and computational expense.

Rural electrification design is not a new problem; as such, models for this purpose are well-covered in the literature. These generally take either local or large-scale approaches, each of which provides useful information at different stages of the design process. Local electrical system design models such as HOMER [212], DER-CAM [214], and HOGA [215] typically focus on optimal sizing and operation of generation technologies and other large infrastructure components (e.g. storage, inverters) to meet local energy needs. Topology and associated conductor costs

are rarely considered in these models. They typically assume that the number of connection points within a system and/or the estimated load profile are known to the designer and can be used as inputs. These models can be quite useful in final design, once these parameters have been ascertained. However, as shown in Section 2.1 of Chapter 2, the locations of potential connection points are not always available in off-grid regions, and so it can be difficult to know how many connection points to consider. Furthermore, it is not always clear where to draw the boundaries between different autonomous local electricity systems, or how to identify which households lie outside their economically-optimal geographic range. As such, while these local models can be very useful when accurate connection point data are available and have been pre-processed to identify candidate areas for grid design, they do not assist in the scoping stage.

Large scale electrification design models such as OnSSET [233] or Network Planner [229] specify least-cost grid types (e.g. mini-grid, grid extension, SHS) and generation types (e.g. solar, wind, diesel) over geographic grid cells based on raster data. Topology is also neglected in these gridded models, as illustrated by the outputs of OnSSET shown in Figure 6.1. Assigning technologies by grid cell area can be useful in preliminary design when scoping the coverage of different electrification technologies. However, without consideration of the distribution of homes within grid cells and the possibilities of interconnection between cells, the outputs of these models miss a significant chunk of the implementation picture.

This leaves a wide gulf over which the designer must leap when progressing from preliminary design in large-scale models to final design in local models. Few models attempt to bridge that gap. The REM [238] is relatively unique in attempting to model at the interface of local and large-scale design by specifying grid types over large areas while simultaneously proposing topologies at home-level. However, REM requires extensive data inputs, is computationally expensive, and is not open source, making it less accessible to practitioners and policymakers designing rural electrification strategies. By endeavouring to create a fully-integrated pipeline from

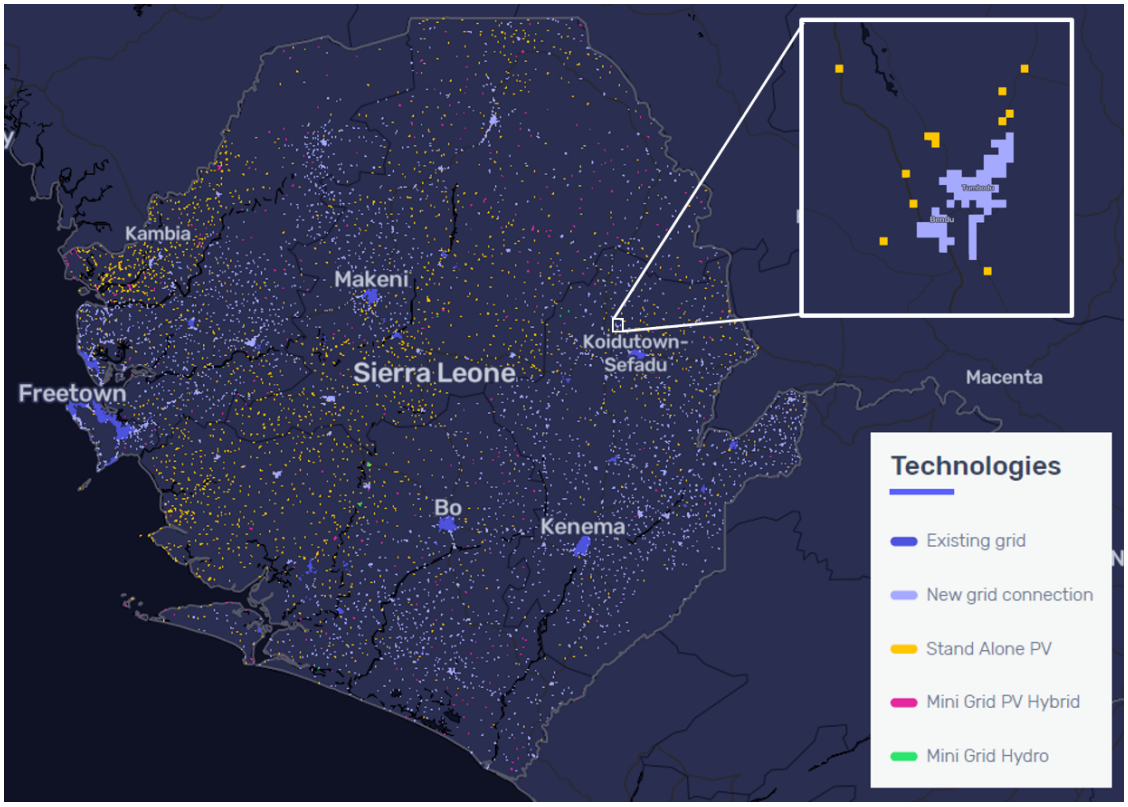


Figure 6.1: Outputs of the OnSSET model from the Global Electrification Platform in Sierra Leone. The inset image is an expansion of results at the indicated area. Note that the technologies are assigned over grid cells without accounting for topology.

raw data to final design, it becomes opaque to users and therefore difficult to operate and trust.

Adding to this difficulty, across both local and large-scale modelling approaches, few methods account for demand evolution in rural off-grid areas [175], as highlighted in **Gap 5** identified in Chapter 2. This is an issue, as access to electricity has been shown to stimulate loads which increase over time, particularly via social pressure to acquire higher-consumption appliances [46]. Planning a growth trajectory for increasing demand over time can be challenging when designing local grids, which will either require oversizing, retrofitting, modular expansion, or interconnection to handle increased demands. However, there is an ethical argument to be made that since grid-connected customers have access to higher power ratings allowing demand growth, the same opportunity for growth should be afforded to local grid and SHS customers. Planning to allow for demand increase allows for this self-determination

while simultaneously helping to minimize stranded assets.

Surveying the current literature, **Gap 6** identified in Chapter 2 becomes evident: practical and accessible spatial design approaches requiring minimal and realistic data inputs which can produce designs with home-level specificity are scarce. Any approach developed to fill this gap should navigate the transition between large-scale preliminary design models and local-scale final design models by: (1) identifying off-grid communities requiring unified grid technologies, (2) selecting best-fit grid types for each, (3) suggesting and costing preliminary topology designs, (4) specifying and sizing generation and storage, and (5) allowing the selection of an appropriate design point while accounting for potential future demand growth.

This chapter argues that a practical feasibility-level spatial electrical design framework can be implemented at low computational expense while still accounting for home-level data. Clustering, graph theory, GIS analysis, cost optimisation, and visualisation are leveraged in this framework to locate candidate communities for electrification, propose grid types and topologies, and size generation and storage. To account for increasing demand over time, a method is developed to map the technically feasible and affordable space of generation and storage design points and to chart possible demand accommodation trajectories. A proof-of-concept for the framework is demonstrated using a case study in Sierra Leone. This framework requires only openly-available input data alongside the connection-point-location and demand data which can be generated as shown in Chapters 3 through 5. The outputs provide a policy-ready feasibility-level electrical design proposal that can be actioned through on-the-ground verification and implementation.

This chapter proceeds as follows. First, the spatial design framework is developed in Section 6.1. This includes: a clustering approach to identify communities best-suited to an integrated electrical topology, and an accompanying grid type decision algorithm; graph theory methods to estimate the most effective electrical topology for each community; GIS analysis to select best-fit generation and storage types for each local grid cluster; and a method to map possible design points in terms of demand and affordability to chart pathways towards modular demand growth. This

framework is demonstrated for a case study area in Sierra Leone and the results are presented and discussed in Section 6.2. Finally, Section 6.3 summarizes the key outcomes and implications of this work in terms of the broader questions of the thesis.

6.1 Methodology

The general methodology for the spatial electrification design process is shown in Figure 6.2 and proceeds as described in the following steps:

1. Potential connection points (as located in Chapters 3 and 4) are clustered into energy communities and grid types are selected for each cluster.

2. Distribution and interconnection grid topologies are proposed for each cluster using graph theory.

3. Site-specific resource potentials and availability are analysed in GIS to select generation types for each cluster.

4. These results are used in combination with regionally tailored demand estimates (as described in Chapter 5) as inputs to a cost optimisation for generation and storage sizing.

5. Potential design points are mapped in terms of technical feasibility and affordability to chart pathways towards higher demand through modular technology expansion or capacity reduction. The outputs of all stages are collated into an actionable spatial feasibility design.

6.1.1 Energy Community Clustering

Household location data produced in Chapters 3 and 4 are first spatially clustered into "electricity communities" which may benefit from a unified electrical system. While clustering algorithms can be based on definitions of data similarity including distribution, density, connectivity, and centroid [291], a density-based approach which allows outliers is best suited to capture natural geographic clusters of homes.

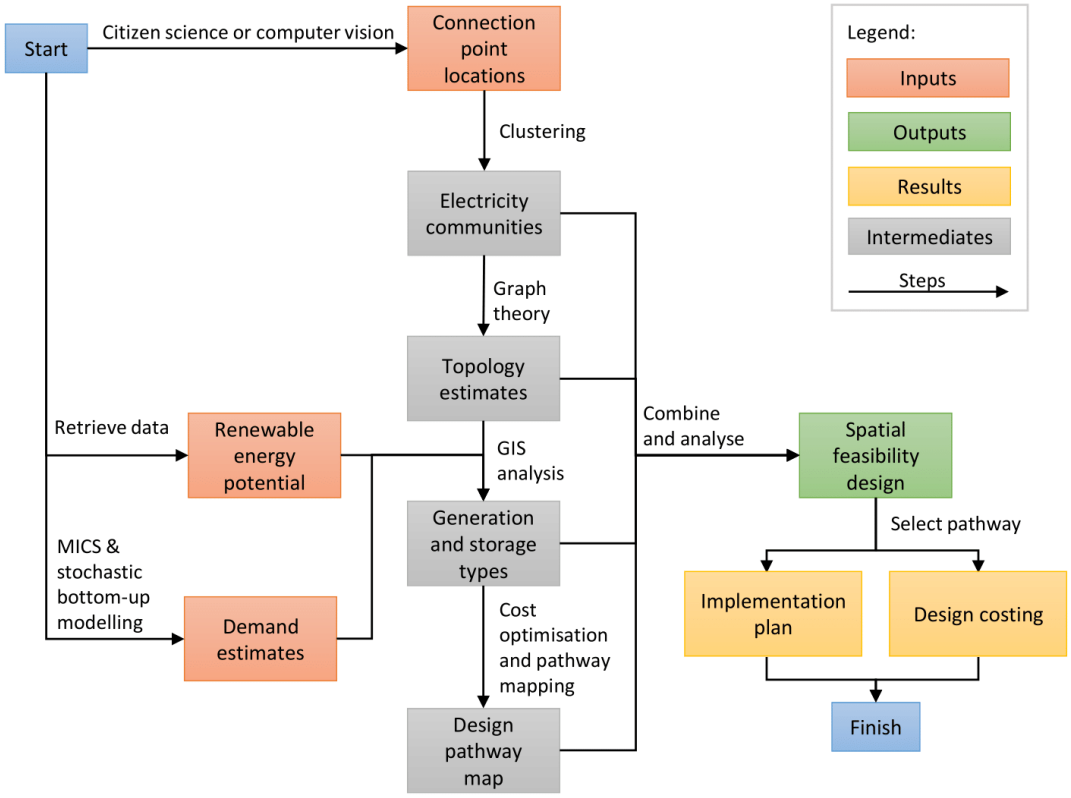


Figure 6.2: General methodology for the spatial electrification design framework.

The Density-Based Clustering of Applications with Noise (DBSCAN) algorithm is well-suited to this task. This algorithm groups data based on density and does not force all points to join clusters. This is useful, as outliers can be identified as distant households which are more likely to benefit from stand-alone systems. Unlike other common clustering methods such as K-means, DBSCAN does not require the expected number of clusters as an input. This is useful here, as the expected number of electricity communities in a large set of homes (i.e. over a region or country) is difficult to specify in advance without counting communities manually. DBSCAN requires the specification of two hyperparameters, but these can be set without tedious manual analysis. For instance, Ordering Points to Identify the Clustering Structure (OPTICS) [292] can be used to investigate the impact of different hyperparameter selections. Here, a tuning heuristic is proposed to configure the DBSCAN hyperparameters to the spatial context under consideration.

In DBSCAN, each point is classified as either a core point, border point, or noise based on the density of points near to it. Core points have at least a defined minimum number of points ($minPts$) within a radius ε, while border points have a number of points $< minPts$ within radius ε, and noise points (i.e. outliers) have no other points within that radius. To build clusters, core points are first identified throughout the data. A random core point is selected as a starting point; each core point within radius ε of that point will then check how many other core points exist within radius ε from themselves to see if they can "grow" the cluster. Each core point added to the cluster repeats this process until no more core points can be found. Then, non-core points within the radius ε of any of these points are added to the cluster as border points. Another non-clustered random core point is then selected and the process repeats until all core points have been clustered [293]. This process is visualised in Figure 6.3.

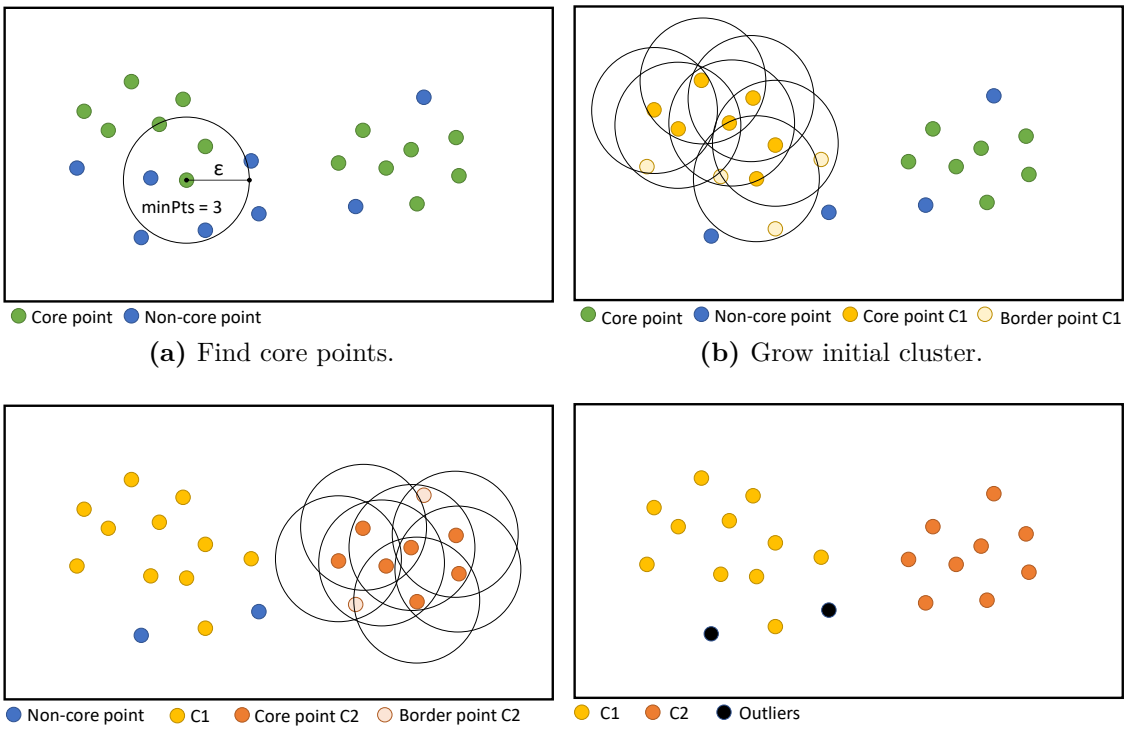
To configure DBSCAN to spatially cluster home locations into electricity communities suited to unified distribution infrastructures, ε and $minPts$ are set as described in the following subsections. Clusters are then sorted into best-fit spatial grid types using the thresholding algorithm described thereafter.

Setting *minPts*

$minPts$ is set to the minimum number of connections accommodated by the smallest communal electrical infrastructure option under consideration. Clusters below this threshold, as well as isolated homes, are marked as outliers to be evaluated for SHS deployment and/or eventual interconnection to a nearby cluster. For instance, if the smallest communal option under consideration when applying this approach is a 15-connection micro-grid, $minPts$ would be set to 15. However, this parameter should be customized based on local policy or planning needs depending on the palette of infrastructures under consideration.

Setting ε

The ε hyperparameter can either be tuned based on the density of households in the region under consideration or set based on under-grid distances defined in



(a) Find core points. **(b)** Grow initial cluster.

(c) Repeat on remaining core points. **(d)** Identify outliers.

Figure 6.3: Visualisation of the DBSCAN algorithm. (a) Core points are identified, and a random core point is selected to begin cluster growth. (b) Each core point within radius ε of that core point looks for additional core points until no more core points are found. Then, border points are added. (c) Steps a and b are repeated until all core points are clustered. (d) Any remaining unclustered points are identified as outliers.

electrification policy. To motivate the usefulness of an ε tuning heuristic, consider the difference between urban and rural settlement styles. Both have clusters of homes, but the typical distance between homes is likely to differ; urban households are likely to be closer together on average. If a distance threshold for cluster formation based on the urban context is used in the rural context, it is likely that the resulting clusters would not follow visible groupings, instead splitting communities into multiple sub-segments or classifying all homes as outliers. This is undesirable and unhelpful in the context of electrical design.

To avoid this, ε is tuned based on the density of homes in the region surrounding the cluster. This region can be defined based on a political boundary (e.g. county, district) or a specified radius. Within this region, the K nearest neighbours to each building are found, and ε is taken as the knee point of these. This tuning

approach is originally suggested in [293], and is applied here in the context of home clustering. In nearest neighbour calculations, K is set to $minPts$, or the size of the smallest communal infrastructure option under consideration. The distance to the Kth nearest neighbour for each building is plotted in ascending order, creating a $k-dist$ plot which tends to make a sharp curve with a reflected L-shape. The knee of this curve is found, and the distance value at that knee point is used as ε.

While building location data do tend to make a single-kneed L-shaped $k-dist$ plot, it could be possible to have multiple-kneed plot in a region with multiple distinct community densities. In this case, a multi-stage application of DBSCAN will be required, using ε values taken at each knee point. An approach that allows for this, such as using OPTICS [292] or Dynamic Method DBSCAN [294], could be applied instead of vanilla DBSCAN to handle this.

Alternatively, ε can be set as the "under-grid" distance (d_{under}) defined in regional or national electricity policy. Many countries undergoing electrification define a maximum distance from existing supply points after which grid electrification shall not be provided and before which a home is considered under-grid. For instance, the Sierra Leone National Electricity Act dictates that electricity shall not be supplied to premises more than 200 m from the nearest supply point [295]. To incorporate these policy constraints into the clustering approach, ε can instead be set to d_{under} instead of being tuned to suit the regional density. Note, however, that setting ε to d_{under} is not a direct translation of the policy's intention. Whereas policy typically refers to a maximum distance *to a supply point* (i.e. transformer, substation), setting ε to d_{under} instead limits the *radius between homes* to form a cluster. Since transformers are not yet placed in off-grid areas, resulting clusters may include homes more than d_{under} from eventual transformer locations. To account for this, a percentage of d_{under} may be used as ε instead, but this percentage must be selected somewhat arbitrarily. Alternatively, after clustering is complete, households outside a d_{under} radius from the cluster centroid could be retroactively assigned as outliers, assuming that the transformer will be placed at the centroid. Here, ε is simply set to d_{under} to explore the impact of this approach without these adjustments.

Grid Type Selection Algorithm

Each cluster is assigned a likely best-fit grid type based on size and distance-from-grid thresholds. As defined in Algorithm 2, each outlier is treated as an individual cluster of size one, and assigned SHS as their best fit technology. Clusters above the minimum communal size $minPts$ are assigned either grid extension or local grid technologies. Grid-extension is assigned if the cluster is less than d_{under} from an existing national grid transformer or substation. Local grid electrification is assigned if they are further than d_{under} from transformers or substations.

Algorithm 2 Best-fit grid type decision process for clusters.

```
d_under ← under-grid distance
for c in clusters:
    n_c ← quantity of homes in c
    if n_c == 1:
        return SHS
    else if n_c > 1:
        d = distance from c to nearest transformer or substation
        if d ≤ d_under :
            return Grid extension
        else if d > d_under :
            return Local grid
```

6.1.2 Topology Estimation

Grid topologies are next estimated for electricity communities. A two-stage graph theory process is applied, wherein lower-voltage distribution grids and higher-voltage interconnection or transmission grids are constructed separately then considered together.

Distribution

The distribution grid for each electricity community requiring grid extension or an autonomous local grid is treated as a graph, where homes are nodes and the paths between them (i.e. conductors) are edges. An undirected fully-connected graph is first calculated as a Delaunay triangulation of all building locations within an electricity community. The edges of this graph are weighted by the distance between

the homes at either end-point. Then, the minimum spanning tree (MST) of the graph is then calculated using Kruskal's algorithm [296]. The MST is treated as the distribution network estimate which uses the minimal possible conductor distance to implement, since it connects all homes in the community at a theoretically minimal distance cost. A purely radial distribution network is also generated for each electricity community, and treated as the maximum bound for conductor distance required. This is generated by connecting each home in an electricity community to its cluster centroid.

