Dominguez, Cristina; Orehounig, Kristina; Carmeliet, Jan

**Article**

# Estimating hourly lighting load profiles of rural households in East Africa applying a data-driven characterization of occupant behavior and lighting devices ownership

Development Engineering

**Provided in Cooperation with:**
Elsevier

*Suggested Citation:* Dominguez, Cristina; Orehounig, Kristina; Carmeliet, Jan (2021) : Estimating hourly lighting load profiles of rural households in East Africa applying a data-driven characterization of occupant behavior and lighting devices ownership, Development Engineering, ISSN 2352-7285, Elsevier, Amsterdam, Vol. 6, pp. 1-21, https://doi.org/10.1016/j.deveng.2021.100073

This Version is available at:
https://hdl.handle.net/10419/299100

Development Engineering 6 (2021) 100073

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

# Estimating hourly lighting load profiles of rural households in East Africa applying a data-driven characterization of occupant behavior and lighting devices ownership

Cristina Dominguez [a,b,*], Kristina Orehounig [a,b], Jan Carmeliet [a]

[a] Chair of Building Physics, Swiss Federal Institute of Technology Zürich (ETHZ), Switzerland
[b] Laboratory for Urban Energy Systems, Swiss Federal Laboratories for Materials Science and Technology, Empa, Duebendorf, Switzerland

ARTICLE INFO

*Keywords:*
Lighting
Energy demand
Electricity consumption
Occupant behavior
Energy access
Rural electrification

ABSTRACT

To design energy access solutions for rural households in developing countries it is important to have an accurate estimation of what their electricity consumption is. Studies reveal that they mainly use electricity to meet their lighting needs, as they cannot afford high power-consuming appliances. However, the scarce data availability and modeling complexity are a challenge to compute correctly the load profiles without collecting data on-site. This paper presents a methodology that computes the hourly lighting load profiles of rural households in East Africa requiring a small amount of publicly available input data. Combining data from household surveys, climate, and satellite imagery, the methodology applies machine learning for determining occupant behavior patterns, and lamps ownership for indoor and outdoor usage. For this, an average prediction accuracy of 80% is reached. After applying lighting requirement functions, load profiles are generated and then validated using measured data from 13 households in Kenya. Results show that the methodology is able to compute the load profiles with an average normalized root mean squared error of 0.7%, which is less compared to existing simulation approaches using on-site data. To demonstrate a broad application, the monthly lighting consumption is computed and projected geospatially for households in Kenya.

## 1. Introduction

Electricity access is an enabler for development, bringing a number of societal benefits such as decreasing the domestic work burden for women, increasing the study hours for children, reducing household air pollution, and supporting productive activities with the mechanization of work (Deshmukh et al., 2013; IEA, 2020). However, 11% of the global population still lacks access to electricity, and this is mostly located in rural areas of developing countries (IEA, 2020). To overcome this challenge and design the best energy access solutions, it is important to estimate accurately what their consumption would be. Studies have revealed that these households mainly use electricity to meet their lighting needs, this is because most of the times they cannot afford having other high power consuming appliances (Deshmukh et al., 2013; Dominguez et al., 2018; McNeil et al., 2010). In Kenya, for example, there is evidence that households spend an approximate of 60% of their energy bill only for lighting (Rom et al., 2020); while in Dominguez et al. (2018) it was found that in an average rural household from sub-Saharan Africa, lighting accounts for more than 50% of their total electricity consumption.

Since the residential lighting consumption depends on human behavior and daylight use, modeling approaches are likely to give uncertain results when predicting it, due to the complexity and randomness involved. To overcome these challenges, some authors proposed models applying stochastic bottom-up approaches instead of statistical top-down approaches. For example, Widén et al. (2009) applied Markov Chains for calculating transition probabilities of activity states using detailed data of households' time use available for Sweden for modeling the lighting load profiles; a similar approach was used in Widén et al. (2010) for estimating the electricity load profiles. Other models make predictions of the residential lighting load profiles based on high-resolution measurements data taken from sampled households over a certain period, such as the one proposed by M. Stokes et al. (2004), in which measurements were taken from 100 houses in the UK. However, if these approaches are applied to estimate the lighting loads of rural households in developing countries, the lack of detailed and

* Corresponding author. Chair of Building Physics, Swiss Federal Institute of Technology Zürich (ETHZ), Switzerland.
*E-mail addresses:* dominguc@ethz.ch (C. Dominguez), kristina.orehounig@empa.ch (K. Orehounig), cajan@ethz.ch (J. Carmeliet).

https://doi.org/10.1016/j.deveng.2021.100073
Received 21 April 2021; Received in revised form 18 October 2021; Accepted 18 October 2021
Available online 21 October 2021

reliable data required as input becomes a challenge; more importantly, the immense socioeconomic gap existing between the countries for which the models were created and other countries must be considered.

The study of residential lighting consumption of rural households in developing countries is mostly limited only to specific on-site case studies in which data are collected. For example, in Rom et al. (2020), authors estimate the share of lighting in the total energy expenditure, as well as the hours of use of solar lighting in Kenya, based on a randomized experimental approach. In Adeoti et al. (2011), authors made a survey-based domestic load assessment determining the lighting requirements as well as the usage of other small appliances such as radios and televisions in a rural village in Nigeria. Other studies have collected measured electricity consumption data from mini-grids and found out with complementary survey data that most of the electricity was used for lighting. For example, in Ahlborg et al. (2015), information was collected from a small-scale hydropower system in Ludewa district in Tanzania, with 1206 customers from which 600 were households using the electricity mainly for lighting and charging their mobile phones. A similar finding was presented in Williams et al. (2018), in which measured load profiles revealed that 644 out of 832 mini-grid customers across Tanzania were using electricity for said purposes. These studies give insights of lighting usage in rural areas for those specific places; but for determining it for other places, the knowledge-transfer or 'rule of thumb' methods become a source of uncertainty leading to over or underestimated estimations.

In Lombardi et al. (2019), a most recent simulation model was specifically developed for simulating high-resolution multi-energy load profiles applied to remote areas using a bottom-up stochastic approach, continuing the work of bib_Mandelli_et_al_2016bMandelli et al. (2016a, b). The model considers specific data from the studied site, such as the households' number of electrical appliances per type, indoor and outdoor lights, their specific times of use and power rating, time frames during the day in which random switch-ons can occur, among other. This model demonstrated to have an outstanding performance when validated with real consumption data from a remote village in Bolivia. However, it entirely relies on on-site data collection to generate the load profiles, and in most of the cases, data collection in rural areas of developing countries is a complicated and resource-consuming task. More recent efforts focused on deliberating a more generalized characterization of residential electricity demand in rural Kenya to facilitate the calculation of the latent energy demand using publicly available data (Falchetta et al., 2020). Authors classified households into different categories based on their appliances ownership and usage patterns, which are pre-defined based on the literature and the authors' field experiences, and these are used as input to simulate their load profiles using Lombardi et al. (2019)'s approach. A more general example of characterization of load profiles in rural areas is found in Moner-Girona et al. (2019), in which a typical load curve is assumed for domestic, social (including health centers and schools) and productive uses. The typical load curves for domestic use are adapted from the Multi-Tier Framework (MTF) proposed by the World Bank's Energy Sector Management Assistance Program (ESMAP) (ESMAP, 2018), and households are then classified into one of the tiers based on sub-locations' poverty rates.

This paper introduces a data-driven methodology that is able to characterize households based on their occupant behavior considering attributes at different spatial scales and estimating their specific lighting devices ownership. This information is then used to compute stochastic lighting load profiles at an hourly resolution, requiring only a small amount of publicly available data as input. It is built on households' survey data deployed by national and international entities, on-site data collection to bridge the data gaps, and other climate and satellite imagery data. The methodology applies unsupervised and supervised machine learning (ML) algorithms to identify the households' typical occupant behavior and to estimate the type of lamps they use (incandescent, fluorescent lamps, compact fluorescent lamps or CFL, and light-emitting diode or LED) and the number of indoor and outdoor lamps they own. After applying lighting requirement conditions based on their activity, occupancy, and daylight availability for each identified cluster, the hourly lighting load profiles are computed. The methodology is then validated using field measurements data from 13 households in Kenya. In addition, the model's performance is compared with the model introduced in Lombardi et al. (2019) and with the daily lighting consumption computed directly from the on-site deployed surveys for further analysis.

The paper is structured as follows; first, background information on East Africa is introduced, followed by a section dedicated to describing the datasets used for building up this methodology. Consequently, the applied methods are introduced in detail; while the results and discussion are presented in the next section, including the validation of the simulated lighting profiles against empirical data, along with the comparison of its performance with existing approaches. A geospatial representation of the lighting consumption of rural households in Kenya is also presented in the latter section. Finally, the study boundaries and limitations are presented, followed by the general conclusions.