Interconnection

To estimate how electricity communities may become interconnected, an undirected fully-connected graph of *community centroids* is next calculated as a Delaunay triangulation of community centroids. An MST of the centroids is then generated using Kruskal's algorithm on this graph, where community centroids are nodes and edges are interconnection lines. The resulting network interconnects the clusters with the theoretical lowest possible cabling distance. A high-end radial estimate is not considered for interconnection, as this is highly unrealistic.

6.1.3 Context-Appropriate Generation and Storage

Generation and storage types are selected to meet demands in communities requiring local grids. To facilitate this, demands are estimated using the methods described in Chapter 5, including the expansion to estimate low, middle, and high-end demands as discussed in Section 5.3.5. GIS-based analysis of generation potentials and weather is then undertaken at cluster locations to select generation and storage types which can meet demands. Solar and wind generation resource potentials are retrieved from GlobalSolarAtlas, GlobalWindAtlas, and Renewables.Ninja [54–56] and analysed for their relative strength and suitability. Potential for hydro development is evaluated based on the International Renewable Energy Agency (IRENA) African Renewable Electricity Profiles for Energy Modelling (AFREP) database [297]. If no opportunities for hydro development are highlighted within a 3 km radius of

the community under consideration in this database, then hydro is disregarded as a possible generation type. Weather patterns at each community location are also retrieved from Renewables.Ninja and visualised to better understand potential impacts of diurnal or intra-annual variance in renewable resource potentials on generation feasibility. While other thermal generation types (e.g. diesel) could be included in this analysis, as stated in Chapter 1, this thesis focuses on renewable electrification, so this option is disregarded.

6.1.4 Demand and Design Pathways

As discussed in Chapter 5, there is a trade-off between generation and storage capacity, expected grid reliability, affordability of electricity to the consumer, and the frequency of grid expansion or retrofitting. A method is therefore developed to map demand evolution trajectories through the "design point space" of generation and storage sizes. This is used to chart paths for demand growth via modular expansion or reliability reduction.

As an industry standard in micro-grid generation and storage optimisation, HOMER is first used to optimize the sizing of the selected generation and storage types for each cluster for lowest net present cost. This is done for low-end, mid-level, and high-end demands calculated using the methods described in Chapter 5. Costs per kW of generation components are retrieved from IRENA [298], and costs per kWh for battery storage are taken from United States National Renewable Energy Laboratory (NREL) benchmarks [299]. These benchmarks define different costs for residential (i.e. 4 kW to 7 kW), commercial (i.e. 100 kW to 2 MW), or utility-scale (i.e. 5 MW to 100 MW) systems; the appropriate level is selected based on the demand estimates for the cluster under consideration. The remainder of parameter configurations in HOMER are routine for local grid optimisation; details are included in Appendix D.

After cost-optimal generation and storage sizes are found, the levelized cost of energy[1] (LCOE) is calculated over a search space of generation and storage extending

[1]Note that LCOE calculations in HOMER do not account for distribution grid installation and conductor costs. As such, they are an underestimate.

outwards. For instance, in the case where a 10 kW of generation and 10 kWh of storage are optimal, a search space from 1-100 kW of solar and 0-99 kWh of batteries is simulated. LCOE at each design point is visualised for each demand estimate. Overlaps in these design point spaces are identified to understand the affordability impact of oversizing a system to accommodate future demands. This enables the selection of an initial design point, either to meet current needs (thus requiring future modular expansion) or to meet future needs (thus reducing present affordability).

The feasible space for each demand level is limited by affordability constraints which represent energy poverty prevention. An energy expenditure budget is defined as a maximum of 10% of household expenditure in the region under study[2] to prevent energy poverty. This is translated into a maximum affordable cost per kWh based on the demand estimate considered and compared to the LCOE results across the design space. Here, LCOE figures greater than this cost per kWh are eliminated from consideration as economically unfeasible. If desired, they could instead be kept under consideration with a caveat that they will necessarily require subsidy to remain feasible.

It may be impossible to design a system which is both affordable in the present and fully reliable in the future. The effect of reducing capacity in high-end demand cases is therefore studied to observe whether the feasible design space for high-end demand can be expanded inwards towards lower capacity generation and storage at lower cost by introducing reliability risk. A capacity shortage percentage is introduced in HOMER simulations to investigate this, defined as the total capacity shortage (in kWh/year) divided by the total annual electric load (in kWh/year). For instance, a capacity shortage of 1% allows for systems which can provide 99% of the energy requirements (kWh) of the input load profile to be considered. Typically, this results in peak shaving on outlier days with low generation potential. For instance, given the stochasticity of the load profiles used, there could be a particular load peak which occurs on a remarkably cloudy day. To provide full reliability (i.e. 0% capacity shortage) on this outlier day would mean installing far more battery

[2]This can be retrieved from an LSMS, IHS, or another context-specific survey.

capacity than is typically needed at high expense. However, by introducing a small acceptable capacity shortage, which would result in a small amount of lost load during this outlier peak, perhaps a much more affordable system design can be implemented while retaining most of the benefit to the community.

Results are analysed to estimate the feasibility of modular generation and storage expansion which preserves affordability as demand grows. Modular expansion pathways are planned by assuming that demands will start closer to the low-end estimate and proceed to mid- and high-level estimates over time. The cost of expansion is calculated based on the cost of the net addition of generation and storage elements required to reach the inner boundary (i.e optimal design point) of the next demand level. As this will occur in the future, this is based on future projections of technology costs, which can be drawn from the NREL annual technology baseline [300] or cost projections [301] among other possible sources.

Finally, these results are augmented with conductor costs and land-use feasibility. Given the expected demands to be met by the system, a distribution voltage can be estimated, and costs for aluminium conductor steel reinforced (ACSR) or other conductor type can be calculated based on distribution grid distances previously generated. These cost can then be used to adjust LCOE to a more expected level based on the full bill of materials cost. As renewable generation installations can require significant land use and particular land types, the land requirements to install the generation required at the initial design point and any subsequent modular expansions are also evaluated and cross-checked with land use data from OSM or Copernicus. The likelihood of adequate land availability near each cluster to implement required generation capacity is evaluated to ensure practicality of the demand trajectory.

6.2 Case Study Results and Discussion

These methods are demonstrated in a case study region in the Northern province of Sierra Leone containing a number of small communities over an area of 32.8 km^2. Homes are located in this area in Chapter 3; there are a total of 335 home locations

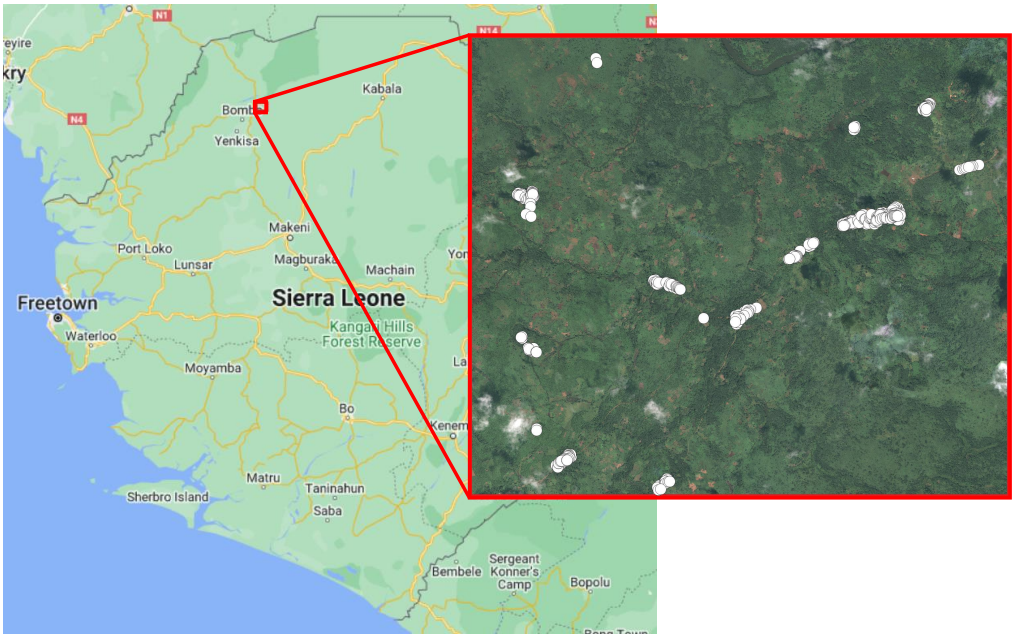


Figure 6.4: Case study area containing home locations (shown as white dots) identified through citizen science in Chapter 3.

identified. The case study area is visualised in Figure 6.4, overlaid with the home-level location data collected in Chapter 3.

Stage 1: Clustering

Home locations are first clustered with DBSCAN as outlined in Section 6.1.1. It is selected to use $minPts = 5$ as a minimum estimate for the amount of homes that may be feasible with the smallest possible communal electrical infrastructure. Using the ε tuning heuristic, the knee of the $k - dist$ plot is found to be 76.8 m, as illustrated in Figure 6.5. As such, home locations are clustered with $\varepsilon = 76.8$ to produce the results shown in Figure 6.6a. This results in 12 natural and distinct visible clusters, and 11 homes classified as outliers.

The homes are also clustered using DBSCAN with ε set to the policy-defined d_{under} of 200 m in Sierra Leone. The clustering results, shown in Figure 6.6b, do not change very much, as the homes in this case are generally grouped in closely-packed communities with large gaps (e.g. > 200 m) in between. However, two small changes can be noted. First, two clusters identified individually in the $\varepsilon = 76.8$ case merge into one cluster when $\varepsilon = 200$. This can be seen as the pink and green clusters

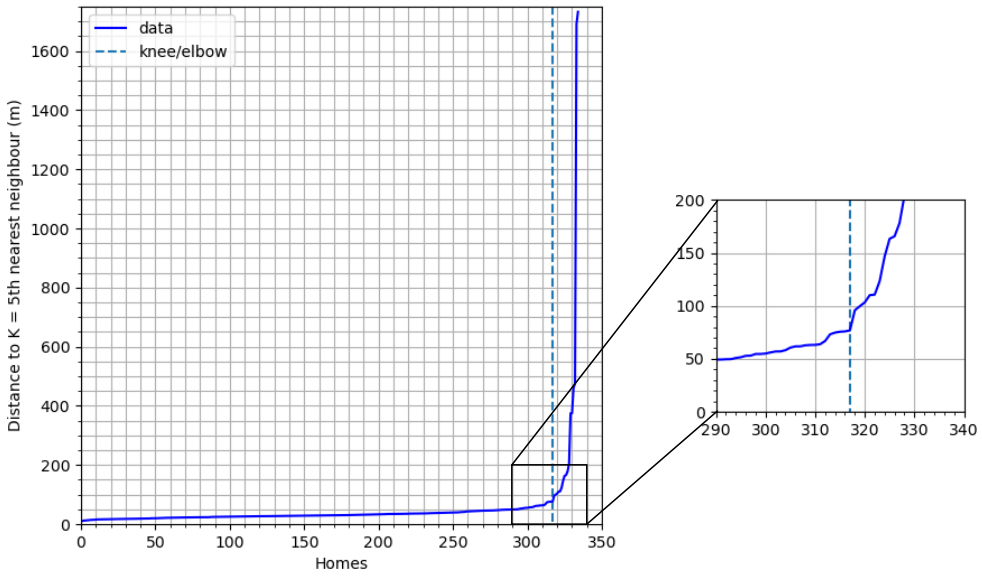
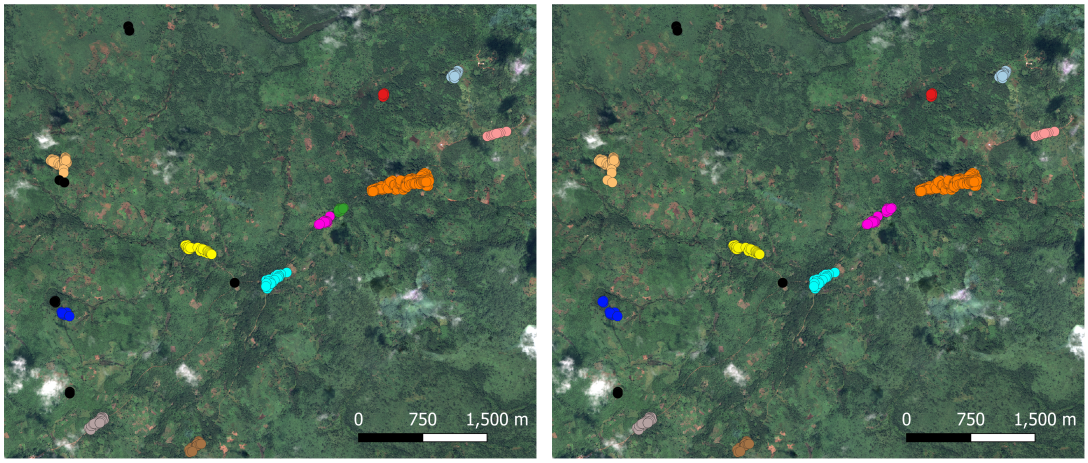


Figure 6.5: K-dist plot for the case study (k = 5 = minPts) with the knee point identified. The knee of 76.8 m is used as ε in DBSCAN.

in Figure 6.6a becoming a single pink cluster in Figure 6.6b. Additionally, some outliers identified in the $\varepsilon = 76.8$ case merge into clusters in the $\varepsilon = 200$ case – see, for instance, the royal blue cluster in Figures 6.6a and 6.6b. Clustering with $\varepsilon = 200$ creates 11 clusters as opposed to the 12 found using the $\varepsilon = 76.8$ and excludes six homes as outliers compared to 11 in the $\varepsilon = 76.8$ case.

While the results for the tuned and under-grid cluster cases are similar, one set of clusters must be used going forward in the design process; the tuned cluster set (i.e. $\varepsilon = 76.8$) is selected to be carried forward.

A spatial best-fit grid type is selected for each of these clusters based on their size and distance from existing grid infrastructure. As outlined in Algorithm 2, all outliers (i.e. cluster size one) are assigned stand-alone SHS systems. Clusters with more than one household are evaluated to see whether they are located within d_{under} of existing transformers and substations. Existing grid infrastructure is not within d_{under} for any of these clusters; so, each is specified as best-suited to a local grid.



(a) Tuned: $\varepsilon = 76.8\,\text{m}$, 12 clusters. **(b)** d_{under}: $\varepsilon = 200\,\text{m}$, 11 clusters.

Figure 6.6: Clusters identified in the case study area using DBSCAN with (a) $\varepsilon = 76.8\,\text{m}$ based on the tuning heuristic and (b) $\varepsilon = 200\,\text{m}$ based on policy-defined under-grid distance in Sierra Leone. Clusters are each a different color; outliers are shown in black.

Stage 2: Topology

Distribution grids are next estimated for these clusters. The MST approach outlined in Section 6.1.2 is first applied to generate estimates which bound the low-end of required conductor distance. A radial approach is also taken to gain a high-end conductor estimate. Resulting conductor lengths required in each cluster for MST and radial approaches are provided in Table 6.1. Finally, an interconnection grid is estimated for the local grid clusters. This is shown in Figure 6.7 alongside two inset examples of the MST-based distribution grids. The interconnection network would require 12 km of conductor to connect all clusters in this area.

To validate the use of the MST as a low-end conductor estimate for distribution grids, a mid-level conductor estimate is also generated by connecting all homes in each cluster via drop lines using conductor following roadways. An example is shown in Figure 6.8. The conductor length for these grids (d_{road}) is compared with the conductor length of the MST-based grid (d_{MST}) for all communities. The results are also shown in Table 6.1. Depending on the cluster considered, d_{road} is between 10% and 90% larger than d_{MST}. The sum of d_{road} across all clusters is 35% longer than the sum of d_{MST}. Amongst clusters, the gap between d_{MST} and

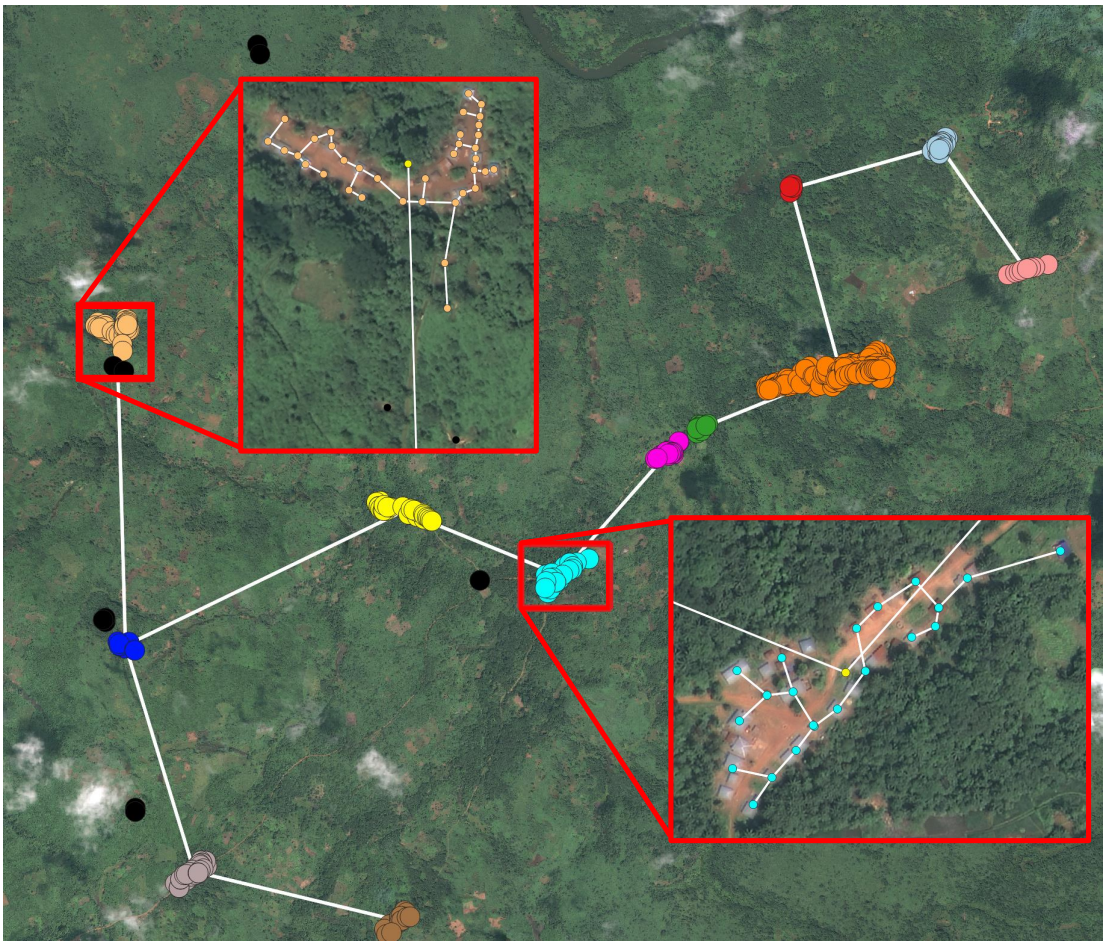


Figure 6.7: Interconnection lines in the case study area with two examples of minimum-spanning-tree-based distribution grids inset.

d_{road} is smaller in more linear communities with homes closer to roads, and larger in communities with homes deeper set from roads. The MST is thus shown to be a good low-end estimate. To convert it to a more realistic estimate based on these results, d_{MST} can be multiplied by a factor of 1.35 on average.

Stage 3: Generation and Storage selection

As the remainder of spatial design is undertaken for each cluster individually, results for one cluster are presented as an example. This is cluster four in Table 6.1, or the orange cluster shown in Figure 6.6 and 6.8, hereafter referred to as C4.

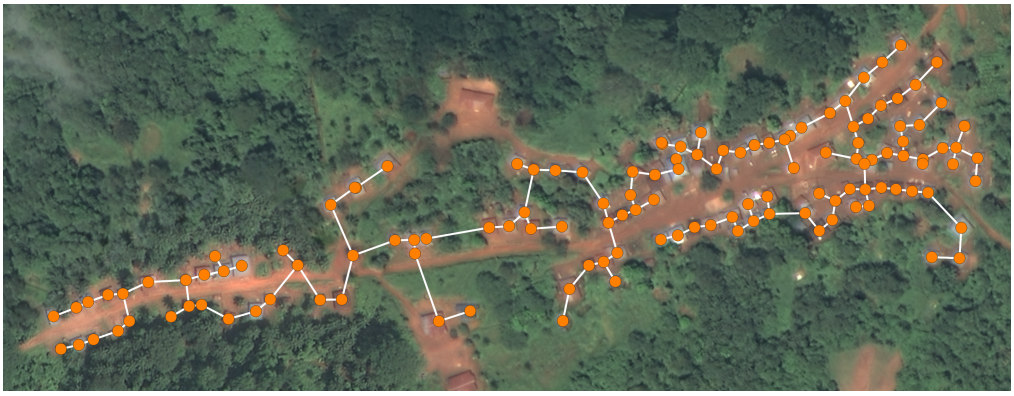
A high-end demand estimate is generated for C4 using the methods proposed in Chapter 5. The alteration to these methods proposed in Section 5.3.5 is used

Table 6.1: Distribution conductor lengths for each cluster using the minimum spanning tree approach (d_{MST}), the road-following (d_{road}) approach, and the radial (d_{rad}) approach. The number of homes in the cluster (n), the length of conductor required per home in the minimum spanning tree case (d_{MST}/n) and the ratio of the road-following length to the minimum spanning tree length (d_{road}/d_{MST}) are also listed.

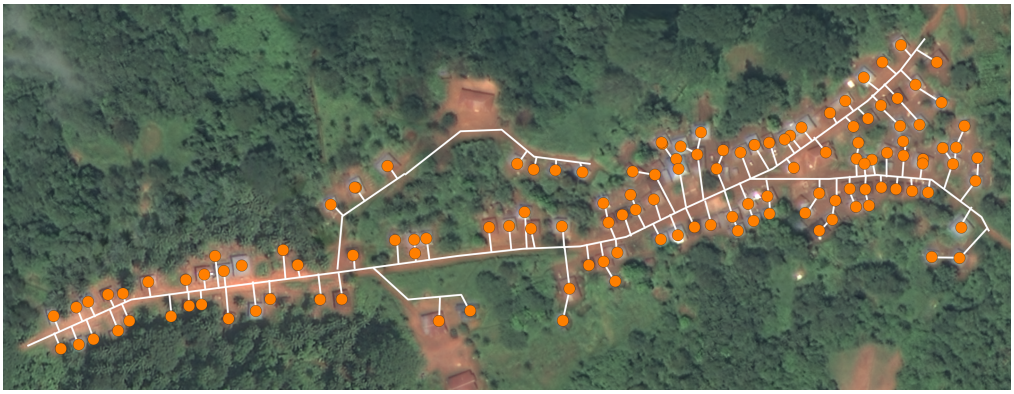
Cluster	**n**	$\mathbf{d_{MST}}$ **(m)**	$\mathbf{d_{road}}$ **(m)**	$\mathbf{d_{rad}}$ **(m)**	$\mathbf{d_{MST}/n}$ **(m)**	$\mathbf{d_{road}/d_{MST}}$
0	18	229	301	574	13	1.3
1	5	66	75	82	13	1.1
2	9	234	434	456	26	1.9
3	36	571	727	2,419	16	1.3
4	133	2,169	3,075	23,472	16	1.4
5	10	138	213	230	14	1.5
6	14	223	303	531	16	1.4
7	26	506	636	2,701	20	1.3
8	22	552	728	1,642	25	1.3
9	10	170	217	348	17	1.3
10	28	421	491	1,640	15	1.2
11	13	299	355	703	23	1.2

to generate low-end and mid-level estimates. The low-end estimate is calculated using the currently expected distribution of wealth quintiles in the Northern region of Sierra Leone, while the mid-level estimate assumes an equal distribution of the population amongst wealth quintiles. Demand is estimated over one year at minute-level resolution. The resulting average daily load profiles for C4 are shown in Figure 6.9. Average daily energy consumption in C4 is estimated at 53 kWh in the low-end case, 102 kWh in the mid-level case, and 301 kWh in the high-end case.

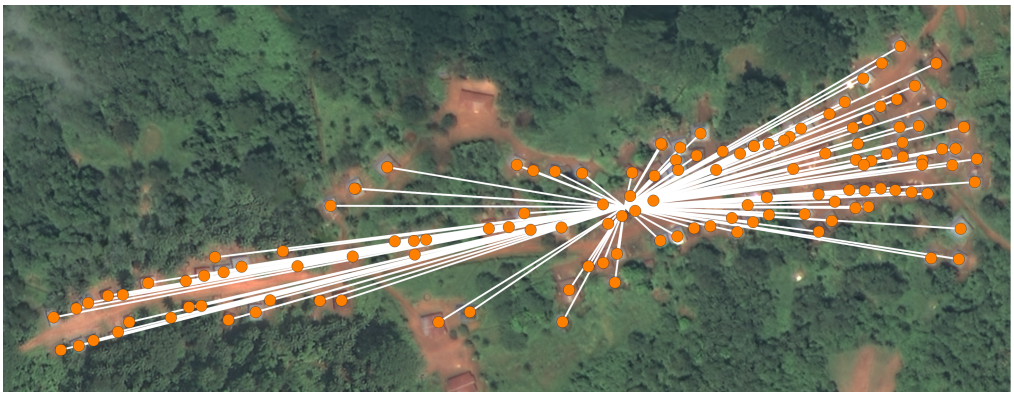
Generation and storage types are next selected as described in Section 6.1.3. Weather, climatic, and resource potential data are retrieved for Sierra Leone and identified at the C4 centroid (i.e. 9°32'47.823", -12°10'9.611") from Renewables.Ninja, GlobalSolarAtlas, and GlobalWindAtlas [54–56], and analyzed to determine the feasibility of different generation types. As there are no opportunities for hydroelectric development identified in the AFREP dataset in a 3 km radius of C4, hydro generation is not viable. For the 10% windiest regions within a 3 km radius of the C4 centroid, the mean wind speed is 3.98 m/s (i.e. 14.3 km/h); this is is quite low for wind turbine operation. The average specific yield of solar PV at this cluster



(a) Minimum spanning tree: 2,169 m



(b) Road: 3,075 m



(c) Radial: 23,472 m

Figure 6.8: Minimum spanning tree, road-based, and radial distribution grid topologies for an example electricity community (C4) in the case study region in Sierra Leone. While the minimum spanning tree result does not exactly resemble the road-based topology, it is a realistic low-end distance estimate.

location is found to be 4.2 kWh/kWp per day [55], as shown in Figure 6.10. Weather patterns in this community vary both intra-annually (based on the dry and rainy seasons) and diurnally. Importantly, irradiance (see Figure 6.11) and cloud cover

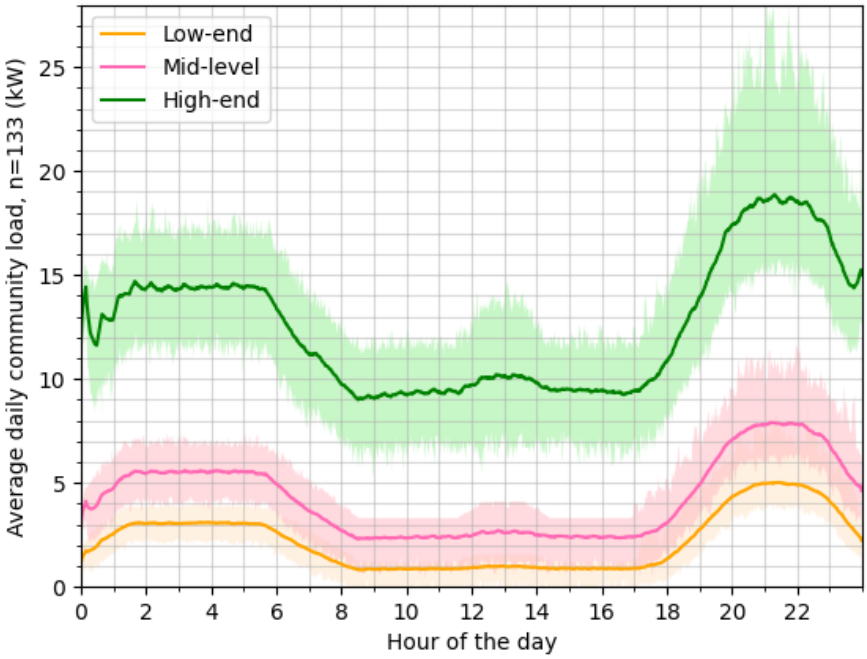


Figure 6.9: High-end, mid-level, and low-end average daily load profile estimates for C4. The full range of data for each case is shown in the lighter semi-transparent band.

(see Figure 6.12) vary throughout the year, resulting in less sunny weather in the rainy season which can limit PV generation. As described in Chapter 5, however, this season is likely to have lighter loads due to lower air circulation needs, which should help to compensate for the lower generation capacity. Grid sizing for C4 therefore proceeds based on PV as the selected generation type. Battery storage is also incorporated since peak demands tend to occur after nightfall, so movement of the solar energy in time will be necessary to satisfy this peak.

Stage 4: Cost Optimisation and Demand Trajectory Planning

PV generation and battery storage sizes for C4 are then cost-optimised in HOMER. The cost for solar PV modules is drawn from the IRENA costs of renewable power generation in 2020 [298]. Since specific costs for Sierra Leone are not included, costs from India are used as an LMIC with costs available. This gives a PV module cost of $224/kW. Battery cost estimates are taken from the NREL benchmarks [299] in the commercial system range. These vary based on battery duration – here, the four

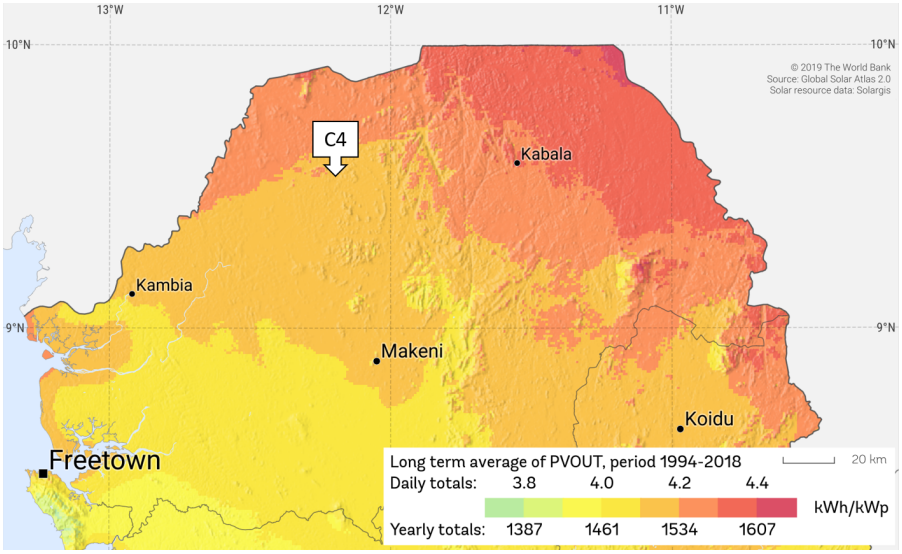
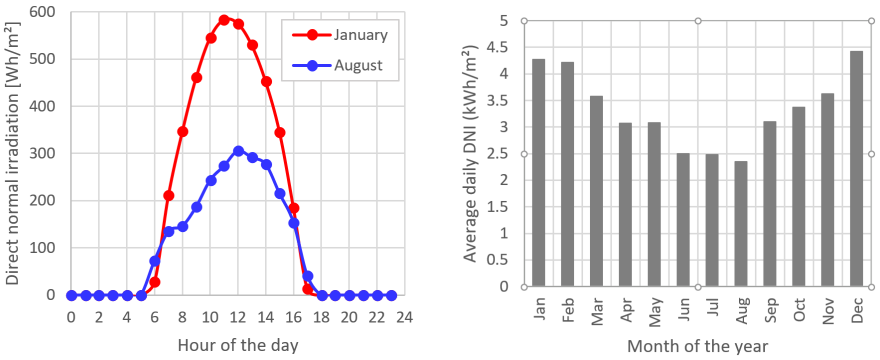
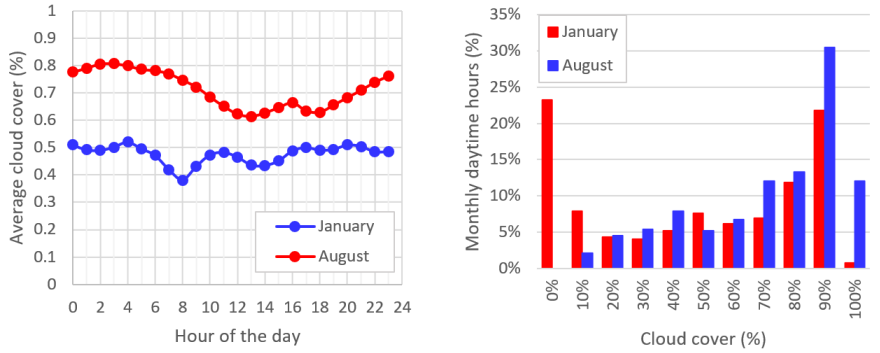


Figure 6.10: Specific yield of solar photovoltaic panels in Sierra Leone, with the approximate location of C4 marked [55].



(a) Diurnal variance **(b)** Intra-annual variance

Figure 6.11: Direct normal irradiance varies (a) diurnally and (b) intra-annually in C4.



(a) Diurnal variance **(b)** Intra-annual variance

Figure 6.12: Cloud cover varies (a) diurnally and (b) intra-annually in C4.

Table 6.2: Cost-optimized battery and storage sizes for each simulated demand level, no capacity shortage.

Demand	PV (kW)	Battery (kWh)	LCOE ($/kWh)
Low-end	24.6	63	0.15
Mid-level	64.4	113	0.15
High-end	191.9	302	0.13

hour duration battery costs of $194/kWh are used. Global horizontal irradiation (GHI) input data for HOMER are retrieved from Renewables.Ninja.

Generation and storage are optimally sized for least net present cost with full reliability in low-end, mid-level, and high-end demand cases. The resulting optimal PV and battery sizes, and LCOE at each size, are listed in Table 6.2. LCOE is then determined for a design point space extending outward from each optimal point.

These spaces are limited based on affordability constraints for energy poverty prevention. As calculated in Table 5.12, the average household in the Northern region of Sierra Leone has a maximum monthly energy budget of 33,400 Le before energy poverty risk. This translates to a different tariff cap depending on the demand level considered; a monthly average usage of 8.0 kWh per household is expected for low-end demands, while 15.5 kWh is expected for mid-level demands and 45.8 kWh for high-end demands. Based on these consumption levels, tariff caps for energy poverty prevention for the average family would need to be 4149 Le/kWh ($0.54/kWh) for the low-end case, 2149 Le/kWh ($0.28/kWh) for the mid-level case, and 729 Le/kWh ($0.09/kWh) for the high-end case. Comparing these tariffs with the LCOE results in Table 6.2, it becomes clear that serving C4 at high-end demand while maintaining a $0.09/kWh tariff cap is likely to be impossible. Building a system to serve this demand will either be financially unsustainable, require subsidy, or risk energy poverty amongst users. Based on these results, the low-end and mid-level design point spaces are limited based on their respective tariff caps (i.e. $0.54/kWh and $0.28/kWh), and it is noted that there is unlikely to be any affordable design point to meet high-end demands at present.

The high-end optimisation is then re-run allowing for capacity shortages from 0% to 25% in 0.1% increments to investigate of whether affordability can be improved by

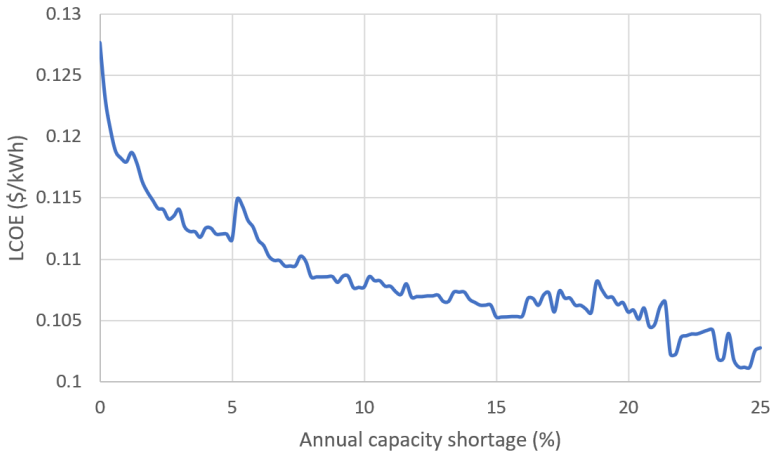
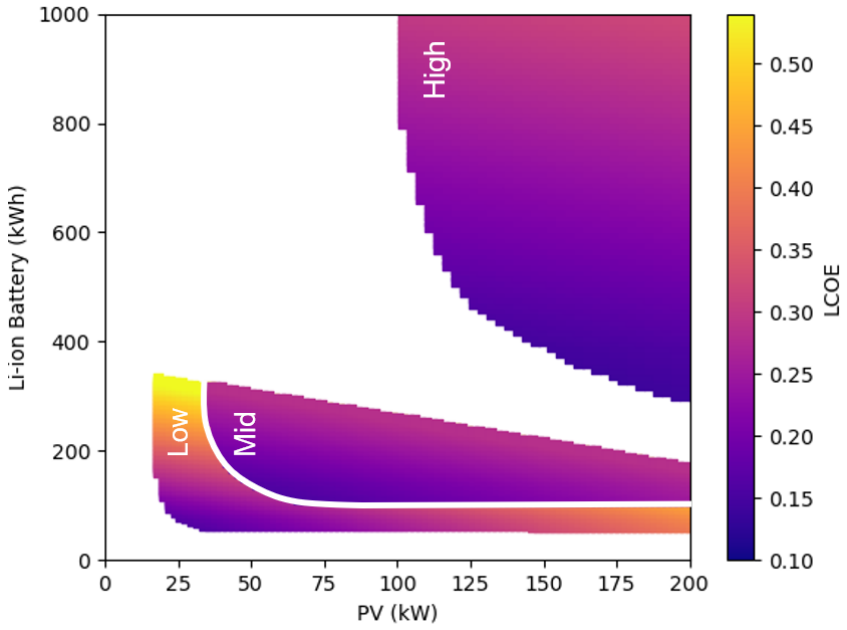


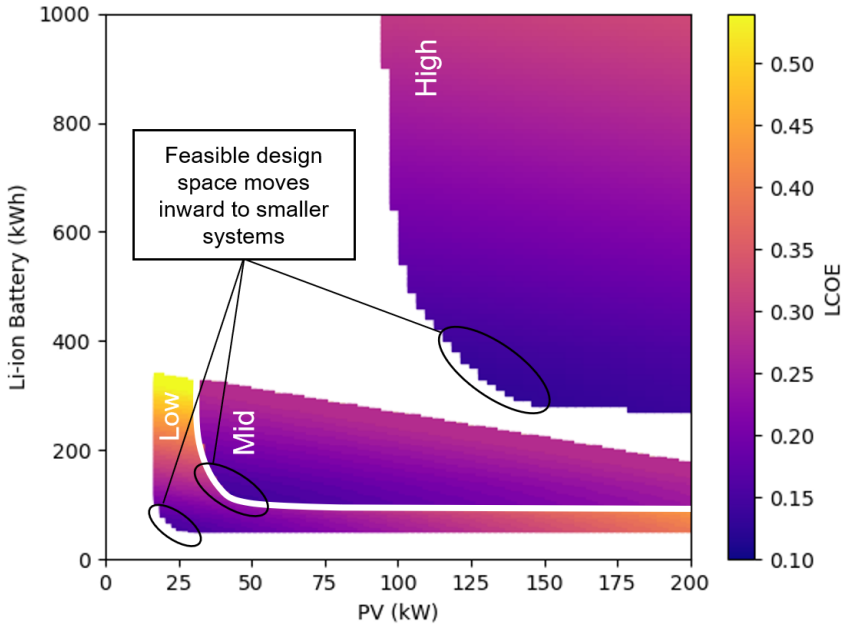
Figure 6.13: Levelized cost of energy (LCOE) achieved at lowest net present cost for capacity shortages between 0% and 25%.

sacrificing reliability. The results are illustrated in Figure 6.13. Even with capacity shortages up to 25%, LCOE does not reduce below \$0.10/kWh. This confirms that system designs to meet the high-end demand estimate are unaffordable to the average household in Northern Sierra Leone at present. As expected, the high-end estimate is aspirational and is more likely to represent future load following income and demand growth. Observing that even a 1% capacity shortage creates a noticeable LCOE decrease in the high-end demand case, however, the same capacity shortage is simulated in the low- and mid-level demand cases to test its effects.

The resulting design point spaces for no capacity shortage and 1% capacity shortage are shown in Figures 6.14a and 6.14b respectively. Note that the optimal corner and boundary of technical feasibility expand inwards towards lower PV and battery sizes slightly in the 1% capacity shortage case. There is significant overlap between the affordable and feasible design point spaces for low-end and mid-level estimated demands. As such, it would be possible to select a design which is affordable and feasible to meet low-end demands and which can also meet mid-level demands as electricity use increases. For instance, in the 1% capacity shortage case, the system could be initially oversized for low-end demand at 50 kW solar and 108 kWh storage. This increases the initial LCOE to \$0.24/kWh – however, this is still within the energy budget for energy poverty prevention at this usage



(a) No capacity shortage.