## 2. The East African context

African countries have the lowest electrification rates in the world (IEA, 2021). In the East African Community countries (Tanzania, Kenya, Uganda, Rwanda, and Burundi) however, electrification rates have significantly improved over the last decade, from an average of 12.2% in 2010 to 39.6% in 2019 (IEA, IRENA, UNSD, World Bank, WHO, 2021). Kenya is the country in which this improvement has been mostly perceived, growing from 19% to 70% for said period (IEA, IRENA, UNSD, World Bank, WHO, 2021). Resulting from different electrification strategies such as the creation of the Rural Electrification Authority (REA) in 2006 (REREC, 2020) and the Last Mile Connectivity Project (LMCP) aiming to connect households within a radius of 600m of each already installed transformer was the final key to their success (LMCP, 2020).

In Tanzania, rural electrification rates have also improved from 2.5% to 18.8%; still, it remains far from its neighbors such as Kenya and Rwanda (World Bank, 2020). Tanzania is the regional leader in mini-grid development, having at least 109 mini-grids registered (Ahlborg et al., 2015). This development resulted from the small power producers (SPP) framework introduced in 2008 (revised in 2015) encouraging investments from the private sector (WRI, 2020). Table 1 presents a comparison of selected socioeconomic indicators and regional average values. By comparing the data availability needed to develop the methodology for each East African country - and especially for fieldwork data availability, Tanzania and Kenya were finally selected.

## 3. Data

For demonstrating the application and value of publicly available data for modeling high-resolution lighting profiles for rural households, the methodology is built on easy-accessible databases, complemented with one set of field-collected data to overcome data gaps. The datasets used in this study are summarized below and found in Table 2. Their detailed description is found in the supplementary material.

- Dataset 1: The 'Time Use' section of the National Household Survey Panel (NHSP) from the Living Standard Measurement Studies (LSMS), performed from 2015 to 2016 in Tanzania (World Bank, 2016) was used for creating the occupant behavior model.
- Dataset 2: The Multi-Tier Framework (MTF) Survey deployed in Kenya from 2016 to 2018 by Energy Sector Management Assistance Program (ESMAP) (ESMAP, 2018) was used for identifying the type of lighting devices owned by households.
- Dataset 3: Survey field data collected from 250 rural Kenyan households in Busia and Siaya counties (Dominguez et al., 2020)

**Table 1**
Selected socioeconomic indicators used for comparison among the East African Community.

| | Tanzania | Kenya | Uganda | Rwanda | Burundi | Average |
|---|---|---|---|---|---|---|
| Total population (millions) | 58 | 52.6 | 42.3 | 12.6 | 11.5 | 35.4 |
| Rural population (%) | 65.5 | 72.5 | 75.6 | 82.7 | 86.6 | 76.6 |
| GDP per capita, PPP (current international $) | 2770.7 | 4509.32 | 2271.6 | 2318.5 | 782.8 | 2358.1 |
| Population in multidimensional poverty (%)[1] | 55.4 | 38.7 | 55.1 | 54.4 | 74.3 | 55.6 |
| Agriculture, forestry, and fishing (% of GDP) | 28.7* | 34.1 | 21.9 | 24.1 | 28.9 | 24.7 |
| Access to electricity (% of rural population)[2] | 19 | 62 | 32 | 26 | 3 | 28.4 |

Note: All values are retrieved from World Bank (2020) and are presented for 2019, except where indicated. [1]The Multidimensional Poverty Index was created by the UNDP to complement the traditional income-based poverty indices by including multiple deprivations that households face at the same time. The values included in this table are retrieved from the 2020 report by UNDP (2020) with a range of years from 2014 to 2017. [2]The rural electrification rates are taken from the SDG7 Tracking Progress 2021 report (IEA, IRENA, UNSD, World Bank, WHO, 2021) and are for 2019. *This value for Tanzania is presented for 2018.

**Table 2**
Summary of the final datasets and their function in the methodology.

| Source | Description | Dataset | Country | Samples | Models |
|---|---|---|---|---|---|
| World Bank (2016) | National Household Survey Panel (NHSP), 2015–2016 | 1 | Tanzania | 461 | Occupant behavior |
| ESMAP (2018) | MTF Global Survey on Energy Access, 2016–2018 | 2 | Kenya | 1043 | Lighting devices |
| Dominguez et al. (2020) | Field-collected: Energy consumption patterns of rural households in Kenya (2019 | 3 | Kenya | 183 | Lighting devices |
| EnergyPlus (2020) | Hourly weather files, 2019 | 4 | Kenya | – | Occupant behavior |
| (C. D. Elvidge et al., 2017; C. D. Elvidge et al., 2021). | Nighttime lights VIIRS composites, 2019 | 5 | Kenya | – | Lighting devices |

were used for creating the models to estimate the number of lighting devices (for indoor and outdoor lights) that households own.

Additionally, other datasets were incorporated to consider specific geospatial attributes related to the households' geographic location:

- Dataset 4: Hourly illuminance profile (availability of sunlight), measured in lux, for a typical day were retrieved from EnergyPlus (2020) with location in Makindu, Kenya, which is the closest station to Tanzania.
- Dataset 5: Global nighttime lights satellite imagery data were retrieved from the Earth Observation Group, the monthly cloud-free average radiance composites for 2019 (C. D. Elvidge et al., 2017; C. D. Elvidge et al., 2021) were used.

Table 2 summarizes the role of each dataset in the methodology, along with final sample sizes. Additionally, the dataset used for validating the methodology comprises measured load profiles data collected from 13 rural households that belong to Dataset 3. These households do not own any electrical appliance other than lighting devices; therefore, they were of special interest for this study. The lighting consumption was recorded during 8 days with a time step of 30 s using electric current clamp meters.[1] For the purpose of this study, a 24 h average lighting profile was computed for each measured household.

[1] The clamp meters used measure the alternating current (AC) True Mean Root Square (TMRS) at a single-phase. The electric current measurements were transformed to power accounting for on-site measurements of the voltage.

## 4. Methods

### 4.1. General description

The proposed methodology computes the lighting hourly load profiles of rural households requiring a reduced amount of publicly available input data at household and sub-county/village scales. The methodology is based on an occupant behavior model that accounts for the household members' daily schedule of activities to identify behavioral patterns that are classified in different groups (clusters); consequently, it applies lighting requirement conditions accounting for the occupancy and activity profiles for each group. In addition, a lighting devices model is created for determining the type of lamps that households use to estimate their potential power rating; as well as for assessing the number of indoor and outdoor lamps they own. Finally, the hourly lighting profiles at a household scale are computed.

Both models are composed by sub-models in which machine learning (ML) algorithms are applied to build predictive models that are able to identify and classify behavioral patterns (as part of the occupant behavior model) and estimate the number and type of lamps owned by households (as part of the lighting devices model). Since the predictive models are evaluated in a standard manner, the predictive models section in supplementary material describes their application in the study, selection, training, and validation procedures. Fig. 1 presents a diagram of the general description of the methodology including the inputs, models, sub-models, and outputs.

### 4.2. Occupant behavior model

To determine the potential hours in which occupants will require lighting, it is important to know their hourly activities inside the house. For this, the model implements a data-driven bottom-up analysis that determines the hourly probabilities of activity inside the house that require lighting (activity profiles) and hourly estimations of the number of people being at home (occupancy profiles). A cluster analysis (unsupervised ML) is first implemented for identifying clusters of behavioral patterns; then, RF is applied creating a predictive supervised ML model for classifying the input data into one of the identified clusters. Finally, lighting requirement functions are formulated based on the activity and occupancy profiles, as well as on the daylight availability.

#### 4.2.1. Cluster analysis

Schedule patterns and their drivers are identified using the diaries from Dataset 1. For this, the Ward's agglomerative hierarchical clustering algorithm (Ward, 1963) was applied, considering a bottom-up approach based on a variance sum-of-squares criterion minimizing the dispersion within groups at each iteration (Murtagh and Legendre, 2014). It initially considers each observation as an independent cluster, and at every iteration, it combines the more similar clusters generating a new one. The process continues until all the observations are assigned to

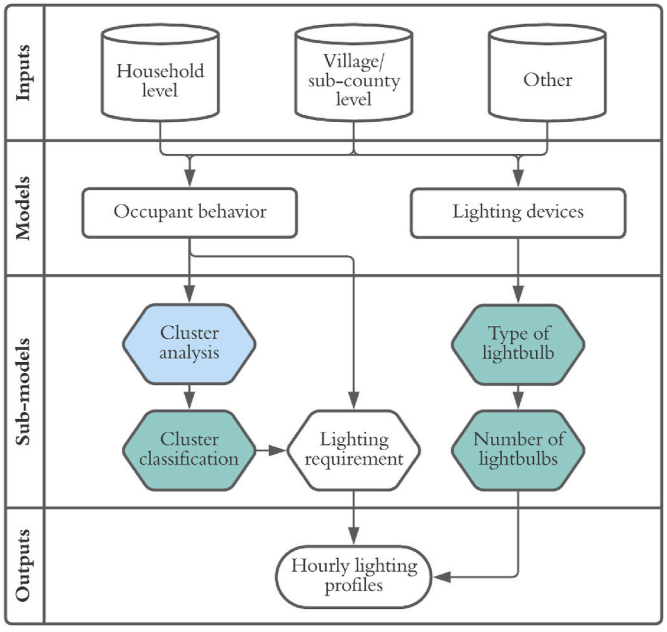


**Fig. 1.** Flow diagram with general description of the methodology. Predictive models are included in the sub-models section and represented by the colored polygons. The blue polygon represents the application of unsupervised ML, while the green ones, the application of supervised ML.

a single cluster. The variances are measured through a distance matrix. Considering that the dataset used in this study contains sequential categorical data (activity states) distributed in time steps, the sequence analysis algorithm TraMineR (Gabadinho et al., 2011) was applied for mining the sequences and determining the most suitable dissimilarity measure to obtain the distance matrix. This dissimilarity depends on the data structure and the study goals; in addition, the methods can be sensitive to specific aspects of a sequence such as the duration of the different successive states, the counts of common attributes, or the order of the different successive states as they appear (Studer et al., 2016). Since the aim of the study is to identify patterns to classify households based on their activity states distributed in a 24 h period, the main interest is to select a method that accounts for dissimilarities in transition rates from one state to the other while considering the states distribution over time and the common attributes among sequences. In the literature, a commonly used method for selecting the optimal dissimilarity measure, while finding at the same time the optimal number of clusters is the silhouette width coefficient (Rousseeuw, 1987; Caliński and Harabasz, 1974). The optimal dissimilarity measure and number of clusters is the one that presents the largest silhouette width coefficient. Table 3 presents a description of the dissimilarity measures tested for comparison.

Linear discriminant analysis (LDA) was applied to study the relationship between the found patterns and the set of variables included as attributes for each observation in the diaries, in other words, to identify the most significant variables that are able to define and characterize the clusters. The Pseudo F, Pseudo R,[2] and Levene tests were applied to measure the statistical significance of each covariate and its importance for defining the clusters, using the t-value and p-value as significance indicators. These tests give information on the ratio of the variance between clusters to the variance within the clusters (Caliński and Harabasz, 1974), the proportion of variance explained by the analyzed covariate (Smith and McKenna, 2013), and the equality of variances for the covariate calculated for two or more groups (Gastwirdth et al., 2009), respectively.

**Table 3**
Dissimilarity measures tested and descriptions (Studer et al., 2016).

| Dissimilarity measure | Description |
|---|---|
| Dynamic Hamming distance (DHD) | Sum of mismatches with position wise state-dependent weights. |
| The length of the longest common subsequence (LCS) | Number of elements in one sequence that can be matched with elements occurring in the same order in the other sequence. |
| Optimal matching based on transition costs (OMtrans) | OM between sequences of transitions. |
| Optimal matching based on spells length costs (OMspells) | OM between sequences of spells. |

Note: The optimal matching methods measure the dissimilarity of two sequences as the minimum total cost of transforming one into the other (Studer et al., 2016).

#### 4.2.2. Cluster classification

After applying the LDA and reducing the dataset to the most significant variables that define the clusters, a predictive model that will classify any new input into one of the identified clusters is created. For this purpose, the RF algorithm is applied, for which the SHAP values for each variable and f1score as respective performance indicator are estimated (see supplementary material for information on SHAP values and f1score).

#### 4.2.3. Lighting requirement

Once clusters of behavioral patterns are identified, it is important to define when people require the use of lighting in their houses. The hourly lighting usage depends on the type of activities performed inside of the house, the daylight availability, and the number of people at home, as more people might require the use of different spaces simultaneously. Hence, lighting requirement functions are proposed accounting for these factors.

*4.2.3.1. Activity-based.* Based on the list of activities included in the diaries from Dataset 1, the activities that are performed inside of the house and that may require the use of lighting were classified and selected for further analysis. The selected activities are specified in the first column of Table 4, while the excluded activities are in the second column. Consequently, the probability of occurrence at every time step P (t) of the selected activities was estimated for each of the identified clusters.

The condition in Eq. (1) was applied for identifying the lighting requirement based on P(t) of the selected activities, hereafter probability of activity, at every time step.

$$Light_{act}(t) = \begin{cases} 1, & P(t) > P_{\min} \\ 0, & P(t) \leq P_{\min} \end{cases} \quad (1)$$

**Table 4**
Classification of activities included in the analysis based on the diaries.

| Activities performed inside/requiring lighting | Activities performed outside/not requiring lighting |
|---|---|
| Taking care of children | Farming |
| Cooking | School |
| Domestic work | Shopping |
| Eating | Sleeping |
| Entertainment | Social activities |
| Exercising | Travelling |
| Own business work | Work as employed (having a salaried job) |
| Personal care | Other |
| Religion (praying) | |
| Weaving | |

[2] Remote-Areas Multi-energy systems load Profiles, open-source code and documentation available in https://github.com/RAMP-project/RAMP.

where $P_{min}$ represents a threshold of minimum probability of activity. If P(t) is larger than $P_{min}$, then lighting is required, $Light_{act}(t) = 1$. To determine the value of $P_{min}$, a range of potential values is set accounting from half of the maximum value to the maximum value of probability of activity during the day from each cluster, $P_{min} = [P_{max}/2, P_{max}]$. Then, a random selection is applied to the defined range for each cluster to select the final value of $P_{min}$.

*4.2.3.2. Daylight availability.* Lighting requirement depends on the daylight availability, as it is less likely that lighting will be required during daytime. Therefore, the direct normal illuminance (measured in lux) was extracted from the weather files of a typical day retrieved from Dataset 4. A typical day is considered because in Tanzania and Kenya the illuminance does not have significant variations over the year, since they are close to the equator. A minimum illuminance $L_{min}$ is defined as the minimal requirement of lighting considering the daylight availability at time t. As suggested in (Widén et al., 2009), $L_{min}$ is estimated as half of the maximum illuminance value of the typical day. If the illuminance at each time step L(t) is less than $L_{min}$, then lighting is required, $Light_L(t) = 1$. The condition for lighting requirement considering the daylight availability is presented in Eq. (2).

$$Light_L(t) = \begin{cases} 1, & L(t) < L_{min} \\ 0, & L(t) \geq L_{min} \end{cases} \quad (2)$$

*4.2.3.3. Occupancy-based.* In Eqs. (1) and (2), it is defined that the lighting devices are used if the conditions $Light_{act}(t) = 1$ and $Light_L(t) = 1$ are fulfilled. However, since there is also probability that household members perform activities that do require the use of lighting even during the daylight hours, then the occupancy of people in the house Occ (t) is also considered in Eq. (3). As for $P_{min}$ and $L_{min}$, a minimum occupancy threshold, $Occ_{min}$, has to be defined dependent on the household size (HHS). Therefore, if there is at least the minimum occupancy in the house and $Light_{act}(t) = 1$, even if $Light_L(t) = 0$, lighting might be required. For this, the occupancy profiles need to be estimated accounting for the number of people present in the house at each time step. As performed in (Dominguez et al., 2018), to estimate the occupancy at time t, Occ(t), the probability of activity P(t), and the average household size HHS of the analyzed cluster are considered, hence, the occupancy at time t equals:

$$Occ(t) = P(t) \cdot HHS \quad (3)$$

As for $P_{min}$, to determine the value of $Occ_{min}$, a range of potential values is set accounting from half of the maximum value to the maximum value of occupancy during the day from each cluster, $Occ_{min} = [Occ_{max}/2, Occ_{max}]$. Then, a random selection is applied to the defined range for each cluster to select the final value of $Occ_{min}$. As a final condition, to determine how many of the total number of lampsa household owns will be turned on at what time, a factor of the number of lamps per person is calculated, that will then be multiplied by Occ(t) at every time step.

### 4.3. Lighting devices

Besides having lamps in the rooms, installing lamps outside of the living space is revealed to be important for rural households, as it provides a sense of security during the nighttime (Dominguez et al., 2021; Mandelli et al., 2016a,b; Van Ruijven et al., 2011), these are often called security lights or outdoor lights in the literature. Knowledge on the amount of indoor and outdoor lamps and their power rating is needed for determining the lighting load profiles. Hence, prediction models were created first for estimating the type of lighting device that households own for identifying ranges of potential power ratings; and then for estimating the amount of both indoor and outdoor lamps. It is important to note that the amount of outdoor lights is calculated separately because they do not have the same hours of use as the indoor lights. While the hours of use of indoor lights depend on the lighting requirement conditions mentioned in the previous section, the ones for the outdoor lights are defined considering that these will be working during the lowest values of illuminance L(t) →0, adding up to approximately 12 h s of use in this case. This amount of hours coincides with the one mentioned in Mandelli et al. (2016) specifying the window of potential use of outdoor lights in a typical rural household.