(b) 1% capacity shortage

Figure 6.14: Feasible design point spaces for C4 at low, mid-level, and high-end demands with: (a) no capacity shortage, and (b) a 1% capacity shortage. In the 1% capacity shortage case, smaller systems become feasible (i.e. the corner of the feasible space moves inward). In each case, the mid-level space is shown superimposed on the low-end space; the location where they meet is marked with a white line for clarity. The inside corner of each space indicates the boundary of technical feasibility, while the upper boundaries represent affordability limits to prevent energy poverty.

level. As demand then increases towards the mid-level estimate, the system will still be able to meet demands thanks to its initial oversizing.

As there is no affordable way to currently meet high-level demands without risking energy poverty, this space can be treated as a future demand target. A modular growth trajectory between the selected design point and the cost optimised system is suggested to eventually meet high-end needs. In the 1% capacity shortage case, modular expansions can be planned between the 50 KW PV, 108 kWh storage system and an eventual 132 kW PV, 302 kWh storage system to meet high-end demands. Once the community load exceeds mid-level to the point where the initial system has an unacceptable amount of unmet load, an expansion of 30 kW PV and 65 kWh batteries could, for instance, be added. Three such expansions would be required to reach high-end load capacity, as shown in Figure 6.15. As the prices of PV and batteries are decreasing over time, the costs of each modular expansion are calculated based on component price forecasts. Using the estimate of $170/kW for PV as per NREL advanced investment scenario [300] and $150/kWh for battery storage as per NREL projections [301], the capital technology cost for the PV and storage to accomplish one modular expansion in 2030 would be $14,850. This is less than would have been paid to install this extra capacity initially (i.e. $19,330), and is likely to be more affordable to consumers by that point.

This modular addition of generation and storage over time assumes that conductor will be initially installed such that it will be able to accommodate the final targeted system usage, under the premise that stranded infrastructure is to be avoided. It could be evaluated whether replacing the conductor with a higher rated cable as needed would be more cost effective despite the wastage – this is left for future study. Nevertheless, in this case, wiring is specified such that it can accommodate the worst-case power (i.e. current drawn) for the highest-demand case. The worst case community load peak is around 30 kW. Assuming distribution at 380 V, the maximum current on any conductor in this network would in the worst case be around 79 A. This can be accommodated by a number of ACSR cabling types, such as the squirrel type [302]. While it is incredibly difficult to source

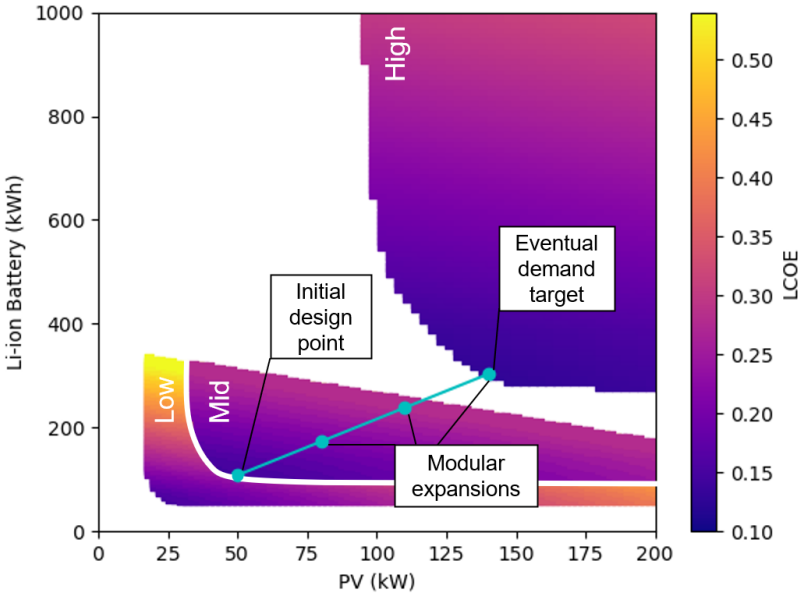


Figure 6.15: Trajectory from low-end demands to high-end demands via modular expansion in the 1% capacity reduction case for C4. Initial design point and infrastructure sizes following each modular expansion marked in cyan.

ACSR costs per meter without requesting a quote for purchase, assuming a rate of \$27.60/m [303] and 2,928 m (based on 2,169 m for the MST as found previously multiplied by 1.35), distribution line for C4 will cost approximately \$81,000.

Finally, land requirements to implement the system and each subsequent modular expansion are cross-checked. Observing either satellite imagery or land-use data from OSM, C4 is surrounded by forested land. To implement ground-mounted PV would require finding or clearing an adequate area to place the panels without shadowing by nearby trees. Assuming approximately 6.4 m^2 of solar modules per 1 kW capacity installed [304], the first 50 kW of PV installed would require 320 m^2 of solar panels. Neglecting the difference in size made by the optimal ground-mounting angle of 13°, this would require an approximate 18×18 m area over which to mount solar panels. Assuming a 50% excess in area is required to allow room to navigate around the installation and prevent shadowing from surrounding trees, these solar panels could be installed in a 22×22 m area. Each 30 kW expansion of installed PV would require an additional 192 m^2 of solar panels, or a 17×17 m area to allowing for 50% excess space beyond solar panels. These area sizes are feasible given the amount

of available uninhabited land surrounding C4, as evident in the previous Figure 6.7.

6.2.1 Results Summary

To synthesize and summarize these results, in this case study area, it is determined that 12 local micro-grids will be the best solution in the near-term, with 11 outlying homes receiving SHS. These micro-grids can be eventually interconnected through a transmission network if desired, which will require a minimum of 11 km of conductor. Each micro-grid will be best served through a solar and battery system, given the lack of neaby high-wind areas and hydroelectric opportunities. For each local grid, low-, mid-, and high-end demand estimates are generated and used to optimally size PV and storage. The LCOE of resulting grids is compared to average household income to understand whether each demand estimate is feasible without introducing energy poverty risk. For the example community C4, the high-end estimate is not affordable in the near-term given local wealth distribution and expenditure levels, while the low- and mid-level demand estimates appear feasible without energy poverty risk. As such, a design point is selected at 50 kW of PV and 108 kWh of storage resulting in an LCOE of $0.24/kWh for low-end demand and $0.14/kWh for mid-level demand, allowing for a 1% capacity shortage. This design point allows for demand growth between levels without energy poverty risk in either demand level. High-end demand is treated as a future goal, given its current unaffordability; three modular expansions of 30 kW PV and 65 kWh are required to meet high-end demand. Based on projected technology costs, one modular expansion in 2030 would require a $14,850 investment in additional PV modules and battery storage. The micro-grid can be wired based on peak expected current at eventual high-end demand to minimize eventual infrastructure stranding using "squirrel" ACSR, assuming 380 V distribution.

6.2.2 Utility of the Framework

The results of this case study application demonstrate a spatial feasibility design framework which presents actionable and concise policy-ready results. A high

level of spatially specific detail is preserved; for instance, a much more detailed picture of implementation requirements and future costs is generated than the results previously presented in Figure 6.1.

This framework is implemented using scalable and accessible methodologies. Clustering and graph theory methods are implemented to identify electricity communities and specify initial grid designs using open-source and user-friendly Python tools. The home-level input data to achieve this can be produced through the scalable methods achieved in Chapters 3 and 4. GIS analysis of generation potentials is undertaken using open-source datasets (i.e. AFREP, Renewables.Ninja), platforms (i.e. GlobalSolarAtlas, GlobalWindAtlas), and software (i.e. QGIS). Optimisation of generation and storage sizing and costs is accelerated through HOMER, an industry-standard micro-grid design tool, for which commercial licenses are available at approximately $500 per year, which should be feasible within the context of governmental budgeting for electrification planning. Demand estimates required to facilitate sizing are produced using openly-available MICS and IHS data, alongside the open-source RAMP modelling framework as described in Chapter 5.

One strength of this framework is that it can be tailored to the design philosophy of the electrification planning team. That is, the design philosophy is not "hard coded" into the modelling framework. The pathway selection presented in the case study above assumes a "least-regrets" philosophy which aims to minimize asset stranding and retrofitting frequency by affordable initial oversizing. This is a choice made by the designer, and not inherent to the methodology. In some cases, stranded assets may indeed result in a least-cost system, and the designer is free to choose to strand assets by instead applying their own design philosophy. Indeed, the same pathway design approach could be applied by selecting the lowest cost point in each demand level, requiring more frequent retrofitting with modular generation and storage increases. Similarly, conductor capacity could instead be selected to only accommodate initial load and not high load, which would in turn require re-wiring the distribution grid at each modular increase but would make

the grid cheaper in earlier stages. These design philosophy choices can be made by the designer, based on their goals and constraints.

Inherent to this framework however is the notion of demand growth autonomy. While the designer may evaluate potential demand growth pathways to estimate costs, this framework necessarily presents spaces of choices and not discrete futures. Modular additions can be implemented if and as needed by the community using this framework, allowing communities to grow their usage as they desire. Low, middle, and high-end demand estimates are not tied to specific times in the future; they simply estimate how demand may grow as wealth shifts over time. The particular trajectory through the feasible design point spaces can be in effect co-designed through the community and the designer over time as needs evolve.

6.2.3 Current Limitations and Future Expansions

Most of the discussions of affordability of energy in these results focused on the notion of a price-per-kWh tariff. However, this may not be the best tariff option for newly electrified communities in practice. A price-per-kWh structure may depress load amongst newly connected poor consumers seeking to minimize their bills, preventing upward demand mobility.

An alternative price-for-peak-power tariff structure can be considered based on the energy budget for the average household in Sierra Leone. For instance, the energy budget of 33,400 Le (i.e. $4.33 USD) could represent a standard, flat-rate monthly tariff for any usage of appliances at or below a designated medium level of peak power consumption (i.e. televisions, fans). For lower-income families, a lower flat-rate tariff could be introduced permitting only low-powered appliances (i.e. lighting and charging), and for high-income families, a higher-power tariff could be introduced permitting higher-powered appliances (i.e. refrigeration and ironing). Importantly, these tariffs could allow *unlimited use* of such appliances to encourage increased energy usage. Another potential modification could be to allow special rates for those who wish to use high-powered appliances *strictly outside of the peak evening window*. This could be useful for, for instance, businesses

seeking to run mills or irrigation pumps during the day, when solar generation is more plentiful and power usage is not peaking. Such tariffs could be considered to encourage upward demand mobility and efficient use of power generated through renewable variable sources, and merit future study.

A number of aspects of topology design merit future study to enhance this approach. First, distribution and interconnection lines are more likely to follow existing roads than to pass through more difficult terrain (i.e. forests) or through private property. This will be lower cost than clear-cutting new paths for pole installation in vegetated areas or acquiring land rights or negotiating construction on privately-owned land. Similarly, routing lines through certain geographic features (i.e. waterways, natural protected areas such as national parks) entails higher costs and more difficult permissions processes than if these are avoided. It can be explored how to incorporate this cost in MST graph construction by raising the weight of Delaunay triangulation edges which intersect with these features using a cost function. This could discourage the MST from selecting these pathways and thereby optimise for a more realistic least-cost system instead of a minimum theoretical least-distance system.

Additionally, the notion that the road-based grid distance was found to represent approximately 135% of the MST of the Delaunay triangulation here was an interesting and useful result. However, obviously, this is found over a small case study area. It would be interesting to study this over a broader region, and over different contexts, to test where it holds true more generally. A set of correction factors for conductor distance could, for instance, be generated over different community configurations and geographic scopes.

There are also improvements to be made in terms of component costing. First, the results presented here assumed ground-mounted solar, but solar PV can also be installed on distributed rooftops. It could be studied, for instance, whether any community cluster has adequate feasible roof space to meet its energy needs through rooftop PV using the sizing of the roof footprints generated in the citizen science effort in Chapter 3. Second, across all grid technologies, cost projections

vary significantly based on the source consulted, and can be difficult to find at all for certain components (i.e. ACSR). As always, the results of a modelling method depend on the input parameters selected, and any uncertainty in these manifests as uncertainty in results. As such, in re-applying these methods, cost figures should be chosen carefully and from robust sources wherever possible. A sensitivity analysis of these methods for variance of inputs is an important avenue for future work. Finally, it is worth noting that the LCOE calculations used to limit system affordability do not take the conductor costs into account. These LCOE calculations, completed in the HOMER optimisation process, are based on the generation and storage bill of materials. Integrating conductor costs would therefore further limit the size of each feasible space, at it would raise the total system cost. This is left as a future expansion.

Further study is also needed on the capacity shortages introduced in mini-grid sizing. While allowing for a small capacity shortage can help to save money, it needs to be studied whether the resulting short outages or unmet loads are acceptable in the community context. Perhaps particular peak loads are absolutely necessary to the community – for instance, peaks caused by higher powered productive use technologies which drive the community economy. Even a small capacity shortage which affects these technologies could have a strong impact on local incomes and may be unacceptable, particularly if outages could damage productive use equipment. No consideration was given to the timing of capacity shortages in this study, or to which loads would be shed to enable them; this was handled as part of the internal optimisation in HOMER. However, future work should consider this explicitly to ensure that load is shed at times which are acceptable to the community.

6.3 Key Outcomes

This chapter has proposed a spatial design framework to account for home-level specificity in scalable spatial electrification design. This framework integrates electricity community clustering, graph-theory-based network estimation, GIS

analysis for generation and storage specification, and cost optimisation for design-point selection and pathway planning.

The framework has been applied in a case study region in Northern Sierra Leone. It is determined that 12 local micro-grids will be the best electrification solution for this region in the near-term, with 11 outlying homes receiving SHS. Within this region, sizing is presented for a specific community cluster, labelled C4. Via cost optimisation and mapping of the design point space for C4, an initial design point of 50 kW of PV and 108 kWh of battery storage is selected to meet both low-end and mid-level demands without energy poverty risk. To accommodate high-end demands in the future, three modular expansions of 30 kW PV and 65 kWh of storage can be installed as demands evolve.

This framework is shown to account for home-level settlement configuration using realistic data inputs and scalable low-cost methodologies. It navigates the transition between preliminary grid type scoping and local detailed planning by offering a feasibility-level scalable design solution.

7
Conclusions

Contents

7.1 Concluding Discussion

This thesis has endeavoured to address the question:

Can we design affordable electricity access systems suited to local spatial context and needs at the scale required to close the global electricity access gap?

Given the enormity of this question, it has been broken down into four constituent sub-questions, three of which address data gaps and one of which addresses design processes. First, can rural off-grid populations be accurately mapped for electrical system design? Second, can methodologies be developed to accomplish this mapping rapidly at a global scale? Third, can the diverse energy needs of these populations be estimated with spatial specificity? And finally, can appropriate least-cost electrical systems be designed to match community spatial context?

In answering these questions, approaches to tackle each have been developed which can be synthesized into a unified and scalable spatial design process, as

illustrated in Figure 7.1. First, accurate and complete home-level spatial data, a key input to the electrical design process, are collected through citizen science. These data are scaled to achieve global coverage rapidly and at low cost through computer vision. Demand is estimated from the bottom-up using existing large-scale datasets accounting for spatial specificity and intra-community diversity. Finally, grid types and technologies are designed using a practical spatial pipeline using clustering, graph theory, GIS, and cost optimisation to generate policy-ready and actionable rural electrification designs at scale.

Locating Potential Consumers

It is impossible to design appropriate electricity access systems without first knowing where off-grid populations are located. Without accurate, complete, and up-to-date location data, the number of potential connection points is unknown, so demand cannot be estimated accurately and generation and storage cannot be sized correctly. Connection point locations are required to design a best-fit grid topology based on the configuration of settlements, which vary hugely in rural off-grid areas of LMICs. Surveying state-of-the-art georeferenced data sources, there is a gap for home-level location data which are accurate, complete, and up-to-date to enable spatial design of electricity access systems tailored to community context.

To address this data gap, this thesis first tests citizen science as a means to map rural homes in LMICs. Satellite imagery is used as input data in a citizen science mapping experiment which engaged members of the public to annotate homes in satellite imagery of rural Kenya, Sierra Leone, and Uganda. The project, called "Power to the People", was executed successfully and mapped approximately 1,267 km^2 at an average rate of 7 km^2/day over the course of 179 days. Citizen science home annotations achieved a recall of 93% and precision of 49%, which could be increased to 69% through clustering. The estimated cost for this approach is \$20.84/km^2. Through an evaluation survey, it was found that this approach adheres to the principles of citizen science by providing an

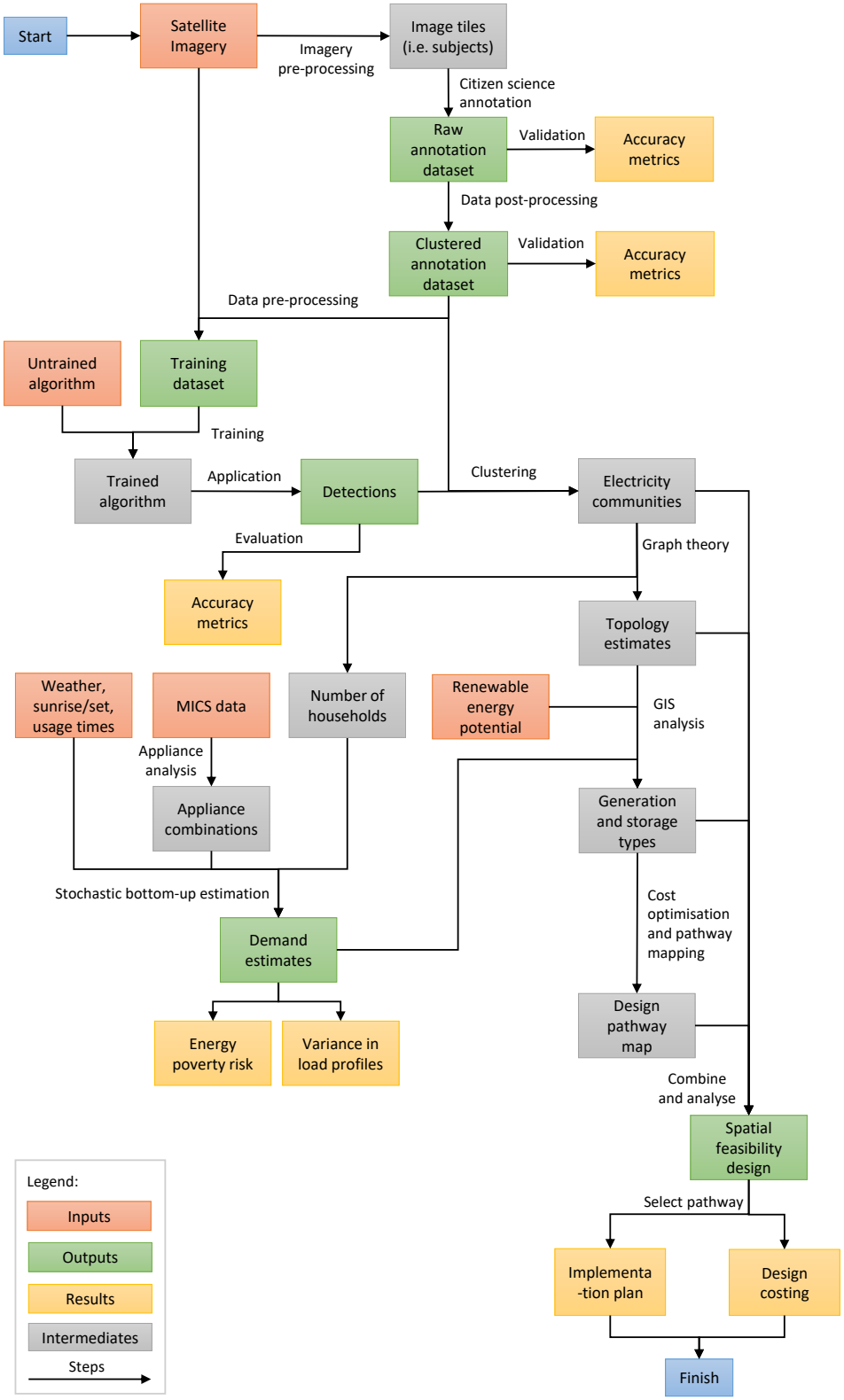


Figure 7.1: Overview of the entire spatial design process proposed in this thesis.

enjoyable, educational, and beneficial experience for citizen contributors. This work answered the first sub-question.

While the proposed citizen science mapping approach offered improvements over conventional methods, it did not achieve speeds and costs appropriate to the size and urgency of the full global electricity access gap. For instance, to map the entire rural area of LMICs globally at these rates would take 30,000 years and cost over $1.6 billion. While it can be argued that this would be financially worthwhile, the infeasibility of the time constraint is hard to debate. Even accounting for the possibility that extra citizen scientists could be recruited to accelerate the process, over 200 million volunteers would need to be mobilized to map the entire rural area of LMICs at home level within a year, which is highly unrealistic compared to the volunteer base on even the most popular citizen science efforts.

Computer vision is therefore explored as a means to accelerate home level mapping at low cost to achieve global scale. A Faster R-CNN object detection model was trained to map rural homes in Kenya, Uganda, and Sierra Leone using satellite imagery and citizen science annotations as training data. The trained detector achieved a precision of 67% and recall of 36% when trained on the clustered citizen science annotations; recall could be improved to 57% by training on raw annotations instead of annotation clusters. It achieved home-level mapping at a rate of 42,938 km^2/day. While an initial investment in training data and computing infrastructure is required for this method, as the area mapped expands, the costs per km^2 rapidly approach the cost of the input satellite imagery (here, $\$10/km^2$). This work answers the second sub-question by generating a proof-of-concept for the ability of computer vision to rapidly map homes at a global scale.