#### 4.3.1. Type of lamps

Dataset 2 was used for estimating the type of lamps households own, as it indicates their access to different types in a binary form. A predictive model that classifies households into different types of lamps is proposed, for which RF was applied, including demographic and socioeconomic characteristics at household and village level. In addition, Dataset 5 containing the nighttime lights average radiance was used for extracting the values for each georeferenced data sample included in Dataset 2. Based on the literature and technical specifications of commercial lamps, the most common ranges of power rating for each type of lamps were defined; consequently, the model selects randomly a power rating value based on the type of lamp and its range of possibilities. Table 5 presents the types of lamps included in this study, the power rating ranges, and the common ranges of luminous efficacy. Due to data limitations in terms of separating the type of lamps owned for indoor and outdoor purposes in a household, the assumption that households use only one type of lamp to meet both purposes is made. However, field studies such as Lombardi et al. (2019) and Mandelli et al. (2016) have shown that households tend to use outdoor lights with more power rating than the ones they use indoors. While in Dominguez et al. (2021) the same trend was observed in households that own solar home systems (SHS); however, for households that have grid-connection the power rating for indoor lights is slightly larger than the outdoor ones.

#### 4.3.2. Indoor and outdoor lamps ownership

Since the number of lamps per household is not included neither in Dataset 1 nor in Dataset 2 and cannot be found in national household surveys for both countries, a method to estimate the number of lamps has to be proposed. Hence, the amount of lamps for indoor and outdoor purposes was extracted from Dataset 3 from the field data collection from 183 households in rural Kenya, creating predictive models applying RF, including household and village level variables. The frequency distribution of both number of indoor and outdoor lamps found in the dataset are presented in Fig. 2 predictive models that estimate the lamps ownership were created as classification problems, as the outcome variable is not continuous.

### 4.4. Lighting profiles validation

The modeled hourly lighting load profiles for the measured households are validated against real consumption data; therefore, indicators are computed to measure their prediction accuracy. These are the normalized root mean squared error (NRMSE), which is a normalization of Eq. (4), divided by the average load value of the measured daily load

**Table 5**
Type of lamps and power rating and luminous efficacy ranges.

| Type of Lighting Device/ Lamp | Power Rating Range[1] (W) | Luminous efficacy[2] (lm/ W) |
|---|---|---|
| Incandescent | (45, 60] | [10.4, 15] |
| Fluorescent | (15, 45] | [60, 100] |
| Compact Fluorescent Light (CFL) | (10, 15] | [46, 75] |
| Light Emitting Diode (LED) | [3, 10] | [30, 120] |

Sources: [1] Philips (2021), U.S. Department of Energy (2021), Mandelli et al. (2016a,b), Adeoti (2011), Mahapatra et al. (2009), Sebitosi & Pillay (2007), Nieuwenhout et al. (1998). [2] Philips (2021), Kumar and Choudhury (2014).

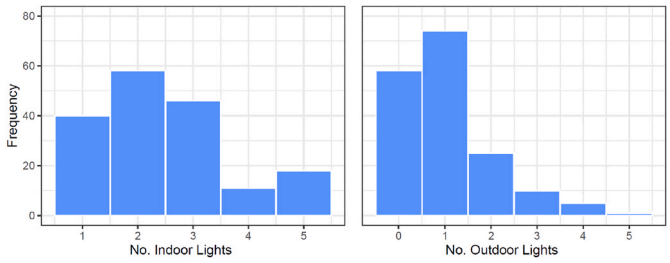


**Fig. 2.** Frequency distribution for number of indoor and outdoor lamps in Dataset 3.

profile, presented in percentage form. Another indicator is the load factor (LF), given by Eq. (4), representing the average load value of the measured daily load profile ($P_{avg}$) divided by the maximum load ($P_{peak}$) over the same period. The daily aggregate consumption (Wh/day) is also considered, which is the sum of the hourly loads in one day. Since the latter indicator can be computed from the survey-collected data for each measured household, this was also used for validation.

$$LF = P_{avg} / P_{peak} \tag{4}$$

For further evaluation, the simulation model introduced in Lombardi et al. (2019) for generating bottom-up stochastic load profiles for remote areas, RAMP,[2] is also implemented for generating each household's lighting profile. In Lombardi et al. (2019), RAMP demonstrated to have an outstanding performance on estimating the electricity profiles for a remote area in Bolivia. This model relies on site-interview-based data as input to simulate the profiles; therefore, the survey-collected data from the measured households, contained in Dataset 3 are used (Table 6). It is important to note that RAMP requires the number of indoor and outdoor lights, and their specific times of use and power rating as inputs for simulating the load profiles; while in the model presented in this paper, these are estimated based on a small number of variables. Therefore, for comparing the performance of both on predicting the measured load profiles, their different modeling approaches need to be considered.

## 5. Results and discussion

### 5.1. Cluster analysis

For selecting the optimal dissimilarity measure to use in the cluster analysis, the silhouette width coefficient was calculated for a range of two to six clusters, as shown in Fig. 3. A maximum silhouette width coefficient is obtained for four clusters (k = 4) for all the evaluated measures.

**Table 6**
Inputs used for RAMP simulation model extracted for the measured households from Dataset 3.

| | Indoor lamps | | | Outdoor lamps | | |
|---|---|---|---|---|---|---|
| | Number | Power rating | Hours of use | Number | Power rating | Hours of use |
| HH1 | 5 | 25 | 7 | 1 | 25 | 7 |
| HH2 | 5 | 8 | 4 | 2 | 20 | 11 |
| HH3 | 1 | 15 | 2 | 1 | 12 | 2 |
| HH4 | 4 | 23 | 4 | 3 | 75 | 8 |
| HH5 | 5 | 18 | 5 | 5 | 24 | 5 |
| HH6 | 2 | 12 | 4 | 0 | 0 | 0 |
| HH7 | 3 | 20 | 3.5 | 2 | 30 | 6 |
| HH8 | 5 | 18 | 4 | 4 | 18 | 8 |
| HH9 | 1 | 30 | 3 | 1 | 20 | 12 |
| HH10 | 3 | 18 | 4 | 2 | 75 | 5 |
| HH11 | 2 | 15 | 4 | 2 | 75 | 4 |
| HH12 | 2 | 9 | 4 | 1 | 9 | 1 |
| HH13 | 4 | 20 | 3 | 4 | 20 | 12 |

Note: For all other input parameters required by RAMP and not found in Dataset 3, the standard values based on the case study in Lombardi et al. (2019) were used.

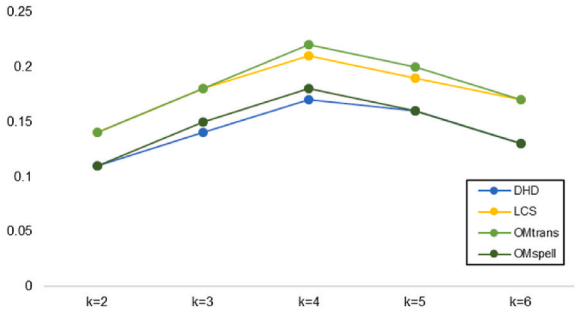


**Fig. 3.** Silhouette width coefficient calculated for a range from two to six clusters.

For having the largest silhouette width coefficient, the optimal dissimilarity measure selected was OMtrans, which classifies all the observations in the dataset into four clusters. Fig. 4 presents the states distribution in each cluster. The results of the linear discriminant analysis (LDA) used for measuring the significance of each covariate for defining the clusters are presented in Table A.1. The final selected covariates are presented in Table 7; for each of them, the percentage of population within each cluster is given.

Cluster 1 (n = 165) characterizes households with at least 50% female population that perform farming activities as income and livestock raising, and their ownership of large and small livestock, and poultry is larger than in the rest of the clusters. They own less large appliances than the rest. Cluster 2 (n = 177) characterizes households with more than 60% of female population. As in cluster 1, they also perform cash farming and livestock raising activities, but almost 20% less. These households mostly perform activities inside of the house, such as taking care of children, cooking, and doing domestic work (Fig. 3). They own more large appliances, and less mobile phones than the ones in cluster 1. This suggests that these households might have other income sources besides farming, which allows them to afford larger appliances. As mobile phones are low power-consuming and affordable devices, this could explain why cluster 1 has a larger share of them compared to cluster 2.