Estimating Demands With Spatial Specificity

To design affordable electricity access systems of an appropriate size, one must also understand the anticipated needs of the off-grid consumers. These vary spatially based on cultural variance and the distribution of wealth among other factors. If these needs are under-estimated, the system could quickly become

damaged by outsized loads, requiring near-immediate retrofitting. If these needs are over-estimated, the resulting system could be unaffordable to consumers and implementers alike, exacerbating system underuse and the increasing the likelihood that community members elect not to connect to the system.

Surveying the literature, stochastic bottom-up engineering-based demand estimation emerges as the most appropriate approach to understand electrical needs in rural off-grid LMIC contexts. However, high-resolution input data for such methods (or means to collect these) are scarce. There is a paucity of accurate and spatially specific appliance time-of-use data in LMICs, which are typically used as an input in stochastic bottom-up demand estimation in HICs. Costly local surveys can be used to fill these data gaps by asking off-grid community members about their energy needs and aspirations directly. However, the resulting data are often inaccurate, as off-grid communities tend to often overestimate their future consumption given their limited experience with electricity. A gap was therefore identified for spatially specific demand estimation approaches which account for the diversity of needs within and between communities, and which produce realistic estimates quickly and cheaply.

To address this, a spatially specific demand estimation approach is developed for rural off-grid communities leveraging existing georeferenced empirical data from large-scale development surveys. MICS data on appliance ownership are used as an input data source for stochastic bottom-up demand estimation. Through a case study application of this approach in Sierra Leone, which leverages the 2018 MICS dataset and the RAMP stochastic demand estimation framework, the underlying premise of spatial variance in demands is validated, and a realistic and spatially specific demand estimates are generated for synthetic communities representing each region of Sierra Leone. This work thereby addresses the third sub-question.

Scalable Home-Level System Design Including Future Pathways

Finally, spatially specific location and demand data are applied in design. Existing models for off-grid electrification design typically fall into one of two categories: (1) local-scale generation and storage sizing and optimisation; or (2) large-scale

grid-type optimisation over units of land area. The first type typically assumes that the number of connections is known to the designer, while the second does not take settlement style or grid topology into account when selecting types. Some methods attempt to do handle both local and large-scale considerations, but these require inordinate data inputs and have high computational complexity.

This thesis therefore proposes a practical spatial design framework to generate feasibility-level grid designs which account for settlement configuration. DBSCAN-based clustering and hyperparameter tuning methods are applied to identify electricity communities best suited to unified architectures. Then, graph theory is used to generate topology estimates, wherein the MST of the Delaunay triangulation of connections is treated as a lower bound on conductor required, and a radial connection to a hub at the community centroid is treated as an upper bound on conductor required. A similar MST between cluster centroids is used to propose an eventual cluster interconnection network. GIS-based analysis of generation potentials is applied to select appropriate generation types at the cluster location. Cost optimal generation and storage sizes are then identified to account for low-end, mid-level, and high-end estimated demands in the cluster, and the boundary of technical feasibility to meet these needs is located. An affordable and feasible space of generation and storage design points is generated for each demand estimate, and LCOE is visualised over this. This allows the identification of overlaps, cost impacts, and pathways for modular grid expansion. This framework is low cost and uses publicly available data. It can be automated over a desired geographic area. This work thus addresses the final sub-question.

Thesis Contributions

This thesis has resulted in the following contributions and publications:

1. A novel dataset of underrepresented home styles in rural Kenya, Sierra Leone, and Uganda. This can be directly applied in GIS analysis or used with accompanying VHR EO satellite imagery to train automated building detection and mapping algorithms (Chapter 3).

- *Leonard, A., Wheeler, S., McCulloch, M. (2022). Rural Home Annotation Dataset Mapped by Citizen Scientists in Satellite Imagery. Data in Brief, 42, doi: 10.1016/j.dib.2022.108262*
- *Leonard, A., Wheeler, S., McCulloch, M., (2022), "Rural Home Annotation Dataset Mapped by Citizen Scientists in Satellite Imagery", Mendeley Data, V1, doi: 10.17632/xw6gr8p2cn.1*

2. The application of citizen science and computer vision to rural off-grid dwelling mapping for electrical system design. These methods are shown to suit the universal electrification problem at differing scales (Chapters 3 and 4).

   - *Leonard, A., Wheeler, S., McCulloch, M. (2022). Power to the people: Applying citizen science and computer vision to home mapping for rural energy access. International Journal of Applied Earth Observation and Geoinformation, 108, 102748, doi: 10.1016/j.jag.2022.102748.*
   - *Leonard, A., Wheeler, S., McCulloch, M. (2022). Evaluating "Power to the People": Best practices for positive community impact in remote mapping citizen science projects. Preprint available at SSRN 4052549, doi: 10.2139/ssrn.4052549.*

3. The application of large-scale socioeconomic and demographic datasets in stochastic bottom-up demand estimation for Sierra Leone (Chapter 5).

   - *Leonard, A., Wheeler, S., McCulloch, M. (2022). Estimating Potential Electricity Demand in Low-Income Countries: A spatially-disaggregated stochastic approach illustrated in Sierra Leone. 17th Conference on Sustainable Development of Energy, Water, and Environment Systems. (Accepted – forthcoming).*

4. A framework of spatial methodologies for spatial feasibility design of rural electrification systems (Chapter 6). This framework uses the previously generated location and demand data.

- *Leonard, A., Wheeler, S., McCulloch, M. (2020, August). Geospatial Clustering and Network Design for Rural Electrification in Africa. In 2020 IEEE PES/IAS PowerAfrica (pp. 1-5). IEEE., doi: 10.1109/PowerAfrica49420.2020.9219908.*

5. An approach to map feasible design points for generation and storage in autonomous rural electrical systems and chart the design implications of demand evolution over time, including effects on affordability (Chapter 6). *(Publication forthcoming.)*

General remarks

Underlying the work of this thesis is the power of spatial data analysis to connect qualitative and quantitative elements of electricity access system design. Spatial data and approaches bridge the gap between the contextual and the quantifiable. They can translate across traditional research disciplines by rooting scientific and social phenomena in a geographic context. This is powerful in the pursuit of universal electrification, given the tension of local specificity and global scale which requires qualitative social awareness and generalizable quantitative methods.

While the full spatial design process presented in this thesis may appear complex (see Figure 7.1), it is earnestly intended to be practical, cost-effective, and scalable, ready to be deployed by policy-makers planning electrification in LMICs. The "recipe" for this analysis can be applied as follows.

1. *Acquire data*: Download satellite imagery, MICS data, OSM data, renewable energy potentials, land use data, and locations of existing grid infrastructure.
2. *Locate homes*: Review the existing OSM building location data. If gaps exist, homes must be located. If appropriate home-identification training data for computer vision is available online, immediately train a detector and map the area with computer vision. If not, run an online citizen science project to map (samples of) the area and then scale the mapping with computer vision.

3. *Identify electricity communities*: Cluster the home locations to identify communities requiring unified electrical systems. Assign each cluster a technology type based on its proximity to grid infrastructure and size.

4. *Study appliance ownership and wealth distribution*: Analyse MICS data to determine spatially-disaggregated appliance ownership patterns and wealth distribution across the region of interest. Analyze regional household expenditure data to set tariff price limits for energy poverty prevention.

5. *Estimate demand*: Estimate demands by applying MICS appliance ownership data and stochastic bottom-up engineering estimation to the number of households in the cluster. Generate additional estimates by adjusting the wealth distribution to represent possible low-, mid-, and high-level demands.

6. *Select generation and storage types*: Analyze generation potentials at the location of interest from spatial climactic and topographic data. Compare site-specific infrastructure costs for each. Select types based on feasibility and cost thresholds specific to the context.

7. *Optimize generation and storage size*: For each cluster and each estimated demand level, find the cost-optimal generation and storage size. Identify the boundary of technical feasibility to meet demands (i.e. the curve through the storage and generation size plane beyond which demands can be met).

8. *Simulate costs over feasible and affordable design space*: Simulate costs extending outward from the boundary of feasibility. Limit the space based on expected affordability of the energy to those in the region. If the space for a certain demand estimate contains no area once it is affordability-limited, it can be assumed that this level of demand is unrealistic at present; instead, it is a future goal for demand expansion after wealth increase.

9. *Chart design pathway*: Select an initial design point based on the desired philosophy (e.g. least present cost, least regrets, longest time before retrofit). Chart a possible modular expansion or reliability reduction pathway to enable

autonomous community-led demand growth. Calculate resulting tariffs using either a price-per-kWh, price-per-peak-power, or other tariff structure.

7.2 Future Work

Throughout and as a result of this work, the following areas have been identified as avenues for further research:

Citizen scientist behaviour and performance: In Chapter 3 it was found that, when locating homes, citizen scientists sometimes interpreted instructions differently or opted not to follow certain steps of the annotation workflow. The user experience of citizen science projects could be studied via A/B testing in future work to determine which tool types and instructions improve their understanding and thereby data quality. Additionally, the annotation abilities of citizen scientists were found to be superior to paid alternatives like Amazon Turk workers. It could be investigated whether this is due to emotional investment in the scientific process, lack of time pressure amongst volunteers compared to paid workers, or other possible reasons through surveys or interviews with contributors.

Citizen science in public service applications: By definition, citizen science typically focuses on academic scientific projects. However, the results of the citizen science mapping effort and subsequent evaluation survey in Chapter 3 indicate that a significant portion of the citizen science community are motivated by the possibility of real-world application and impact. As such, it would be interesting to experiment with citizen-science-style projects in the public service realm to enable humanitarian response, infrastructure design, and so on. This is an exciting avenue for future study.

Computer vision performance improvements: To improve the results achieved by the computer vision algorithm in Chapter 4, a number of training data post-processing choices could be tested. For instance, since crowded images had the lowest-quality citizen science annotations, it would be interesting to exclude these from training, and test whether computer vision algorithms trained only on non-crowded images could still accurately detect homes in crowded images. Alternatively,

it could be tested whether clipping images closely around the annotations used for training could minimize the number of non-labelled homes the algorithm was exposed to as "false" samples. This could also create class imbalance issues between positives and negatives, and so this must be accounted for in experiment and testing. Since the detector trained in this work also performed poorly in certain, but not all, agricultural contexts, it could be investigated which contexts present particular challenges for home detection, and whether this is due to data imbalances, challenging backgrounds, or other factors. This could be accomplished by evaluating detector accuracy across particular data subsets. Additionally, separate detectors could be trained on a per-country or per-context basis (e.g. agricultural, refugee, clustered community) to determine whether this would improve accuracy..

Appliance type diversity: In the demand estimation approach developed in Chapter 5, each instance of each appliance was modelled identically for each type of user. No variance was included to account for different versions of the same appliance (e.g. with different power ratings). Such variation is likely in practice, and its impact on demand estimates represents an interesting avenue for future study.

Appliance acquisition priority: The results of the appliance ownership analysis completed in Chapter 5 indicate a potential sequence of appliance acquisition. However, this was not confirmed, as the date of appliance purchase was not recorded in MICS. It would be interesting to investigate whether this appliance acquisition order holds true, and the time between acquisitions, to enable better future demand growth planning. Additionally, it was noted that rural customers were more likely to not acquire appliances (besides lighting) compared to urban customers. It should be investigated whether affordability, accessibility, cultural norms, or some other factor are encouraging rural Sierra Leoneans not to acquire appliances following electrical connection.

Sensitivity analysis: An analysis can be undertaken to verify the sensitivity of the stochastic demand estimation approach developed in Chapter 5 to its input data. While MICS data is likely to be highly accurate, data collected through

expert interviews are less certain. By evaluating whether the approach adopted is particularly sensitive to certain parameters, time can be better invested in verifying the most critical input data to obtain more precise demand estimates.

Conductor length approximation: In Chapter 6, road-based distribution network estimates were found to be 35% longer on average than the MST of the Delaunay triangulation of homes. It would be interesting to study whether this proportionality holds true in different contexts, and if not, to generate a series of proportionality constants appropriate to different settlement types to enable more accurate conductor distance estimates. Alternative distance types, such as the Manhattan distance, could also be explored in terms of their similarity to the road-based estimate.

Interconnection to accommodate demand growth: The demand trajectory mapping in Chapter 6 primarily investigated modular expansion and capacity or affordability reduction as mechanisms to accommodate demand growth. One factor which has not been explored is interconnecting initially autonomous systems to meet growing demands. With the increased diversity from a larger consumer base, there is the possibility that increased demands could be met without additional retrofitting or with lower-cost expansions. This option should be investigated in future study.

Tariff structures to promote demand growth: The demand growth planned for in Chapter 6 can be limited by standard price-per-kWh tariffs, which can depress the demand of poor households living hand-to-mouth. Novel tariff structures based on peak allowed power usage, time of day in which usage is allowed, or both can be investigated as mechanisms to allow demand growth without major tariff increase, promoting upward energy mobility and consequent beneficial development outcomes.

Capacity shortage timing: Introducing a small capacity shortage during mini-grid sizing can create more affordable energy access, as discussed in Chapter 6. However, it should be studied whether such shortages are acceptable to the community. If there are particular productive use technologies critical to the community economy which can be damaged by load shedding, it should be

investigated whether adequate non-essential load can be dropped to keep this equipment online despite the capacity shortage.

Appendices

A

Evaluation Survey Questionnaire

Power to the People: Your feedback

Page 1: The Power to the People team needs your feedback!

We are evaluating the Power to the People project to better understand the following:

- Who is contributing to Power to the People and why do they choose to engage?
- What are the impacts of engaging with Power to the People for our contributors?
- What are the benefits and challenges of contributing to Power to the People?

We want to hear your feedback so we can improve Power to the People and do better citizen science in the future. Please complete this short survey and let us know your thoughts. It should take around 5-10 minutes to complete.

How will my data be used?
We will use your data to better understand who is engaging with the Zooniverse Power to the People project and explore your experiences with the project. This feedback will inform improvement and evaluation of Power to the People. The results will be published in a case study that will be publicly shared through the Power to the People online platform and University of Oxford webpages. Your data will also inform academic publications (i.e., scientific journal articles) about the project. Please note, your responses will be anonymous - no names will be collected or published. For Zooniverse's User Agreement and Privacy Policy, please visit www.zooniverse.org/privacy. For further details about how your data will be used in this survey, see our Privacy Notice.

Who is conducting this evaluation and why?
Alycia Leonard (Energy and Power Group, University of Oxford) is conducting this evaluation to explore the impact that Power to the People has on contributors, the benefits and challenges of engaging with the project, and to learn about the diversity of contributors.

The Energy and Power Group wants to learn as much as possible from your experiences to conduct more, and better, citizen science studies in the future. More information about this survey is available on the participant information sheet. If you would like to add any additional comments or thoughts to your responses, or you have any concerns about this evaluation, please get in touch at alycia.leonard@eng.ox.ac.uk. Find out more about the Energy and Power Group at the University of Oxford here: https://epg.eng.ox.ac.uk/

Can everyone complete this survey?
This survey is intended for contributors to the Power to the People project. You must be at least 11 years old to complete this survey. Consent from a parent or guardian will be needed for participants under the age of 18. Please note that participant information sheets specifically prepared for younger participants and their parents/guardians are available. If you are under the age of 11, thank you for contributing to Power to the People but please do not complete this survey. You can instead send any feedback you have to alycia.leonard@eng.ox.ac.uk with parental permission.

Consent
Please indicate below your consent to completing the survey, and to your responses being used in our research.
C.1: Please note that you may only participate in this survey if you are 11 years of age or over.

- I certify that I am between 11 and 18 years of age.
- I certify that I am 18 years of age or over.

C.2: If you have read the information above and agree to participate with the understanding that the data (including any personal data) you submit will be processed accordingly, please check the relevant box below to get started.

- Yes, I agree to take part.

C.3: If under 18: Please show your parent or guardian the information above. If they agree to your participation, please have them tick the relevant box below so that you can get started. If they do not agree to your participation, you cannot proceed.

- Yes, I consent for my child to take part.

Page 2: Your experience on Power to the People

1: How would you rate your experience contributing to Power to the People? (Select one)

- Excellent
- Good
- Average
- Below Average
- Poor

2: How frequently do you contribute to Power to the People? (Select one)

- Daily

- Several times a week
- Once a week
- 2-3 times a month
- Once a month
- Less than once a month
- Only once
- Not sure
- Other (type)

3: How did you find out about Power to the People?

- Zooniverse page
- Social media (i.e. Twitter)
- From a friend or colleague
- Through my school
- Through an extra-curricular or volunteer programme
- Other

4: Why do you engage with Power to the People? (Select all that apply)

- I enjoy learning about energy and electricity access.
- I enjoy learning about satellite imagery analysis and/or computer vision for rural mapping.
- I want to contribute to scientific research.
- I am generally interested in science and/or engineering.
- I enjoy identifying homes and finding interesting features in images.
- I enjoy being part of a like-minded research community.
- I want to contribute to projects with real-world impact.
- I am contributing as part of a class or a volunteer programme.
- I find it entertaining.
- Other (Type)

5: Did you learn anything through taking part in Power to the People? (Select one)

- Yes
- No
- Maybe

5a: [If 5 = Yes] What did you learn? (Type)

5b: [If 5 = No OR maybe] Tell us why. (Type)

6: What Power to the People information pages have you visited? (Select all that apply)

- About or Learn More pages (i.e. Research, The Team, Education, FAQ)

- Field Guide
- Tutorial
- Power to the People Statistics
- Other (type)

7: Have you used Power to the People Talk? (Power to the People Talk is an online chat forum for discussion about the project) (Select one)

- Yes
- No
- Not sure

7a: [If 7 = Yes] What did you use Talk to do? (Select all that apply)

- To ask a question.
- To post something interesting you have found.
- To talk to a researcher.
- To talk to other volunteers.
- Other (type)

7b: [If 7 = No] Why not? (Select all that apply)

- I did not know it existed.
- It does not interest me.
- I do not have the time.
- It is not helpful to me.
- Other (Type)

8: Is there anything that prevents you from spending time on Power to the People? (Select one)

- Yes
- No

8a: [If 8 = Yes] Please explain what prevents you from spending time? (Type)

Page 3: Power to the People and beyond!

9: Have you promoted or shared Power to the People with others? (Select one)

- Yes
- No
- Not sure

10: Since engaging in Power to the People, have you volunteered in other projects on Zooniverse? (Select one)

- Yes
- No
- Not sure

10a: [If 10 = Yes] Select any other Zooniverse projects you have volunteered with since engaging with Power to the People (Select all that apply from list of all projects)

11: Has Power to the People prompted you investigate any concepts further by looking things up, seeking further information, or doing your own research? (Select one)

- Yes
- No

11a: [If 11 = Yes] What did you decide to investigate further? (Select all that apply)

- Energy access in sub-Saharan Africa
- Electrical grid design
- Rural mapping
- Satellite imagery
- Computer vision and AI
- Information about Uganda, Kenya, or Sierra Leone
- Rural housing styles
- Other (Type)

11b: [If 11 = Yes] How did you do your own investigations? (Select all that apply)

- Internet search
- Watching videos
- Reading books
- Reading news or magazine articles
- Listening to podcasts
- Asking teachers, professors, or experts
- Discussing with friends and family
- Doing a real-life experiment
- Other (Type)

Page 4: Demographic information

Please complete the following demographic information to help us understand who is engaging with the Power to the People project. If you do not wish to provide this information, please leave the answers blank.

12: What country do you live in? (Select one from a drop down list)

13: What is your highest level of education (Select one)

- No formal education

- Secondary/high school or equivalent
- Vocational training (i.e. job-specific training leading to a certificate or diploma)
- Bachelor's degree (e.g. BA, BS)
- Master's degree (e.g. MSc, MA, MRes)
- PhD or other advanced professional degree
- Prefer not to answer
- Other (type)

14: What is your current employment status? (Select one)

- Employed (full time)
- Employed (part time)
- Retired
- Student
- Unemployed
- Prefer not to answer
- Other (type)

14a: [If 14 = Employed full time or part time] Please describe your occupation: (Type)

15: Do you have a background in science or engineering?

- Yes
- No
- Not sure
- Prefer not to answer

15a: [If 15 = Yes] Please explain a bit about your background in science or engineering: (Type)

15b: [If 15 = Not sure] Please specify: (Type)

16: What is your age? (Select one)

- Under 18
- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 55-64 years old
- 65-74 years old
- 75 years or older
- Prefer not to answer

17: What best describes your gender? (Select one)

- Woman
- Man
- Non-binary
- Prefer to self-describe (Type)
- Prefer not to say

Page 5: Final thoughts

18: Do you have any advice for us to make Power to the People more engaging? (For example, things you want more or less of, things you thought were missing, etc.) (Type)

19: Are there any aspects of Power to the People that you found particularly enjoyable? (Type)

20: Would you like to tell us anything further about your experience of Power to the People or Zooniverse? (Type)

Page 6: Survey Complete

Thank you for completing this survey. Your response has been recorded.
Please download your completion receipt here. You will need to provide the information on this receipt if you decide to withdraw your participation at any point in the future. We will use your feedback to better understand who is engaging with the project and why, as well as exploring contributor experiences with the project. This feedback will inform the evaluation and improvement of Power to the People. The results will be made available to the Power to the People community on the project webpage and through the University of Oxford webpages.
If you have any questions about this survey, please email alycia.leonard@eng.ox.ac.uk.