Cluster 3 and 4 represent the lowest shares in the total sampled population (n = 62 and n = 57 respectively) showing different patterns of behavior compared to the other two. These clusters characterize households in which the main economic activity is having a business at home (cluster 3) and a salaried job (cluster 4). They both have males representing more than 60% of their population. Households in cluster 3 perform more cash farming activities than those in cluster 4, but less livestock raising activities, owning almost 10% less large livestock and poultry, and 1% less small livestock. It is interesting to note that cluster 1 and 2 have more in common on their demographics and economic activities, dedicating their time mostly to farming activities and livestock raising. As for clusters 3 and 4, they have diverse sources of income, which explains their high ownership of mobile phones and large appliances.

For cluster 1, activities performed inside the house such as taking care of children, domestic work and entertainment have three peaks during the day, at 09:00, 16:00 and at 21:00, respectively. The latter alone have their highest peak at 23:00. Conversely, these are performed throughout the day for cluster 2. Instead of having three peaks as cluster 1, it has a predominant one at 10:00 and another at 22:00, having the activity peak in morning hours, while social activities are performed between the morning and night peaks. Households from cluster 3 allocate their time to their own business from 10:00 to 21:00, having similar peaks of activities inside the house, as cluster 1, with the difference that in cluster 3, the morning peak is larger and the afternoon peak is more distributed over time. Cluster 4 presents two well defined peaks of

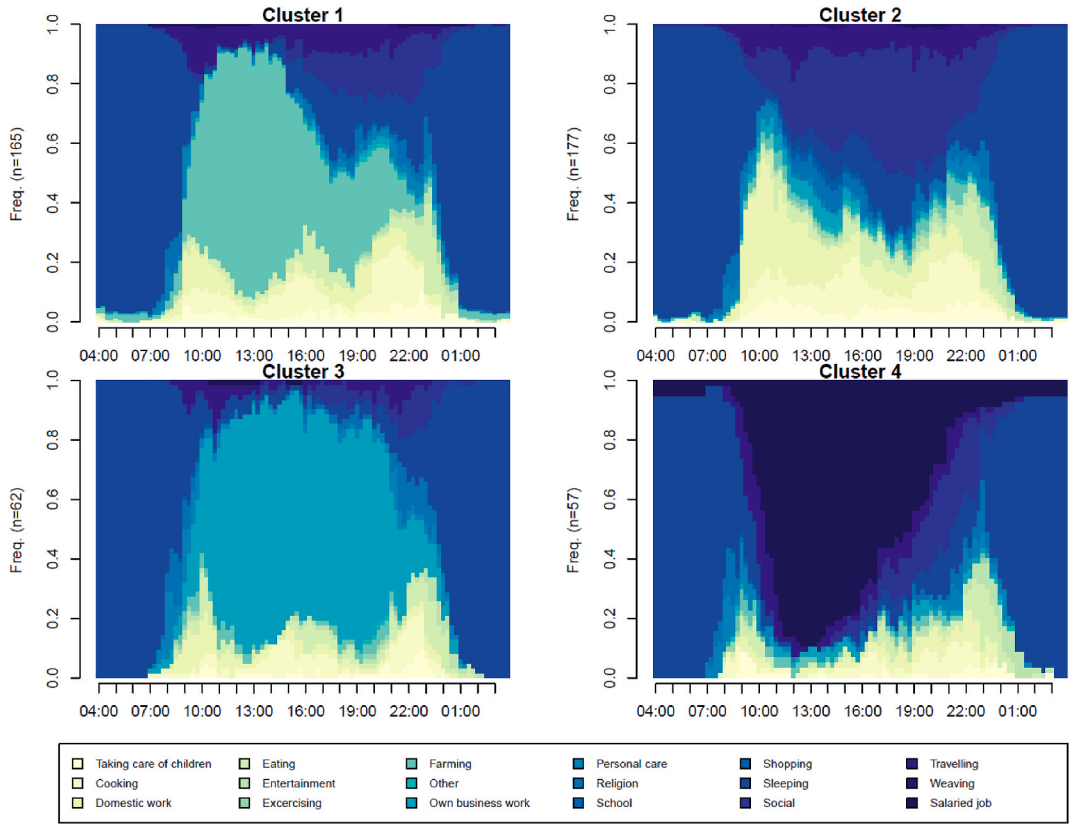


**Fig. 4.** Cluster's grouped schedule patterns of activity.

**Table 7**
Variables selected from discriminant analysis and their characteristics by cluster.

| Variable | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Respondent gender * (% of females) | 50.3 | 66.7 | 37.1 | 31.6 |
| Relationship to HH head * (% of household heads) | 55.8 | 46.3 | 71.0 | 78.9 |
| Cash farming activities[a](%) | 60.0 | 43.5 | 37.1 | 24.6 |
| Livestock raising (%) | 84.2 | 64.4 | 45.2 | 54.4 |
| Large livestock ownership[b] (%) | 41.2 | 27.1 | 9.7 | 19.3 |
| Small livestock ownership[c] (%) | 43.6 | 29.9 | 16.1 | 17.5 |
| Poultry ownership (%) | 75.2 | 63.3 | 41.9 | 52.6 |
| Large appliances ownership[d] (%) | 18.8 | 25.4 | 29.0 | 40.4 |
| Mobile phone ownership (%) | 77.0 | 73.4 | 82.3 | 82.5 |

Note: *Information from respondent, the rest of information corresponds to the household as a whole.
[a] They sell what they harvest.
[b] Oxen, cattle.
[c] Goats, pigs, sheep.
[d] Televisions, refrigerators.

activities inside the house, one at 09:00 as they leave for work and one at 23:00.

#### 5.1.1. Cluster classification

RF was applied to create the predictive model for classifying the identified clusters. Due to the existence of class imbalance among the clusters found in the training dataset (Fig. 5, left), weighting was applied to give all classes an equal weight. The training parameters used, and f1score as performance indicator are found in Table 8, and the confusion matrix is presented in Fig. A2. Mean absolute SHAP values for each variable considered in the model are included in Fig. 5 at the right, while the SHAP values (with the positive and negative contributions) are found in Fig. A1 in the Appendix.

Based on this figure, the most influential variable to define the clusters is the small livestock ownership, having a negative contribution for clusters 1 and 4, and positive contribution for clusters 2 and 3 if the household does not own small livestock. The variable with the least contribution is the ownership of poultry, which affects positively clusters 1 and 3, and negatively clusters 2 and 4 especially if the household owns poultry (refer toFig. A1 Figure A.1 for further explanation).Fig. A2 Figure A.2 presents the confusion matrix as a visualization of the model's performance. The matrix columns represent the true class households belonging to each cluster, while the rows show the households that were predicted to belong to each cluster. Meaning that the sum of the diagonal values are the households that were predicted correctly for each cluster.

### 5.2. Lighting requirement

After the application of the lighting requirement conditions given by Eqs. (1)–(3), the probability of activity at home requiring lighting was computed for all four clusters. For illustrative purposes, in Fig. 6, the direct normal illuminance for a typical day (a), together with number of people in the house (b) and their activity profiles (c), are presented for cluster 1. The activity profiles for the remaining clusters are included in Fig. A3.

Similarities were found in the activity profiles between clusters 1 and 4, since they are mostly performing activities outside the house (farming and having a salaried job, respectively). In addition, similarities were found for clusters 2 and 3, since they are mostly performing activities inside the house (cooking and domestic work, and having a business at home, respectively).

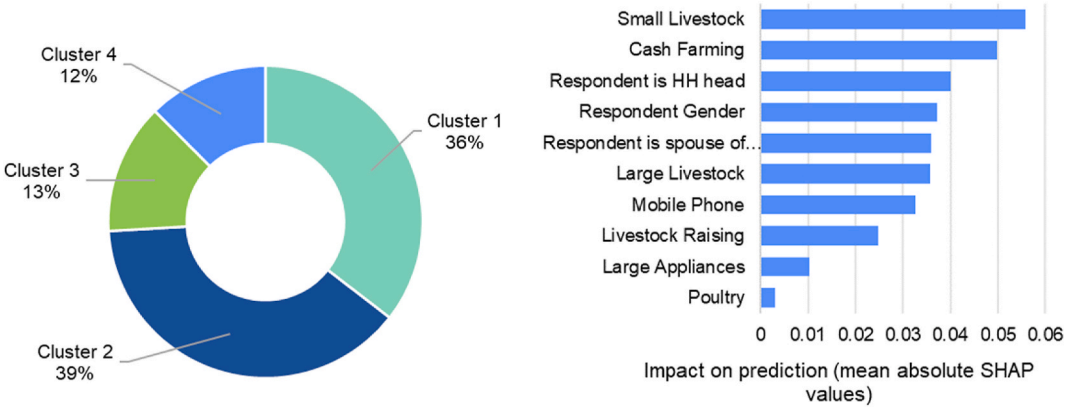


**Fig. 5.** Left, class imbalance towards Cluster 1 and Cluster 2. Right, impact of each variable on the model's prediction, expressed in mean absolute SHAP values.

**Table 8**
Training parameters and performance of the predictive model for cluster classification.

| | Cluster Identification |
|---|---|
| Samples | 459 |
| Training | 346 |
| Testing | 113 |
| CV Folds | 10 |
| CV Repeats | 3 |
| Variables Sampled (mtry) | 5 |
| Trees Number (ntrees) | 500 |
| Node Size (sizenode) | 0.1 |
| f1score (training) | 0.88 |
| f1score (testing) | 0.78 |

### 5.3. Type of lamps

For estimating the type of lamps that households own, RF was applied as a classification problem, which hyperparameters are defined as in Table 4, and the confusion matrix is presented in Fig. A2.Fig. 7 introduces the variables that were included in the final model, showing the contribution on the predictions in mean absolute SHAP values. Both household and village level variables were included in the analysis; however, only household attributes showed to be relevant for defining the type of lamps. These attributes range from household size, housing materials (especially the walls), and assets such as mobile phones, motorbikes, and poultry. In Fig. A4 the positive and negative contributions of each variable are observed. Interestingly, the nighttime lights are relevant for predicting the type of lamps used by households; strong nighttime lights captured by satellite images (having a high average radiance) indicate the use of fluorescent and CFL lamps. Having a negative contribution on the use of incandescent and LED lamps. When comparing the ownership distribution for each type of lamp against the households' NL values from both Datasets 2 and 3 (Fig. 7), it is noted that with higher NL values, the distribution of samples owning fluorescent and CFL lamps is slightly more concentrated. It is important to consider that households in these databases are located in areas with very low NL values (especially the ones in Dataset 2), which makes it difficult to identify a clear trend. The VIIRS DNB satellite (providing the

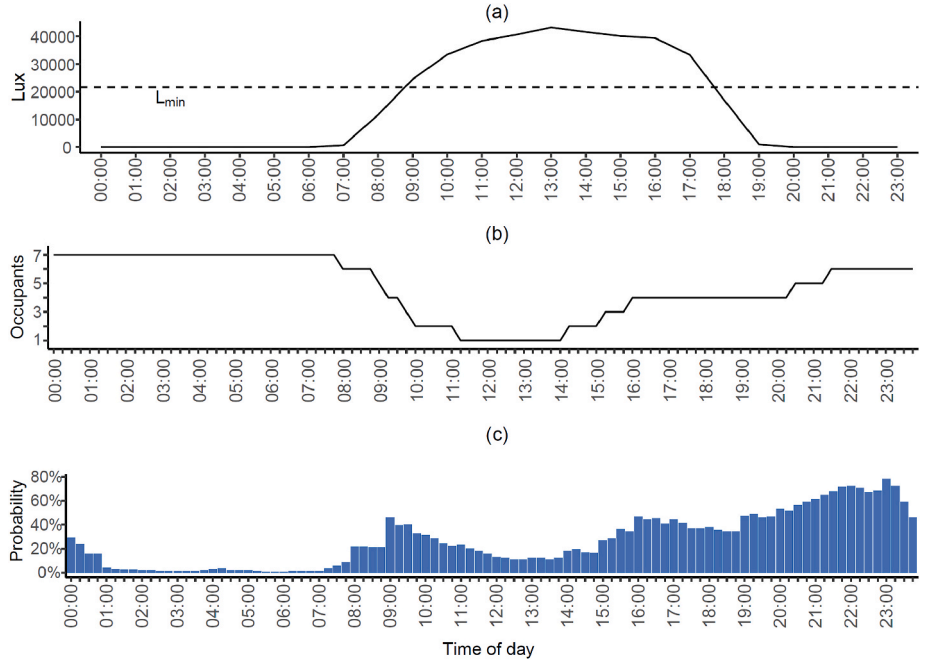


**Fig. 6.** (a) Direct normal illuminance (in lux) used for representing a typical day of the year with $L_{min}$ used for this study. (b) Occupancy profile (number of people at home at time t) for cluster 1. (c) Activity profile of cluster 1 representing the probability of activity at home that requires lighting.

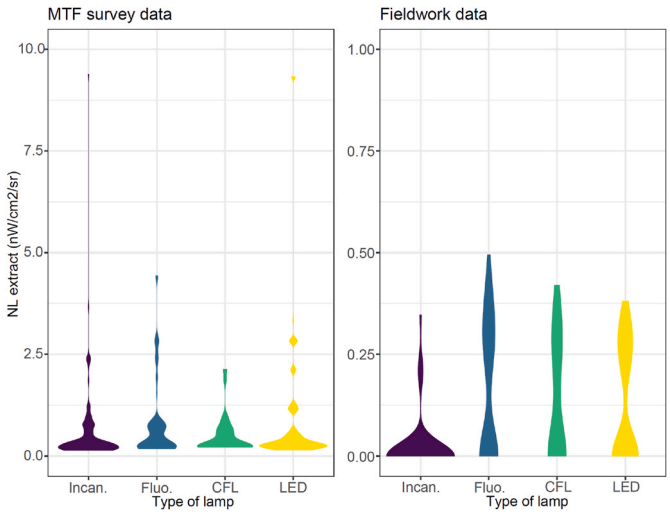


**Fig. 7.** Distribution of ownership for each type of lamp against the households' NL values from both datasets 2 and 3.

NL images used in this study) has a spectral response between 400 and 900 nm, which can be able to capture many human-made light sources (C.D. Elvidge et al., 2017). The relationship between different types of lamps with satellite images has been studied in the literature. For example, Sanchez de Miguel et al. (2019) and N. Levin et al. (2020), reported that these images are capable to provide information on the color of artificial lighting during the nighttime. However, without knowing the technology that is producing the lights it is not possible to identify their luminous efficacy due to the broadband spectra that can be produced by a single technology (especially LED). The VIIRS DNB satellite, for example, is unable to capture the incoming band in the blue band (N. Levin et al., 2020). Furthermore, there are studies focused on identifying the type of lighting technology from remote sensing images. For example, by taking ground-measurements using visible and near infrared hyperspectral imaging, Dobler, et al., 2016 showed that the images captured mostly fluorescent and LED technologies, while incandescent lamps were largely undetected, attributing this to their low signal amplitude and primary usage for indoor lighting, which is barely detected.

### 5.4. Indoor and outdoor lamps ownership

Predictive models applying RF were created for estimating the indoor and outdoor lamps ownership considering all household and village level attributes included in the Dataset 3. For determining the number of indoor lights, a classification problem was formulated. This was also the first approach for determining the number of outdoor lamps. However, after testing different problem formulations, the correlation between this variable with the rest of covariates – with different combinations and sampling methods, was not evident; moreover, the models had no statistical significance (p-value over 5%), reaching unsatisfactory accuracy levels (f1score below 50%). Consequently, a different approach was used formulating a regression problem in which the dependent variable is continuous, representing a fraction of the number of outdoor per indoor lamps. The final hyperparameters used in the models and their performance are presented in Table 9.Fig. A2 Figure A.2 presents the confusion matrix for the indoor lamps model, and the correlation plot of modeled and predicted values for the outdoor lamps model.

In Fig. 8, the variables with the greatest impact on the final predictions (in mean absolute SHAP values) for both indoor and outdoor lamps are presented. Their ownership depends not only on household level, but also on village and sub-county level variables, such as the electrification rate, streetlights access rate, and population density.

**Table 9**
Training parameters and performance of the predictive models for the number of indoor and outdoor lights.

| | Type of lamp | Indoor lamps | Outdoor lamps |
|---|---|---|---|
| Samples | 1043 | 183 | 148 |
| Training | 785 | 138 | 103 |
| Testing | 258 | 45 | 45 |
| CV Folds | 10 | 10 | 10 |
| CV Repeats | 3 | 3 | 3 |
| Variables Sampled (mtry) | 9 | 11 | 5 |
| Trees Number (ntrees) | 500 | 500 | 103 |
| Node Size (sizenode) | 0.1 | 0.1 | 0.1 |
| f1 Score (training) | 0.96 | 0.75 | – |
| RMSE (training) | – | – | 0.35 |
| R2 (training) | – | – | 0.77 |
| f1 Score (testing) | 0.80 | 0.75 | – |
| RMSE (testing) | – | – | 0.26 |
| R2 (testing) | – | – | 0.76 |

Interesting trends are found when analyzing the positive and negative contributions of each variable, presented in Fig. A5 and A6. For example, having large appliances has a positive contribution on owning more indoor and outdoor lamps, while having access to a solar home system (SHS) increases the number of indoor lamps, but decreases the ownership of outdoor lamps. Generally, having good quality walls and floor materials increases the ownership of both and having a male as household head increases the ownership only of indoor lamps. The type of lamp also has a role on defining the number of outdoor lamps, as the ownership of incandescent and fluorescent lamps increases their ownership. Furthermore, high access to streetlights and low population density has a negative contribution on their ownership.