B

Data Availability

The dataset produced during the Power to the People citizen science project is made publicly available online for research use on Mendeley Data.

It can be accessed at: `https://www.doi.org/10.17632/xw6gr8p2cn.1`

C

Demand Estimation Parameters

The following information was used to set the demand estimation parameters listed in Table 5.7 which were not specified by the MICS data results or other data sources detailed in Chapter 5. Expert opinions were sourced from a Sierra Leonean energy systems expert with significant experience in the power sector.

Lights: Five indoor lights and four outdoor lights were included in each household. The expert consulted indicated that one outdoor light was required per corner of the house for security and safety from animals such as snakes and reptiles. Indoor lights are assumed to be used for two hours (with 20% variability) each night between sunset to midnight (with 35% variability) and left on for a minimum of 10 minutes after switch-on. Outdoor lights are assumed to be used for nine hours (with 20% variability) each night between sunset to sunrise (with 20% variability) and kept on for a minimum of 30 minutes after switch-on. All outdoor lights in a household are modelled as turning on and off simultaneously.

Mobile telephone: This refers to mobile phone charging, not mobile phone usage once charged. Phones are assumed to be charged each night. Based on expert opinion, phones are typically plugged in around 10 pm (i.e. before bed) and not earlier, as agricultural workers (i.e. the dominant industry in rural Sierra Leone) will use their phone in the evening upon returning from the fields and charge as they sleep. It is assumed that phones will charge for two hours (with 20% variability) between 10 PM and 2 AM (with 35% variability) and are charged for a minimum of 20 minutes after being connected.

Radios, televisions, and computers: These three appliances are modelled similarly. While radios can be plug-in, rechargeable, or powered with dry-cell batteries, the expert consulted indicated that radios used during working hours (e.g. by farmers) are likely to be powered by dry-cell batteries whereas those in the home are more likely to be plugged in and used in the evening. Dry cell batteries obviously do not affect the demand on a household electrical system; so, it is assumed that all radios are plug-in and used in the evening for entertainment. It can similarly be questioned whether computers are usually used at work (i.e. not contributing to the household demand) or in the evening for entertainment. It is similarly assumed that all computers are used in the evening for entertainment. Finally, televisions may be used by certain businesses in day-time hours

(e.g. to screen football games) – however, again, here these are assumed to be used in the evening for entertainment. These assumptions err on the side of household demand overestimation. Each of these appliances is modelled as used for 90 minutes (with 20% variability) in a window between sunset and midnight (with 35% variability), and left on for a minimum of 30 minutes once turned on. They are modelled as used on 80% of days, to account for the fact that households can have different activities and social engagements in the evenings which can disrupt their usage patterns.

Fans: The expert source consulted indicated that precipitation has a strong impact on fan use in Sierra Leone, with fans used much more in the hotter dry season (i.e. late October through March as shown in Figure C.1) than the cooler wet season. Demand estimation experiments therefore modelled a month in the dry season to err on the side of overestimation. Additionally, it was indicated by the expert that fans are used differently on weekdays and weekends. While on weekdays they are likely to be used in the early evening (i.e. 6-7pm) and overnight, on the weekends they are also likely to be used from 12-2pm, the hottest point of the day, as people tend to be at home at this time on weekends. As such, fans are modelled differently on weekdays and weekends. On weekdays, they are modelled with twelve hours of usage (with 20% variability) between 6PM and the following sunrise (with 20% variability), and used for at least an hour once turned on. On weekends, an additional possible usage window is added between 12pm and 2pm (with 20% variability) and total usage is increased to 14 hours.

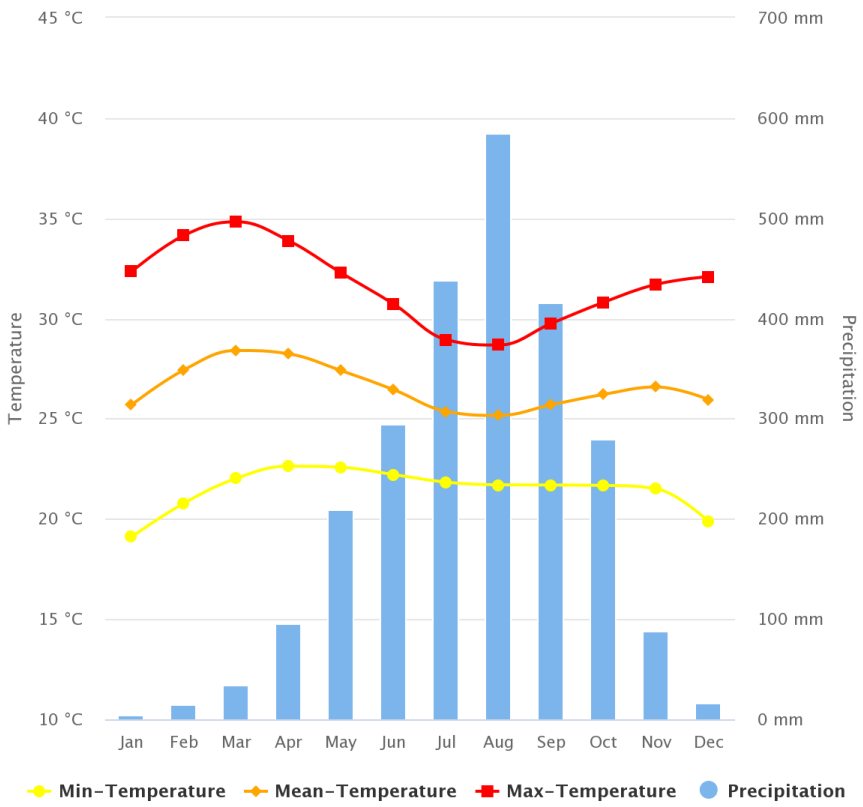


Figure C.1: Monthly climatology of temperature and precipitation in Sierra Leone based on 1991-2020 data [305].

Irons: Irons were modelled as used only on weekends. Based on expert opinion, irons are most likely to be used to press church clothes before Sunday morning service, with the most likely usage window being Saturday evening. While a smaller group of people may iron their clothes each day (e.g. some students ironing their clothes before school), this is likely to be a small minority. Irons were therefore modelled as a weekend-only appliance, and only used 80% of the time in designated weekend windows. They are modelled with a usage time of 30 minutes (with 20% variability) between sunset and midnight (with 35% variability) and used for a minimum of five minutes once turned on. As a thermal appliance, they are modelled with a 30% power variability, to account for differing iron settings.

Refrigerators/freezers: These appliances were assigned two duty cycles based on daily temperature fluctuation [54]. A light duty cycle was modelled during cooler night-time hours and a heavier duty cycle was modelled during warmer day-time hours. Parameters for these duty cycles are shown in Table 5.6. These devices are modelled with 24-hour usage, every day, with no variability.

D

HOMER Simulation Parameters

The following information was used to configure the HOMER simulations and optimisations completed in Chapter 6.

While specific infrastructure models can be modelled in HOMER, generic models are used in the absence of country-specific data on which technologies might be available. The costs of these generic technology models are altered based on the most up-to-date data from IRENA [298].

A discount rate of 8% and inflation rate of 2% are assumed in cost calculations, as well as a project lifetime of 25 years. However, different lifetimes are also assumed for individual components (e.g. batteries and inverter at 15 years, PV at 25 years). Operation and maintenance expenses are assumed at $10/year/kWh for storage and $10/year/kW for generation. Where unspecified, other parameters are left at HOMER defaults.

References

[1] World Bank. *Access to electricity (% of population)*. data retrieved from World Development Indicators, https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS. 2021.

[2] World Bank. *Population, total*. data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SP.POP.TOTL. 2021.

[3] International Energy Agency. *SDG7: Data and Projections*. Tech. rep. Paris, France: IEA, 2020. URL: https://www.iea.org/reports/sdg7-data-and-projections.

[4] World Bank. *State of Electricity Access Report 2017*. Tech. rep. Washington DC: World Bank Group, 2017. URL: http://documents.worldbank.org/curated/en/364571494517675149/full-report.

[5] International Renewable Energy Agency. *Tracking SDG 7: The Energy Progress Report (2021)*. Tech. rep. Washington DC: World Bank, 2021. URL: https://www.irena.org/publications/2021/Jun/Tracking-SDG-7-2021.

[6] Ipsita Das, Pamela Jagger, and Karin Yeatts. "Biomass Cooking Fuels and Health Outcomes for Women in Malawi". In: *EcoHealth* 14.1 (2017), pp. 7–19.

[7] J Parikh, K Smith, and V Laxmi. "Indoor air pollution: A reflection on gender bias". In: *Economic and Political Weekly* 34.9 (1999), pp. 539–544.

[8] Nicholas L. Lam et al. "Kerosene: A review of household uses and their hazards in low-and middle-income countries". In: *Journal of Toxicology and Environmental Health - Part B: Critical Reviews* 15.6 (2012), pp. 396–432.

[9] Makoto Kanagawa and Toshihiko Nakata. "Assessment of access to electricity and the socio-economic impacts in rural areas of developing countries". In: *Energy Policy* 36.6 (2008), pp. 2016–2029.

[10] S. S. Chandel, M. Nagaraju Naik, and Rahul Chandel. "Review of solar photovoltaic water pumping system technology for irrigation and community drinking water supplies". In: *Renewable and Sustainable Energy Reviews* 49 (2015), pp. 1084–1099.

[11] Jem Porcaro et al. *Modern Energy Access and Health*. Tech. rep. Washington DC: World Bank, 2017. URL: http://documents.worldbank.org/curated/en/756131494939083421/pdf/BRI-P148200-PUBLIC-FINALSEARSFHealthweb.pdf.

[12] Ines Havet. "Linking women and energy at the local level to global goals and targets". In: *Energy for Sustainable Development* 7.3 (2003), pp. 75–79.

[13] William Jack and Tavneet Suri. "Mobile Money: The Economics of M-PESA". In: *National Bureau of Economic Research Working Papers* 16721 (2011). URL: https://www.nber.org/papers/w16721.

[14] Giacomo Falchetta et al. "Data from: A High-Resolution Gridded Dataset to Assess Electrification in Sub-Saharan Africa". In: *Mendeley Data* 4 (2019). Accessed through Energy Access Explorer, 2021-09-17. www.energyaccessexplorer.org. URL: http://dx.doi.org/10.17632/kn4636mtvg.4.

[15] Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. *High Resolution Settlement Layer (HRSL)*. 2016. URL: https://www.ciesin.columbia.edu/data/hrsl/.

[16] AJ Tatem et al. *Pilot high resolution poverty maps, University of Southampton/Oxford*. Available at https://www.worldpop.org/geodata/summary?id=1262. Accessed through Energy Access Explorer, 2021-09-27. www.energyaccessexplorer.org. 2013.

[17] James P Dorian, Herman T Franssen, and Dale R Simbeck. "Global challenges in energy". In: *Energy policy* 34.15 (2006), pp. 1984–1991.

[18] Intergovernmental Panel on Climate Change. *Global Warming of 1.5°C. An IPCC Special Report*. Tech. rep. Intergovernmental Panel on Climate Change (IPCC), 2018. URL: https://www.ipcc.ch/sr15/.

[19] Navin Singh Khadka. "Climate change: Low-income countries 'can't keep up' with impacts". In: *BBC News* (Aug. 8, 2021). URL: https://www.bbc.co.uk/news/world-58080083.

[20] Stephane Hallegatte, Jun Rentschler, and Julie Rozenberg. *LIFELINES: The resilient infrastructure opportunity*. World Bank Publications, 2019.

[21] Lauren Sommer. "Developing nations say they're owed for climate damage. Richer nations aren't budging". In: *National Public Radio (NPR)* (Nov. 11, 2021). URL: https://www.npr.org/2021/11/11/1054809644/climate-change-cop26-loss-and-damage?t=1647526903784.

[22] Adriaan Zomers. "Remote access: Context, challenges, and obstacles in rural electrification". In: *IEEE Power and Energy Magazine* 12.4 (2014), pp. 26–34.

[23] Abeeku Brew-Hammond. "Energy access in Africa: Challenges ahead". In: *Energy policy* 38.5 (2010), pp. 2291–2301.

[24] Harald Winkler et al. "Access and Affordability of Electricity in Developing Countries". In: *World Development* 39.6 (2011). Microfinance: Its Impact, Outreach, and Sustainability, pp. 1037–1050.

[25] James D Fearon. "Ethnic and cultural diversity by country". In: *Journal of economic growth* 8.2 (2003), pp. 195–222.

[26] Energy Sector Management Assistance Program (ESMAP). *Mini Grids for Half a Billion People: Market Outlook and Handbook for Decision Makers - Executive Summary*. Tech. rep. Washington DC: World Bank, 2019. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/31926/Mini-Grids-for-Half-a-Billion-People-Market-Outlook-and-Handbook-for-Decision-Makers-Executive-Summary.pdf.

[27] Subhes C. Bhattacharyya. "Energy access programmes and sustainable development: A critical review and analysis". In: *Energy for Sustainable Development* 16.3 (2012), pp. 260–271.

[28] Raluca Golumbeanu and Douglas Barnes. “Connection Charges and Electricity Access in Sub-Saharan Africa -”. In: *Policy Research Working Paper* WPS6511 (2013). URL: http://econ.worldbank.org.

[29] Kenneth Lee et al. “Electrification for "under Grid" households in Rural Kenya”. In: *Development Engineering* 1 (2016), pp. 26–35.

[30] Nicolas Maennling and Perrine Tolenado. *The Renewable Power of the Mine: Accelerating Renewable Energy Integration*. Tech. rep. December. Columbia Center on Sustainable Investment, 2018. URL: https://ccsi.columbia.edu/content/renewable-power-mine-0.

[31] Subhes C. Bhattacharyya. “Mini-grids for the base of the pyramid market: A critical review”. In: *Energies* 11.4 (2018).

[32] Daniel Burmester et al. “A review of nanogrid topologies and technologies”. In: *Renewable and Sustainable Energy Reviews* 67 (2017), pp. 760–775.

[33] M. Rezwan Khan and Edward D. Brown. “DC nanogrids: A low cost PV based solution for livelihood enhancement for rural Bangladesh”. In: *Proceedings of 2014 3rd International Conference on the Developments in Renewable Energy Technology, ICDRET 2014* (2014), pp. 1–5.

[34] Adam Hirsch, Yael Parag, and Josep Guerrero. “Microgrids: A review of technologies, key drivers, and outstanding issues”. In: *Renewable and Sustainable Energy Reviews* 90.September 2017 (2018), pp. 402–411.

[35] Liang Che et al. “Optimal Interconnection Planning of Community Microgrids With Renewable Energy Sources”. In: *IEEE Transactions on Smart Grid* 8.3 (2017), pp. 1054–1063.

[36] Njeri Wamukonya. “Solar home system electrification as a viable technology option for Africa’s development”. In: *Energy Policy* 35.1 (2007), pp. 6–14.

[37] Njeri Wamukonya and Mark Davis. “Socio-economic impacts of rural electrification in Namibia: comparisons between grid, solar and unelectrified households”. In: *Energy for Sustainable Development* 5.3 (2001), pp. 5–13.

[38] Chukwuma Leonard Azimoh et al. “Illuminated but not electrified: An assessment of the impact of Solar Home System on rural households in South Africa”. In: *Applied Energy* 155 (2015), pp. 354–364.

[39] Jonathan Bowes, Campbell Booth, and Scott Strachan. “System interconnection as a path to bottom up electrification”. In: *2017 52nd International Universities Power Engineering Conference (UPEC)*. 2017, pp. 1–5.

[40] Bartosz Soltowski et al. “A Simulation-Based Evaluation of the Benefits and Barriers to Interconnected Solar Home Systems in East Africa”. In: *2018 IEEE PES/IAS PowerAfrica*. 2018, pp. 491–496.

[41] Abdelhay A Sallam and Om P Malik. *Electric distribution systems*. John Wiley & Sons, 2018.

[42] Shafiqa Keddar et al. “An Overview of the Technical Challenges Facing the Deployment of Electric Cooking on Hybrid PV/Diesel Mini-Grid in Rural Tanzania—A Case Study Simulation”. In: *Energies* 14.13 (2021).

[43] Liang Che et al. “Optimal Planning of Loop-Based Microgrid Topology”. In: *IEEE Transactions on Smart Grid* 8.4 (2017), pp. 1771–1781.

[44] N.W.A. Lidula and A.D. Rajapakse. "Microgrids research: A review of experimental microgrids and test systems". In: *Renewable and Sustainable Energy Reviews* 15.1 (2011), pp. 186–202.

[45] Lixiao Zhang et al. "Spatial Variation and Distribution of Urban Energy Consumptions from Cities in China". In: *Energies* 4.1 (2011), pp. 26–38.

[46] Nicholas Nixon Opiyo. "How basic access to electricity stimulates temporally increasing load demands by households in rural developing communities". In: *Energy for Sustainable Development* 59 (2020), pp. 97–106.

[47] Elizabeth Shove and Gordon Walker. "What Is Energy For? Social Practice and Energy Demand". In: *Theory, Culture & Society* 31.5 (2014), pp. 41–58.

[48] Allison Hui and Gordon Walker. "Concepts and methodologies for a new relational geography of energy demand: Social practices, doing-places and settings". In: *Energy Research and Social Science* 36.August 2017 (2018), pp. 21–29.

[49] EED Advisory Kenya & World Resources Institute. *Population Census Data Report: Lighting, Cooking & Household Assets Data*. Accessed Through Energy Access Explorer, 28th June 2021. `www.energyaccessexplorer.org`. 2020.

[50] Kenya National Bureau of Statistics (KNBS). *Kenya Population and Housing Census Volume IV: Distribution of Population by Socio-Economic Characteristics*. Accessed Through Energy Access Explorer, 28th June 2021. `www.energyaccessexplorer.org`. 2019.

[51] Copernicus Global Land Service. *Land Cover*. 2021. URL: `https://land.copernicus.eu/global/products/lc`.

[52] OpenStreetMap contributors. *OpenStreetMap*. 2021. URL: `https://www.openstreetmap.org/`.

[53] Russ Garrett and OpenStreetMap contributors. *Open Infrastructure Map*. 2021. URL: `https://openinframap.org/`.

[54] Stefan Pfenninger and Iain Staffell. *Renewables.Ninja*. 2021. URL: `https://www.renewables.ninja/`.

[55] The World Bank Group, Solargis, and Energy Sector Management Assistance Program (ESMAP). *Global Solar Atlas*. 2021. URL: `https://globalsolaratlas.info`.

[56] Technical University of Denmark (DTU) et al. *Global Wind Atlas*. 2021. URL: `https://globalwindatlas.info`.

[57] United Nations. *Principles and Recommendations for Population and Housing Censuses, Revision 3*. Tech. rep. New York, 2017. URL: `https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf`.

[58] Institut national de la statistique et Secretariat national du recensement, Republique du Zaire. *Combiens sommes nous: Résultats provisoires du Recensement Scientifique de la Population*. Tech. rep. Kinshasa, 1984. URL: `https://ireda.ceped.org/inventaire/ressources/cod-1984-rec-o1_resultats_provisoires.pdf`.

[59] Emmanuel N. Amadife and James W. Warhola. "Africa's Political Boundaries: Colonial Cartography, the OAU, and the Advisability of Ethno-National Adjustment". In: *International Journal of Politics, Culture, and Society* 6.4 (1993), pp. 533–554.

[60] Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4)*. Palisades, 2016. URL: http://dx.doi.org/10.7927/H4SF2T42.

[61] *WorldPop, open data for spatial demography*. 2017. URL: https://doi.org/10.1038/sdata.2017.4.

[62] Google Earth, CNES/Airbus, Maxar Technologies. *Mapoton, Katiri County, Sierra Leone: 9°32'47.823", -12°10'9.611"*. 2022. URL: https://goo.gl/maps/3L6LnW2GK5k3pib39.

[63] Facebook. *High Resolution Population Density Maps*. 2018. URL: https://data.humdata.org/dataset/highresolutionpopulationdensitymaps.

[64] Humanitarian OpenStreetMap Team. *Humanitarian OpenStreetMap Team (HOT)*. URL: https://www.hotosm.org/.

[65] Geofabrik GmbH. *Download OpenStreetMap data for this region: Kenya*. http://download.geofabrik.de/africa/kenya.html. 2018.

[66] Google Earth, CNES/Airbus, Maxar Technologies. *Baringo County Kenya: 0°34'01"N 35°47'37"E*. 2022. URL: https://goo.gl/maps/pVJBq4rXydNwUAcK8.

[67] Nuno Limao and Anthony J Venables. "Infrastructure, geographical disadvantage, transport costs, and trade". In: *The world bank economic review* 15.3 (2001), pp. 451–479.