In Fig. 9, the jitter[3] representation of the distribution of indoor and outdoor lamps in the dataset, classified by type of lamp is shown. LED lamps are used in wide-ranging quantities for indoor purposes, while their use is quite restricted for outdoor purposes. For the latter, fluorescent are preferred. This might be triggered by the price per lamp. According to different local markets consulted during the field study in Kenya, the most common prices are 1US$, 2US$, 2-3US$ and 5-8US$, for incandescent, fluorescent, CFL, and LED lamps, respectively. Therefore, for having cheap (considering only the purchase price) and strong outdoor lights for bringing security to households, fluorescent lamps are their most reasonable choice.

### 5.5. Application: generating lighting profiles for measured households

The lighting load profiles are generated for the 13 households with measured profiles. For this, the variable inputs needed are extracted from Dataset 3, first for determining to which of the identified clusters they belong, second for defining the type of lamps that they use, and third for estimating how many indoor and outdoor lights they own. To account for the nighttime lights, the average monthly radiance values were extracted from the satellite image with approximate geographic coordinates for each household, while the streetlights access was taken from Dataset 2 for each village. The overall required input variables for the proposed methodology are described in Table 10, the detailed inputs for the sampled households are found in Table 1 and the simulated results for each sub-model are presented in Table 2, both presented in supplementary material. In Fig. 10, the simulated lighting load profiles for each household are grouped by cluster. Different trends can be observed by cluster, for which the most evident is the daytime lighting usage between clusters 1 and 4, and clusters 2 and 3. The simulations show that households from cluster 1 and 4 do not present any lighting

[3] Data visualization technique of adding random noise to prevent plotting data on top of each other when they have the same coordinates range.

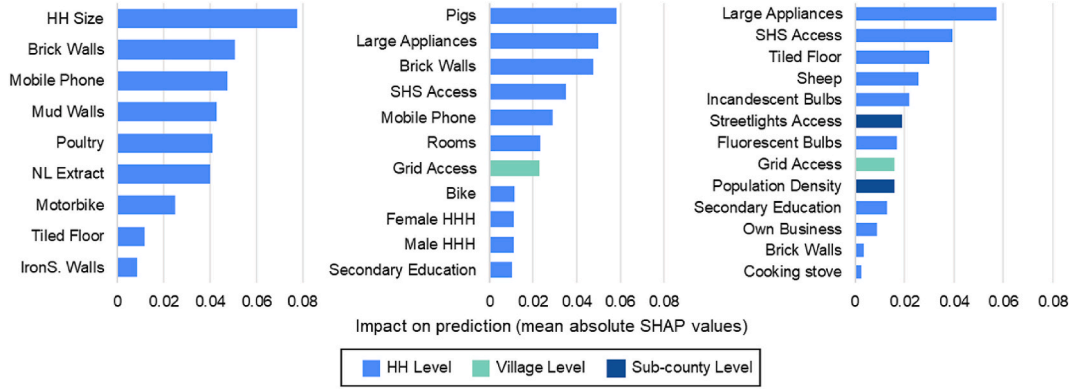


**Fig. 8.** Impact of each variable on each model's prediction, expressed in mean absolute SHAP values. Left, for type of lamps; center, for indoor lamps; right, for outdoor lamps.

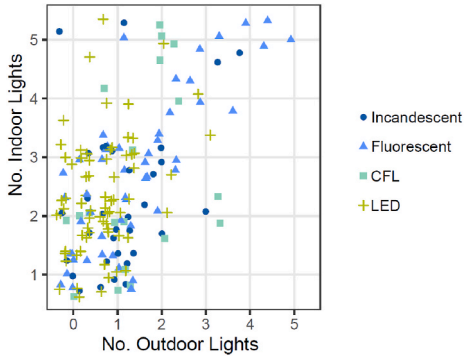


**Fig. 9.** Distribution of number of indoor lights versus number of outdoor lights, per type of lamp.

use during daytime; oppositely, households from cluster 2 present the most lighting activity variation during these hours, especially from 07:00 to 11:00. However, the peak power for all clusters is mostly happening between 19:00 to 22:00.

### 5.6. Validation with empirical data

To validate and test the accuracy of the methodology presented in this paper for predicting hourly lighting load profiles for rural households, the simulated profiles from the previous section were compared to the on-site measured profiles from 13 households in rural Kenya. In addition, RAMP model was used for generating these profiles on an hourly basis for a typical day using as inputs the survey-collected data (Dataset 3), presented previously in Table 6. The values in this table were also used for estimating the aggregate consumption (Wh/day) for the survey comparison method. To facilitate the comparison, the simulated load profiles were averaged by cluster. This procedure was also applied to the ones generated by RAMP and to the measured profiles. The cluster means comparison is presented in Fig. 11, while the results for each indicator used for comparison are presented for each cluster in Table 11.

The first observation from Fig. 11 is that the mean lighting load profile from cluster 1 is overestimated by RAMP, this is due to the inconsistency between the information provided by the household through the survey and the real measured data. Especifically, because HH4 (belonging to cluster 1) reported in to consume 4.13 times more electricity than it actually consumes; which affects the cluster's mean value. The errors caused by the discrepancy between survey and measured load profiles data have been previously explored in studies such as Hartvigsson et al. (2018) and Blodgett et al. (2017).

**Table 10**
Required overall input variables for the proposed methodology, their description, and the sub-models in which they are applied.

| Input variable | Description | Sub-models[1] |
|---|---|---|
| Household head gender | Male or female | 3 |
| Respondent gender | If survey respondent is male or female | 1 |
| Relationship to HH head | If survey respondent is the household (HH) head, spouse or other | 1 |
| Access to secondary education | If survey respondent had access or not to secondary education | 3, 4 |
| Main occupation | Occupation in which the main income relies, it can be farming, own business, or other | 1, 4 |
| Livestock raising | If the household raises livestock (small or large) | 1 |
| Household size | Number of people in the household | 2 |
| Number of rooms | Number of rooms in the household | 3 |
| Type of cooking stove | It can be open fire or other | 4 |
| Walls material | Predominant walls material, it can be bricks, mud, corrugated iron sheets, or other | 2, 3, 4 |
| Floor material | Predominant floor material, it can be ceramic tiles or other | 2, 4 |
| Access to SHS | If the household has access or not to a solar home system (SHS) | 3, 4 |
| Large livestock ownership | If the household owns oxen or cattle | 1 |
| Small livestock ownership | If the household owns goats, pigs or sheep*, and how many pigs** and sheep*** | 1*, 3**, 4*** |
| Poultry ownership | If the household owns poultry*, and how many** | 1*, 2** |
| Large appliances ownership | If the household owns television, DVD, refrigerator, sound equipment, sewing machine, portable computer*, and how many** | 1*, 3**, 4** |
| Mobile phones ownership | If the household owns mobile phones*, and how many** | 1*, 2**, 3** |
| Motorbikes ownership | Number of motorbikes owned by the household | 2 |
| Bicycle ownership | Number of bicycles owned by the household | 3 |
| Nighttime lights extract | Average monthly radiance (nW/cm2/sr) | 2 |
| Village electricity grid access | Percentage (%) of households with access to grid-electricity | 3, 4 |
| Sub-county streetlight access | Percentage (%) of households reporting that their street has access to streetlights (sub-county/division values from Dataset 2) | 4 |
| Sub-county population density | Number of people per km2 | 4 |

Note [1]: The sub-models are presented as follows: 1 = Cluster classification, 2 = Type of lamp, 3 = Indoor lamps ownership, 4 = Outdoor lamps ownership.