[68] Gavin Capps. "Tribal-landed property: The value of the chieftaincy in contemporary Africa". In: *Journal of Agrarian Change* 16.3 (2016), pp. 452–477.

[69] Stig Enemark et al. *Fit-for-purpose land administration*. International Federation of Surveyors and World Bank, 2014. URL: https://fig.net/resources/publications/figpub/pub60/figpub60.asp.

[70] Gerald K. Moore. "What is a picture worth? a history of remote sensing". In: *Hydrological Sciences Bulletin* 24.4 (1979), pp. 477–485.

[71] United States Office of the Historian. *Sputnik, 1957*. Accessed: 2022-03-17. https://history.state.gov/milestones/1953-1960/sputnik. 2017.

[72] Laura Rocchio, Michael Taylor, and Jeffrey Masek. *History: Landsat Science*. 2019. URL: https://landsat.gsfc.nasa.gov/about/history/.

[73] Bradley C Reed et al. "Measuring phenological variability from satellite imagery". In: *Journal of Vegetation Science* 5.5 (1994), pp. 703–714.

[74] Vernon Dvorak. "Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery". In: *Monthly Weather Review* 103 (1975), pp. 420–430.

[75] David J Rogers et al. "Satellite imagery in the study and forecast of Malaria". In: *Nature* 415.6872 (2002), pp. 710–715.

[76] Richard P. Stumpf, Kristine Holderied, and Mark Sinclair. "Determination of water depth with high-resolution satellite imagery over variable bottom types". In: *Limnology and Oceanography* 48.1part2 (2003), pp. 547–556.

[77] Neal Jean et al. “Combining satellite imagery and machine learning to predict poverty”. In: *Science* 353.6301 (2016), pp. 790–794.

[78] The United States Geological Survey. *EarthExplorer*. 2021. URL: https://earthexplorer.usgs.gov/.

[79] European Space Agency. *Sentinel Data via Copernicus Access Hub*. 2021. URL: https://scihub.copernicus.eu/dhus/#/home.

[80] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. “SpaceNet: A Remote Sensing Dataset and Challenge Series”. In: *CoRR* abs/1807.01232 (2018). arXiv: 1807.01232. URL: http://arxiv.org/abs/1807.01232.

[81] Maxar Technologies. *Open Data Program (Webpage)*. 2021. URL: https://www.maxar.com/open-data.

[82] Humanitarian OpenStreetMap Team and Open Imagery Network contributors. *OpenAerialMap*. 2021. URL: https://openaerialmap.org/.

[83] Glenn Vancauwenberghe and Bastiaan van Loenen. “Exploring the Emergence of Open Spatial Data Infrastructures: Analysis of Recent Developments and Trends in Europe”. In: *User Centric E-Government: Challenges and Opportunities*. Ed. by Saqib Saeed, T. Ramayah, and Zaigham Mahmood. Cham: Springer International Publishing, 2018, pp. 23–45.

[84] National Geographic. *Resource Library Encyclopedic Entry: Citizen Science*. https://www.nationalgeographic.org/encyclopedia/citizen-science/. 2021.

[85] European Citizen Science Association (ECSA). *Ten principles of citizen science*. Berlin, 2015. URL: http://doi.org/10.17605/OSF.IO/XPR2N.

[86] Janis L. Dickinson, Benjamin Zuckerberg, and David N. Bonter. “Citizen science as an ecological research tool: Challenges and benefits”. In: *Annual Review of Ecology, Evolution, and Systematics* 41 (2010), pp. 149–172.

[87] Raffael Heiss and Jörg Matthes. “Citizen science in the social sciences: A call for more evidence”. In: *GAIA-Ecological Perspectives for Science and Society* 26.1 (2017), pp. 22–26.

[88] Behzad Esmaeilian et al. “Use of Citizen Science to Improve Student Experience in Engineering Design, Manufacturing and Sustainability Education”. In: *Procedia Manufacturing* 26 (2018). 46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA, pp. 1361–1368.

[89] Grant Van Horn et al. “The iNaturalist Species Classification and Detection Dataset”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), pp. 8769–8778. arXiv: 1707.06642.

[90] Joseph S. Wilson et al. “More eyes on the prize: an observation of a very rare, threatened species of Philippine Bumble bee, Bombus irisanensis, on iNaturalist and the importance of citizen science in conservation biology”. In: *Journal of Insect Conservation* 24.4 (2020), pp. 727–729.

[91] Maria Aristeidou et al. “Exploring the participation of young citizen scientists in scientific research: The case of iNaturalist”. In: *PLoS ONE* 16.1 January (2021), pp. 1–13.

[92] Benjamin Herfort, Melanie Eckle, and João Porto De Albuquerque. "Being specific about geographic information crowdsourcing: A typology and analysis of the Missing Maps project in South Kivu". In: *Proceedings of the International ISCRAM Conference* May 2016 (2016).

[93] Michal Givoni. "Between micro mappers and missing maps: Digital humanitarianism and the politics of material participation in disaster response". In: *Environment and Planning D: Society and Space* 34.6 (2016), pp. 1025–1043.

[94] Stefan Scholz et al. "Volunteered geographic information for disaster risk reduction-the missing maps approach and its potential within the Red Cross and Red Crescent movement". In: *Remote Sensing* 10.8 (2018).

[95] Jacques Charmes. "A Review of Empirical Evidence on Time Use in Africa from UN-Sponsored Surveys". In: *Gender, time use, and poverty in Sub-Saharan Africa*. Ed. by C. Mark Blackden and Quentin Wodon. World Bank Working Paper No. 73. Washington, D.C.: The International Bank for Reconstruction and Development/The World Bank, 2006. Chap. 3. URL: https://doi.org/10.1596/978-0-8213-6561-8.

[96] Aslihan Kes and Hema Swaminathan. "Gender and Time Poverty in Sub-Saharan Africa". In: *Gender, time use, and poverty in Sub-Saharan Africa*. Ed. by C. Mark Blackden and Quentin Wodon. World Bank Working Paper No. 73. Washington, D.C.: The International Bank for Reconstruction and Development/The World Bank, 2006. Chap. 2. URL: https://doi.org/10.1596/978-0-8213-6561-8.

[97] UNESCO. *UNESCO Institute for Statistics Adult Literacy Rate*. World Bank Open Data. 2018. URL: https://data.worldbank.org/indicator/SE.ADT.LITR.ZS.

[98] UNDESA. *United Nations Department of Economic and Social Affairs, Population Division: World Urbanization Prospects*. World Bank Open Data. 2018. URL: https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.

[99] ITU. *International Telecommunication Union World Telecommunication/ICT Indicators Database*. World Bank Open Data. 2019. URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS.

[100] Stephanie Hirmer et al. "Building Representative Corpora from Illiterate Communities: A Reviewof Challenges and Mitigation Strategies for Developing Countries". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2176–2189.

[101] James Hendler. "Avoiding another AI winter". In: *IEEE Intelligent Systems* 23.02 (2008), pp. 2–4.

[102] Zhongming Zhao Yanfeng Wei. "Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection". In: *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*. Vol. 3. IEEE, 2004, pp. 2008–2010.

[103] Beril Sirmacek and Cem Unsalan. "Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory". In: *IEEE Transactions on Geoscience and Remote Sensing* 47.4 (2009), pp. 1156–1167.

[104] Salman Ahmadi et al. “Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours”. In: *International Journal of Applied Earth Observation and Geoinformation* 12.3 (2010), pp. 150–157.

[105] Jing Peng, Dong Zhang, and Yuncai Liu. “An improved snake model for building detection from urban aerial images”. In: *Pattern Recognition Letters* 26.5 (2005), pp. 587–595.

[106] Sebastien Lefevre, Jonathan Weber, and David Sheeren. “Automatic Building Extraction in VHR Images Using Advanced Morphological Operators”. In: *2007 Urban Remote Sensing Joint Event*. 2007, pp. 1–5.

[107] Li Sun, Yuqi Tang, and Liangpei Zhang. “Rural Building Detection in High-Resolution Imagery Based on a Two-Stage CNN Model”. In: *IEEE Geoscience and Remote Sensing Letters* 14.11 (2017), pp. 1998–2002.

[108] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin. “Artificial Neural Networks: A Tutorial”. In: *Computer* 29.3 (1996), pp. 31–44.

[109] Jürgen Schmidhuber. “Deep Learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117.

[110] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[111] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[112] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. San Diego: Computational and Biological Learning Society, 2015, pp. 1–14. arXiv: `1409.1556`.

[113] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1–9.

[114] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[115] Gao Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: `1608.06993 [cs.CV]`.

[116] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Columbus: IEEE, 2014, pp. 580–587.

[117] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015, pp. 1440–1448.

[118] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149.

[119] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016, pp. 779–788.

[120] Adam Van Etten. “You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery”. In: *CoRR* abs/1805.09512 (2018). arXiv: 1805.09512. URL: http://arxiv.org/abs/1805.09512.

[121] T. Nathan Mundhenk et al. “A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 785–800.

[122] Darius Lam et al. “xView: Objects in Context in Overhead Imagery”. In: *CoRR* abs/1802.07856 (2018). arXiv: 1802.07856. URL: http://arxiv.org/abs/1802.07856.

[123] Jacob Shermeyer and Adam Van Etten. “The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

[124] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *CoRR* abs/1411.4038 (2014). arXiv: 1411.4038. URL: http://arxiv.org/abs/1411.4038.

[125] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[126] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *CoRR* abs/1606.00915 (2016). arXiv: 1606.00915. URL: http://arxiv.org/abs/1606.00915.

[127] Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki. “Multiple Object Extraction from Aerial Imagery with Convolutional Neural Networks”. In: *Journal of Imaging Science and Technology* 60.1 (2016), pp. 104021–104029.

[128] Volodymyr Mnih. “Machine Learning for Aerial Image Labeling”. PhD thesis. University of Toronto, 2013.

[129] Jiangye Yuan. “Learning Building Extraction in Aerial Scenes with Convolutional Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.11 (2018), pp. 2793–2798.

[130] Emmanuel Maggiori et al. “Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark”. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017.

[131] Yaning Yi et al. “Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network”. In: *Remote Sensing* 11.15 (2019).

[132] Yuchu Qin et al. “Semantic Segmentation of Building Roof in Dense Urban Environment with Deep Convolutional Neural Network: A Case Study Using GF2 VHR Imagery in China”. In: *Sensors* 19.5 (2019).

[133] Picterra. *Geospatial intelligence for enterprise: Picterra (Webpage)*. 2021. URL: https://picterra.ch/.

[134] Petuum. *Petuum: AI for All (Webpage)*. 2021. URL: https://petuum.com/.

[135] Mapwith.ai. *Mapwith.ai: Using AI to map the world (Webpage)*. 2021. URL: https://mapwith.ai/.

[136] Nicholas Weir et al. "SpaceNet MVOI: A Multi-View Overhead Imagery Dataset". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

[137] Jacob Shermeyer et al. "SpaceNet 6: Multi-Sensor All Weather Mapping Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020.

[138] Adam Van Etten et al. "The Multi-Temporal Urban Development SpaceNet Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 6398–6407.

[139] Ilke 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.

[140] Sharada Prasanna Mohanty et al. "Deep Learning for Understanding Satellite Imagery: An Experimental Survey". In: *Frontiers in Artificial Intelligence* 3 (2020).

[141] Adrian Boguszewski et al. "LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2021, pp. 1102–1110.

[142] Seyed Majid Azimi et al. "SkyScapes - Fine-Grained Semantic Understanding of Aerial Scenes". In: *CoRR* abs/2007.06102 (2020). arXiv: 2007.06102. URL: https://arxiv.org/abs/2007.06102.

[143] Global Facility for Disaster Reduction and Recovery (GFDRR). *Open Cities AI Challenge Dataset*. https://doi.org/10.34911/rdnt.f94cxb. 2020.

[144] Evan Mills. *The specter of fuel-based lighting*. 2005.

[145] Evan Mills. "Identifying and reducing the health and safety impacts of fuel-based lighting". In: *Energy for Sustainable Development* 30 (2016), pp. 39–50.

[146] Evelyn L. Rhodes et al. "Behavioral Attitudes and Preferences in Cooking Practices with Traditional Open-Fire Stoves in Peru, Nepal, and Kenya: Implications for Improved Cookstove Interventions". In: *International Journal of Environmental Research and Public Health* 11.10 (2014), pp. 10310–10326.

[147] Radhika Khosla et al. "Cooling for sustainable development". In: *Nature Sustainability* 4.3 (2021), pp. 201–208.

[148] Giacomo Falchetta and Malcolm N. Mistry. "The role of residential air circulation and cooling demand for electrification planning: Implications of climate change in sub-Saharan Africa". In: *Energy Economics* 99 (2021), p. 105307.

[149] Labelle Prussin. "An Introduction to Indigenous African Architecture". In: *Journal of the Society of Architectural Historians* 33.3 (1974), pp. 183–205.

[150] Narasimha D. Rao and Kevin Ummel. "White goods for white people? Drivers of electric appliance growth in emerging economies". In: *Energy Research and Social Science* 27 (2017), pp. 106–116.

[151] Miguel Poblete-Cazenave and Shonali Pachauri. "A model of energy poverty and access: Estimating household electricity demand and appliance ownership". In: *Energy Economics* 98 (2021), p. 105266.

[152] Efficiency for Access Coalition. *2021 Appliance Data Trends*. Tech. rep. January. Efficiency for Access, 2021.

[153] Lukas G. Swan and V. Ismet Ugursal. "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques". In: *Renewable and Sustainable Energy Reviews* 13.8 (2009), pp. 1819–1835.

[154] Asligul Serasu Duran and Feyza G. Sahinyazan. "An analysis of renewable mini-grid projects for rural electrification". In: *Socio-Economic Planning Sciences* 77 (2021), p. 100999.

[155] A. Capasso et al. "A bottom-up approach to residential load modeling". In: *IEEE Transactions on Power Systems* 9.2 (1994), pp. 957–964.

[156] Runming Yao and Koen Steemers. "A method of formulating energy load profile for domestic buildings in the UK". In: *Energy and Buildings* 37.6 (2005), pp. 663–671.

[157] Ian Richardson et al. "Domestic electricity use: A high-resolution energy demand model". In: *Energy and Buildings* 42.10 (2010), pp. 1878–1887.

[158] George Tsagarakis, Adam J. Collin, and Aristides E. Kiprakis. "Modelling the electrical loads of UK residential energy users". In: *Proceedings of the Universities Power Engineering Conference* (2012).

[159] Eoghan McKenna and Murray Thomson. "High-resolution stochastic integrated thermal-electrical domestic demand model". In: *Applied Energy* 165 (2016), pp. 445–461.

[160] David Fischer, Andreas Härtl, and Bernhard Wille-Haussmann. "Model for electric load profiles with high time resolution for German households". In: *Energy and Buildings* 92 (2015), pp. 170–179.

[161] Joakim Widén and Ewa Wäckelgård. "A high-resolution stochastic model of domestic activity patterns and electricity demand". In: *Applied Energy* 87.6 (2010), pp. 1880–1892.

[162] Govinda R. Timilsina and Kalim U. Shah. "Filling the gaps: Policy supports and interventions for scaling up renewable energy development in Small Island Developing States". In: *Energy Policy* 98 (2016), pp. 653–662.

[163] Katherine A Collett and Stephanie A Hirmer. "Data needed to decarbonize paratransit in Sub-Saharan Africa". In: *Nature Sustainability* (2021), pp. 1–3.

[164] Arun Jacob. "Mind the Gap: Analyzing the Impact of Data Gap in Millennium Development Goals' (MDGs) Indicators on the Progress toward MDGs". In: *World Development* 93 (2017), pp. 260–278.

[165] Jevgenijs Steinbuks et al. *Forecasting Electricity Demand: An Aid for Practitioners*. 2017. URL: `http://www.worldbank.org/energy/livewire`.

[166] Catherine Wolfram, Orie Shelef, and Paul Gertler. "How will energy demand develop in the developing world?" In: *Journal of Economic Perspectives* 26.1 (2012), pp. 119–138.

[167] Subhes C. Bhattacharyya and Govinda R. Timilsina. "Modelling energy demand of developing countries: Are the specific features adequately captured?" In: *Energy Policy* 38.4 (2010), pp. 1979–1990.

[168] P. Boait, V. Advani, and R. Gammon. "Estimation of demand diversity and daily demand profile for off-grid electrification in developing countries". In: *Energy for Sustainable Development* 29 (2015), pp. 135–141.

[169] Philip Sandwell et al. "Analysis of energy access and impact of modern energy sources in unelectrified villages in Uttar Pradesh". In: *Energy for Sustainable Development* 35 (2016), pp. 67–79.

[170] Stefano Mandelli, Marco Merlo, and Emanuela Colombo. "Novel procedure to formulate load profiles for off-grid rural areas". In: *Energy for Sustainable Development* 31 (2016), pp. 130–142.

[171] S. Mandelli et al. "Novel LoadProGen procedure for micro-grid design in emerging country scenarios: Application to energy storage sizing". In: *Energy Procedia* 135 (2017), pp. 367–378.

[172] Fabio Riva et al. "Modelling long-term electricity load demand for rural electrification planning". In: *2019 IEEE Milan PowerTech, PowerTech 2019* (2019).

[173] Francesco Lombardi et al. "Enabling combined access to electricity and clean cooking with PV-microgrids: new evidences from a high-resolution model of cooking loads". In: *Energy for Sustainable Development* 49 (2019), pp. 78–88.

[174] Francesco Lombardi et al. "Generating high-resolution multi-energy load profiles for remote areas with an open-source stochastic model". In: *Energy* 177 (2019), pp. 433–444.

[175] Fabio Riva et al. "Long-term energy planning and demand forecast in remote areas of developing countries: Classification of case studies and insights from a modelling perspective". In: *Energy Strategy Reviews* 20 (2018), pp. 71–89.

[176] Fabio Riva et al. "Soft-linking energy demand and optimisation models for local long-term electricity planning: An application to rural India". In: *Energy* 166 (2019), pp. 32–46.

[177] Elias Hartvigsson and Erik O. Ahlgren. "Comparison of load profiles in a mini-grid: Assessment of performance metrics using measured and interview-based data". In: *Energy for Sustainable Development* 43 (2018), pp. 186–195.

[178] Courtney Blodgett et al. "Accuracy of energy-use surveys in predicting rural mini-grid user consumption". In: *Energy for Sustainable Development* 41 (2017), pp. 88–105.

[179] Henry Louie and Peter Dauenhauer. "Effects of load estimation error on small-scale off-grid photovoltaic system design, cost and reliability". In: *Energy for Sustainable Development* 34 (2016), pp. 30–43.

[180] Anna Clements et al. "The Service Value Method for Design of Energy Access Systems in the Global South". In: *Proceedings of the IEEE* (2019).

[181] Stephanie Hirmer and Peter Guthrie. "The benefits of energy appliances in the off-grid energy sector based on seven off-grid initiatives in rural Uganda". In: *Renewable and Sustainable Energy Reviews* 79.July 2016 (2017), pp. 924–934.

[182] Douglas Henry Fabini et al. "Mapping Induced Residential Demand for Electricity in Kenya". In: *Proceedings of the Symposium on Computing for Development (ACM DEV 5)*. May 2016. San Jose: ACM, 2014, pp. 43–52.

[183] "Analyzing grid extension and stand-alone photovoltaic systems for the cost-effective electrification of Kenya". In: *Energy for Sustainable Development* 25.2015 (2015), pp. 75–86.

[184] UNICEF. *About MICS*. 2021. URL: `https://mics.unicef.org/about`.

[185] UNICEF. *MICS 6 Tools*. 2021. URL: `https://mics.unicef.org/tools`.

[186] UNICEF. *Multiple Indicator Cluster Surveys: MICS6 Household Questionnaire*. 2018. URL: `http://mics.unicef.org/tools#survey-design`.

[187] The DHS Program. *Demographic Health Surveys: Model Household Questionnaire*. 2017.

[188] World Bank Group. *Living Standards Measurement Survey 2012 - Household Questionnaire (Part II)*. URL: `https://microdata.worldbank.org/index.php/catalog/1970/related-materials`.

[189] Mikul Bhatia and Niki Angelou. *Beyond Connections: Energy Access Redefined*. Tech. rep. 008/15. ESMAP Technical Report. Washington, DC: Energy Sector Management Assistance Program (ESMAP), 2015. URL: `https://openknowledge.worldbank.org/handle/10986/24368`.

[190] Christopher M. Raymond et al. "Integrating local and scientific knowledge for environmental management". In: *Journal of Environmental Management* 91.8 (2010), pp. 1766–1777.

[191] Fikret Berkes, Johan Colding, and Carl Folke. "Rediscovery of Traditional Ecological Knowledge as Adaptive Management". In: *Ecological Applications* 10.5 (2000), pp. 1251–1262.

[192] P. Palaiologou et al. "Wind characteristics and mapping for power production in the Island of Lesvos, Greece". In: *Computers and Geosciences* 37.7 (2011), pp. 962–972.

[193] Stefano Grassi, Ndaona Chokani, and Reza S. Abhari. "Large scale technical and economical assessment of wind energy potential with a GIS tool: Case study Iowa". In: *Energy Policy* 45 (2012), pp. 73–85.

[194] Shahid Hussain Siyal et al. "Wind energy assessment considering geographic and environmental restrictions in Sweden: A GIS-based approach". In: *Energy* 83.2015 (2015), pp. 447–461.

[195] Juan M. Sánchez-Lozano et al. "GIS-based photovoltaic solar farms site selection using ELECTRE-TRI: Evaluating the case for Torre Pacheco, Murcia, Southeast of Spain". In: *Renewable Energy* 66 (2014), pp. 478–494. URL: `https://doi.org/10.1016/j.renene.2013.12.038`.

[196] Adel Gastli and Yassine Charabi. "Solar electricity prospects in Oman using GIS-based solar radiation maps". In: *Renewable and Sustainable Energy Reviews* 14.2 (2010), pp. 790–797.

[197] Mevlut Uyan. "GIS-based solar farms site selection using analytic hierarchy process (AHP) in Karapinar region Konya/Turkey". In: *Renewable and Sustainable Energy Reviews* 28 (2013), pp. 11–17.

[198] S. Szabó et al. "Sustainable energy planning: Leapfrogging the energy poverty gap in Africa". In: *Renewable and Sustainable Energy Reviews* 28 (2013), pp. 500–509.

[199] Dimitrios Mentis et al. "Assessing the technical wind energy potential in africa a GIS-based approach". In: *Renewable Energy* 83 (2015), pp. 110–125.

[200] Christopher N.H. Doll and Shonali Pachauri. "Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery". In: *Energy Policy* 38.10 (2010), pp. 5661–5670.

[201] Fabian Huneke et al. "Optimisation of hybrid off-grid energy systems by linear programming". In: *Energy, Sustainability and Society* 2.1 (2012), pp. 1–19.

[202] YA Katsigiannis, PS Georgilakis, and ES Karapidakis. "Multiobjective genetic algorithm solution to the optimum economic and environmental performance problem of small autonomous hybrid power systems with renewables". In: *IET Renewable Power Generation* 4.5 (2010), pp. 404–419.

[203] BK Bala and Saiful Azam Siddique. "Optimal design of a PV-diesel hybrid system for electrification of an isolated island—Sandwip in Bangladesh using genetic algorithm". In: *Energy for Sustainable Development* 13.3 (2009), pp. 137–142.

[204] SM Moghaddas-Tafreshi, HA Zamani, and SM Hakimi. "Optimal sizing of distributed resources in micro grid with loss of power supply probability technology by using breeding particle swarm optimization". In: *Journal of Renewable and Sustainable Energy* 3.4 (2011), p. 043105.

[205] Lingfeng Wang and Chanan Singh. "Multicriteria Design of Hybrid Power Generation Systems Based on a Modified Particle Swarm Optimization Algorithm". In: *IEEE Transactions on Energy Conversion* 24.1 (2009), pp. 163–172.

[206] Yiannis A. Katsigiannis, Pavlos S. Georgilakis, and Emmanuel S. Karapidakis. "Hybrid Simulated Annealing–Tabu Search Method for Optimal Sizing of Autonomous Power Systems With Renewables". In: *IEEE Transactions on Sustainable Energy* 3.3 (2012), pp. 330–338.

[207] S. Jonnavithula and R. Billinton. "Minimum cost analysis of feeder routing in distribution system planning". In: *IEEE Transactions on Power Delivery* 11.4 (1996), pp. 1935–1940.

[208] N.G. Boulaxis and M.P. Papadopoulos. "Optimal feeder routing in distribution system planning using dynamic programming technique and GIS facilities". In: *IEEE Transactions on Power Delivery* 17.1 (2002), pp. 242–247.

[209] P.C. Paiva et al. "Integral planning of primary-secondary distribution systems using mixed integer linear programming". In: *IEEE Transactions on Power Systems* 20.2 (2005), pp. 1134–1143.

[210] N. C. Koutsoukis, P. S. Georgilakis, and N. D. Hatziargyriou. "A Tabu search method for distribution network planning considering distributed generation and uncertainties". In: *2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. 2014, pp. 1–6.

[211] T.W Lambert and D.C Hittle. "Optimization of autonomous village electrification systems by simulated annealing". In: *Solar Energy* 68.1 (2000), pp. 121–132.

[212] UL LLC. *HOMER Software*. 2021. URL: https://www.homerenergy.com/.

[213] Tom Lambert, Paul Gilman, and Peter Lilienthal. "Micropower system modeling with HOMER". In: vol. 1. 1. John Wiley & Sons New York, NY, USA, 2006, pp. 379–385.

[214] Energy Technologies Area, Berkeley Lab. *The Distributed Energy Resources Customer Adoption Model (DER-CAM)*. 2021. URL: https://gridintegration.lbl.gov/der-cam.

[215] Universidad Zaragosa. *IHOGA / MHOGA*. 2021. URL: https://gridintegration.lbl.gov/der-cam.

[216] D. Thevenard, G. Leng, and S. Martel. "The RETScreen model for assessing potential PV projects". In: *Conference Record of the Twenty-Eighth IEEE Photovoltaic Specialists Conference - 2000 (Cat. No.00CH37036)*. 2000, pp. 1626–1629.

[217] Government of Canada. *RETScreen*. 2021. URL: https://www.nrcan.gc.ca/maps-tools-and-publications/tools/modelling-tools/retscreen/7465.

[218] Carlos Mateo Domingo et al. "A Reference Network Model for Large-Scale Distribution Planning With Automatic Street Map Generation". In: *IEEE Transactions on Power Systems* 26.1 (2011), pp. 190–197.

[219] C. Monteiro, J.T. Saraiva, and V. Miranda. "Evaluation of electrification alternatives in developing countries-the SOLARGIS tool". In: *MELECON '98. 9th Mediterranean Electrotechnical Conference. Proceedings (Cat. No.98CH36056)*. Vol. 2. 1998, 1037–1041 vol.2.

[220] J. Amador and J. Domínguez. "Application of geographical information systems to rural electrification with renewable energy sources". In: *Renewable Energy* 30.12 (2005), pp. 1897–1912.

[221] Thomas Huld, Magda Moner-Girona, and Akos Kriston. "Geospatial Analysis of Photovoltaic Mini-Grid System Performance". In: *Energies* 10.2 (2017).

[222] Elizabeth Kaijuka. "GIS and rural electricity planning in Uganda". In: *Journal of Cleaner Production* 15.2 (2007), pp. 203–217.

[223] Paul Bertheau, Catherina Cader, and Philipp Blechinger. "Electrification Modelling for Nigeria". In: *Energy Procedia* 93 (2016). Africa-EU Symposium on Renewable Energy Research and Innovation, pp. 108–112.

[224] Catherina Cader, Philipp Blechinger, and Paul Bertheau. "Electrification Planning with Focus on Hybrid Mini-grids – A Comprehensive Modelling Approach for the Global South". In: *Energy Procedia* 99 (2016). 10th International Renewable Energy Storage Conference, IRES 2016, 15-17 March 2016, Düsseldorf, Germany, pp. 269–276.

[225] Eduardo Alejandro Martinez-Cesena et al. *Using Mobile Phone Data for Electricity Infrastructure Planning*. 2015. arXiv: 1504.03899 [physics.soc-ph].

[226] S Szabó et al. "Energy solutions in rural Africa: mapping electrification costs of distributed solar and diesel generation versus grid extension". In: *Environmental Research Letters* 6.3 (2011), p. 034002.

[227] Yakubu Abdul-Salam and Euan Phimister. "The politico-economics of electricity planning in developing countries: A case study of Ghana". In: *Energy Policy* 88 (2016), pp. 299–309.

[228] Yakubu Abdul-Salam and Euan Phimister. "How effective are heuristic solutions for electricity planning in developing countries". In: *Socio-Economic Planning Sciences* 55 (2016), pp. 14–24.

[229] Quadracci Sustainable Engineering Lab, Columbia University. *Network Planner*. 2017. URL: `https://qsel.columbia.edu/network-planner/`.

[230] Francis Kemausuor et al. "Electrification planning using Network Planner tool: The case of Ghana". In: *Energy for Sustainable Development* 19.1 (2014), pp. 92–101.

[231] Sanusi Ohiare. "Expanding electricity access to all in Nigeria: a spatial planning and cost analysis". In: *Energy, Sustainability and Society* 5.1 (2015), pp. 1–18.

[232] Vijay Modi et al. *Liberia power sector capacity building and energy master planning, Final Report. Phase 4: National Electrification Master Plan*. 2014.

[233] Dimitrios Mentis et al. "Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa". In: *Environmental Research Letters* 12.8 (2017), p. 085003.

[234] Alexandros Korkovelos et al. "The role of open access data in geospatial electrification planning and the achievement of SDG7. An onsset-based case study for Malawi". In: *Energies* 12.7 (2019).

[235] Mounirah Bissiri et al. "A geospatial approach towards defining cost-optimal electrification pathways in West Africa". In: *Energy* 200 (2020).

[236] Moksnes Nandi et al. "Electrification pathways for Kenya–linking spatial electrification analysis and medium to long term energy planning". In: *Environmental Research Letters* 12 (2017), p. 95008.

[237] J. G. Peña Balderrama et al. "Incorporating high-resolution demand and techno-economic optimization to evaluate micro-grids into the Open Source Spatial Electrification Tool (OnSSET)". In: *Energy for Sustainable Development* 56 (2020), pp. 98–118.

[238] Pedro Ciller et al. "Optimal electrification planning incorporating on-and off-grid technologies: the Reference Electrification Model (REM)". In: *Proceedings of the IEEE* 107.9 (2019), pp. 1872–1905.

[239] Douglas Douglas Austin Ellman. "The reference electrification model: a computer model for planning rural electricity access". PhD thesis. Massachusetts Institute of Technology, 2015.

[240] Pedro Ciller Cutillas. "Clustering-related improvements in the Reference Electrification Model". PhD thesis. Master's Thesis, School of Engineering, Universidad Pontificia Comillas ..., 2016.

[241] Vivian Li et al. "The local reference electrification model: Comprehensive decision-making tool for the design of rural microgrids". PhD thesis. Massachusetts Institute of Technology, 2016.

[242] Matthew Daniel Brusnahan. "Minigrids for electrification: Policies to promote industry growth". PhD thesis. Massachusetts Institute of Technology, 2018.

[243] Turner Cotterman. "Enhanced techniques to plan rural electrical networks using the Reference Electrification Model". PhD thesis. Massachusetts Institute of Technology, 2017.

[244] The World Bank Group. *Off-Grid Market Opportunities Tool*. 2016. URL: https://offgrid.energydata.info/.

[245] ECOWAS Observatory for Renewable Energy and Energy Efficiency (ECREEE). *ECOWREX Tool*. 2018. URL: http://www.ecowrex.org/mapView/.

[246] World Resources Institute. *Energy Access Explorer*. www.energyaccessexplorer.org. 2021.

[247] The World Bank Group. *Global Electrification Platform*. 2021. URL: https://electrifynow.energydata.info/.

[248] Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11*. Palisades, NY, 2018. URL: https://doi.org/10.7927/H49C6VHW.

[249] Kwabena Asiama, Rohan Bennett, and Jaap Zevenbergen. "Participatory land administration on customary lands: A practical VGI experiment in Nanton, Ghana". In: *ISPRS International Journal of Geo-information* 6.7 (2017), p. 186.

[250] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.

[251] The Zooniverse Team. *Welcome to the Zooniverse*. 2021. URL: https://www.zooniverse.org/.

[252] Samuel Sigal. "Citizen science is booming during the pandemic". In: *Vox* (2021). URL: https://www.vox.com/future-perfect/22177247/citizen-science-amateur-backyard-birding-astronomy-covid-pandemic.

[253] Northwestern University. "Citizen Science Project Spotlights". In: *SustainNU* (2020). URL: www.northwestern.edu/sustainability/news/2020/citizen-science-project-spotlights.html.

[254] Paul Aparna. "Regeneron ISEF attendees contribute 900 volunteer hours". In: *Society for Science* (2020). URL: https://www.societyforscience.org/blog/regeneron-isef-attendees-contribute-900-volunteer-hours/.

[255] The Zooniverse. *Fulfilling service hour requirements through Zoonvierse*. https://blog.zooniverse.org/2020/03/26/fulfilling-service-hour-requirements-through-zooniverse/. 2020.

[256] D. Bonafilia et al. *Mapping the world to help aid workers, with weakly, semi-supervised learning*. 2019. URL: https://ai.facebook.com/blog/mapping-the-world-to-help-aid-workers-with-weakly-semi-supervised-learning.

[257] GeospatialWorld.net News Desk. *Price breakthrough in high resolution satellite imagery via Soar platform*. 2020. URL: https://www.geospatialworld.net/news/price-break-through-for-new-tasked-high-resolution-satellite-imagery-via-the-soar-platform/.

[258] Salary.com. *Hourly Wage for Researcher I - Academic Salary in the United States*. 2022. URL: https://www.salary.com/research/salary/benchmark/researcher-i-academic-hourly-wages.

[259] Stephen James Lee et al. "Adaptive electricity access planning". PhD thesis. Massachusetts Institute of Technology, School of Engineering, Institute for ..., 2018.

[260] Kenya Ministry of Energy. *Kenya National Electrification Strategy: Key Highlights*. Tech. rep. Nairobi, Kenya: Government of Kenya, 2018, p. 36. URL: http://pubdocs.worldbank.org/en/413001554284496731/Kenya-National-Electrification-Strategy-KNES-Key-Highlights-2018.pdf.

[261] World Bank. *Rural land area (sq. km)*. data retrieved from World Development Indicators and provided by CIESIN, https://data.worldbank.org/indicator/AG.LND.TOTL.RU.K2. 2021.

[262] Yifan Sun et al. "Summarizing CPU and GPU Design Trends with Product Data". In: *CoRR* abs/1911.11313 (2019). arXiv: 1911.11313. URL: http://arxiv.org/abs/1911.11313.

[263] United Nations Office for Outer Space Affairs. *Online Index of Objects Launched into Outer Space*. https://www.unoosa.org/oosa/osoindex/search-ng.jspx. 2022.

[264] Marco Sozzi et al. "Economic Comparison of Satellite, Plane and UAV-Acquired NDVI Images for Site-Specific Nitrogen Application: Observations from Italy". In: *Agronomy* 11.11 (2021).

[265] Alexandra Swanson et al. "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna". In: *Scientific data* 2.1 (2015), pp. 1–14.

[266] Fiona M Jones et al. "Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project". In: *Scientific data* 5.1 (2018), pp. 1–13.

[267] EdjeElectronics. *generate_tfrecord.py*. https://github.com/EdjeElectronics/TensorFlow-Object-Detection-API-Tutorial-Train-Multiple-Objects-Windows-10/blob/master/generate_tfrecord.py. 2020.

[268] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[269] Yiming Shi and Vivek Rathod. *TensorFlow 1 Detection Model Zoo*. 2021. URL: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md.

[270] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

[271] Tensorflow contributors. *faster_rcnn_inception_v2_pets.config*. https://github.com/tensorflow/models/blob/master/research/object_detection/samples/configs/faster_rcnn_inception_v2_pets.config. 2020.

[272] Omkar M Parkhi et al. "Cats and dogs". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 3498–3505.

[273] Common Objects in Context (COCO). *COCO: Detection Evaluation*. 2021. URL: https://cocodataset.org/#detection-eval.

[274] World Bank. *Low & middle income*. https://data.worldbank.org/country/XO. 2021.

[275] Jukka V. Paatero and Peter D. Lund. "A model for generating household electricity load profiles". In: *International Journal of Energy Research* 30.5 (2006), pp. 273–290.

[276] The World Bank Group. *Measuring Energy Access in Multidimensional Way through Household Surveys Multi-tier Energy Access Tracking Framework Global Surveys*. 2020. URL: https://www.worldbank.org/en/results/2020/11/10/measuring-energy-access-in-multidimensional-way-through-household-surveys-multi-tier-energy-access-tracking-framework-global-surveys.

[277] NOAA Earth System Research Laboratories, Global Monitoring Laboratory. *NOAA Solar Calculator*. https://gml.noaa.gov/grad/solcalc/. 2021.

[278] Jeffrey James. "The distributional effects of leapfrogging in mobile phones". In: *Telematics and Informatics* 29.3 (2012), pp. 294–301.

[279] Chris Mullen and Neal Wade. *Appliance Data and Multi-Tier Framework for household electrical load modelling*. Tech. rep. Working Paper 03 Dec 2020. v.1.2. Loughborough, UK: Newcastle University, 2020. URL: https://mecs.org.uk/wp-content/uploads/2020/12/Appliance-Data-and-Multi-Tier-Framework-_working-paper_201203_NOCOMMENTS-002.pdf.

[280] Louis-Benoit Desroches et al. "Computer usage and national energy consumption: Results from a field-metering study". In: (2015).

[281] Smarter Business. *How Much Energy Do My Appliances Use?* https://smarterbusiness.co.uk/blogs/how-much-energy-do-my-appliances-use-infographic/. 2019.

[282] Simone Fobi et al. "A longitudinal study of electricity consumption growth in Kenya". In: *Energy Policy* 123 (2018), pp. 569–578.

[283] Kobina Aidoo and Ryan C. Briggs. "Underpowered: Rolling blackouts in Africa disproportionately hurt the poor". In: *African Studies Review* 62.3 (2019), pp. 112–131.

[284] Statistics Sierra Leone. *Sierra Leone Integrated Household Survey (SLIHS) Report 2018*. Tech. rep. 2019. URL: https://www.statistics.sl/images/StatisticsSL/Documents/SLIHS2018/SLIHS_2018_New/sierra_leone_integrated_household_survey2018_report.pdf.

[285] Statistics Sierra Leone. *Sierra Leone Integrated Household Survey (SLIHS)*. https://www.statistics.sl/index.php/sierra-leone-integrated-household-survey-slihs.html. 2021.

[286] Exchange Rates UK. *US Dollar to Sierra Leone Leone Spot Exchange Rates for 2018*. 2022. URL: https://www.exchangerates.org.uk/USD-SLL-spot-exchange-rates-history-2018.html.

[287] U.S. Energy Information Administration. *Electricity explained: Factors affecting electricity prices*. 2021. URL: https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php.

[288] Sierra Leone Electricity & Water Regulatory Commission. *EDSA Tariff*. 2022. URL: https://ewrc.gov.sl/wp-content/uploads/2020/10/EDSA-Tariff.pdf.

[289] Sierra Leone Electricity & Water Regulatory Commission. *Electricity Minigrid Tariffs*. 2022. URL: https://ewrc.gov.sl/wp-content/uploads/2021/10/ELECTRICITY-MINI-GRIDS-TARIFF-1.pdf.

[290] World Bank. *Doing Business 2020: Economy Profile Sierra Leone*. Tech. rep. Washington, DC: World Bank, 2020. URL: https://www.doingbusiness.org/content/dam/doingBusiness/country/s/sierra-leone/SLE.pdf.

[291] Dongkuan Xu and Yingjie Tian. "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2 (2015), pp. 165–193.

[292] Mihael Ankerst et al. "OPTICS: Ordering points to identify the clustering structure". In: *ACM Sigmod record* 28.2 (1999), pp. 49–60.

[293] Martin Ester et al. "A Density-Based Clustering Algorithms for Discovering Clusters". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Vol. 96. 34. 1996, pp. 226–231.

[294] Mohammad T. Elbatta and Wesam M. Ashour. "A Dynamic Method for Discovering Density Varied Clusters". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6.February 2013 (2012), pp. 1–8.

[295] Ibrahim Sasay. *The National Electricity Act, 2011*. 2012.

[296] Joseph B. Kruskal. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem". In: *Proceedings of the American Mathematical Society* 7.1 (1956), pp. 48–50.

[297] IRENA. *African Renewable Electricity Profiles for Energy Modelling database: Hydropower*. Tech. rep. Data retrieved from IRENA Global Atlas platform: https://globalatlas.irena.org/. Abu Dhabi, 2021.

[298] IRENA. *Renewable Power Generation Costs in 2020*. Tech. rep. Abu Dhabi, 2021.

[299] David Feldman et al. *US solar photovoltaic system and energy storage cost benchmark: Q1 2020*. Tech. rep. National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2021.

[300] National Renewable Energy Laboratory (NREL). *2021 Annual Technology Baseline*. Tech. rep. National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2021. URL: https://atb.nrel.gov/.

[301] Wesley Cole, A Will Frazier, and Chad Augustine. *Cost projections for utility-scale battery storage: 2021 update*. Tech. rep. National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2021.

[302] Ashish Seth. *ACSR Conductor Size Chart*. 2021. URL: https://electricalsells.com/acsr-conductor-size-chart/.

[303] Electrical Engineering Portal. *Overhead vs Underground Residential Distribution Circuits. Which one is better?* 2017. URL: https://electrical-engineering-portal.com/overhead-vs-underground.

[304] YES Energy Solutions. *How much energy do solar panels produce for your home?* 2022. URL: https://www.yesenergysolutions.co.uk/advice/how-much-energy-solar-panels-produce-home.

[305] The World Bank Group and the Climatic Research Unit of the University of East Anglia. *Climate Change Knowledge Portal: Sierra Leone - Climatology*. 2021. URL: https://climateknowledgeportal.worldbank.org/country/sierra-leone/climate-data-historical.