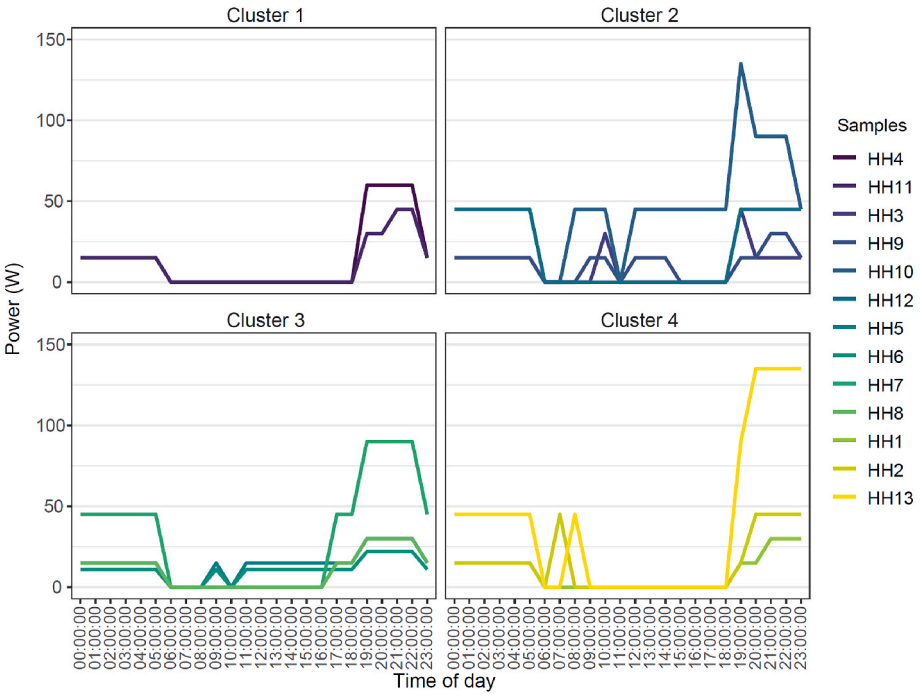


**Fig. 10.** Results from the simulated lighting load profiles for households in each cluster.

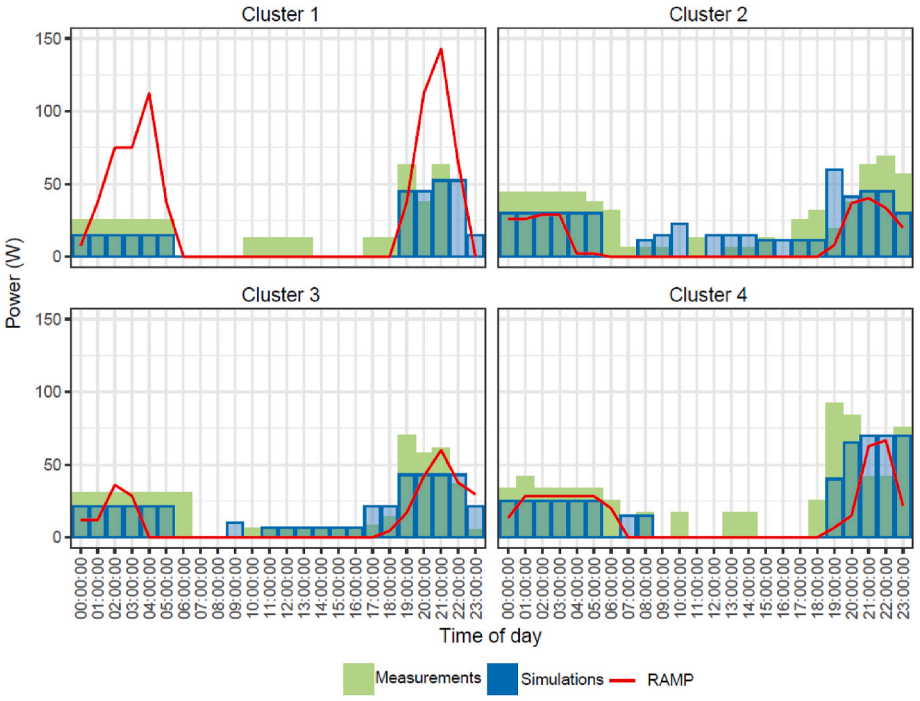


**Fig. 11.** Results from the comparison of mean lighting load profiles from the measurements of 13 HH, the simulations, and from RAMP model, by cluster.

In Hartvigsson et al. (2018), the inability of interview-based load profiles for identifying the daytime hours of electricity use due to bias in the respondents' gender is discussed. During the data-collection in Kenya, this was considered on varying the interview hours and days (including weekends) to achieve a 50:50 proportion of male and female respondents. In the measured lighting load profiles, a slight usage of lighting during the daytime is registered, especially for households from clusters 2 and 3. These loads are not identified by RAMP, while they are identified by the simulated profiles. It is important to highlight that RAMP is based on user-entered time frame information such as the time of use per day of each appliance (in this case, lamps), minimum time in which these are kept on once they are switched-on, and time frames in which random switch-ons can occur. While the first parameter was included in the survey-collected data, the other two were not; therefore, the standard parameters set in the open-source model – that were also applied in the case study in Lombardi et al. (2019) for Bolivia were used.

From Table 11 it is observed that for predicting the real LF, generally the simulations outperform the results from RAMP in all clusters with errors ranging from −1.54 to −7.83%, except for cluster 3 with 33.21%. However, both models underestimate the peak load for all clusters

**Table 11**
Results of comparison by cluster.

| | Real | Sim | % Err Sim | RAMP | % Err RAMP | Survey | % Err Survey |
|---|---|---|---|---|---|---|---|
| Cluster 1 | | | | | | | |
| Load Factor | 0.26 | 0.24 | −7.83 | 0.20 | −20.79 | | |
| Peak Load (W) | 62.50 | 52.50 | −16.00 | 143.00 | 128.80 | | |
| Aggregate Consumption (Wh/day) | 387.50 | 300.00 | −22.58 | 702.25 | 81.23 | 1444.00 | 272.65 |
| NRMSE (%) | | 0.89 | | 2.23 | | | |
| Cluster 2 | | | | | | | |
| Load Factor | 0.39 | 0.38 | −3.88 | 0.26 | −32.84 | | |
| Peak Load (W) | 68.75 | 60.00 | −12.73 | 40.25 | −41.45 | | |
| Aggregate Consumption (Wh/day) | 643.75 | 540.00 | −16.12 | 253.13 | −60.68 | 357.75 | −44.43 |
| NRMSE (%) | | 0.66 | | 0.77 | | | |
| Cluster 3 | | | | | | | |
| Load Factor | 0.30 | 0.40 | 33.21 | 0.19 | −35.54 | | |
| Peak Load (W) | 70.00 | 43.00 | −38.57 | 60.00 | −14.29 | | |
| Aggregate Consumption (Wh/day) | 505.00 | 413.25 | −18.17 | 279.00 | −44.75 | 663.00 | 31.29 |
| NRMSE (%) | | 0.56 | | 0.84 | | | |
| Cluster 4 | | | | | | | |
| Load Factor | 0.30 | 0.29 | −1.54 | 0.22 | −27.39 | | |
| Peak Load (W) | 91.67 | 70.00 | −23.64 | 66.67 | −27.27 | | |
| Aggregate Consumption (Wh/day) | 658.33 | 495.00 | −24.81 | 347.67 | −47.19 | 950.00 | 44.30 |
| NRMSE (%) | | 0.65 | | 1.01 | | | |

(except for RAMP on cluster 1), generally, the simulations show a better performance, except again for cluster 3, in which RAMP has an error of −14.29%. For the aggregate consumption, the survey method was also included for comparison, and it is interesting to note that the simulations outperform the surveys and RAMP on this indicator, with an average error of −20.42%. In the case of RAMP it is mainly due to the unidentified daytime loads; as for the surveys, they tend to overestimate the consumption in most cases, except for cluster 2. The NRMSE indicator evaluates the performance of both models considering the hourly estimations, for this, the simulations and RAMP presented an average performance below 2.5%, still the simulations obtained the lowest errors, ranging from 0.56 to 0.89%. It is important to note that the performance for RAMP could have been improved if the specific required inputs would have been included in the field data collection; meaning that RAMP is useful to apply when field data are collected from specific sites being considered for electrification projects.

### 5.7. Geospatial representation of lighting consumption

Kenya was selected as an example of how the methodology can be applied to geospatially project the lighting demand of rural households using publicly available data. The variables needed as input for the predictive models were extracted from each household in Dataset 2 for computing their typical load profiles and then calculating their monthly aggregate lighting consumption (kWh/month). This dataset was used because households are geo-referenced, which allows mapping their final results. It also contains information about their actual monthly electricity consumption (accounting for the total consumption, not only lighting), which allows its comparison with the simulated lighting consumption results. In Fig. 12, the top left figure presents the results of the simulated monthly lighting consumption, while the top right shows the actual monthly electricity consumption reported in the surveys. Overall, it was found that the average share of lighting in their total electricity consumption corresponds to 40.85%, with an average aggregate lighting consumption of 380.71 Wh/day (11.42 kWh/month) per household; however patterns are identified by province and by county. Previous research has explored the usage of geospatial data and application on economic development (Goldblatt et al., 2019). More widely, it has been studied the correlation between nighttime lights and economic activity (Ishizawa et al., 2017; Mellander et al., 2015), as economic development could be correlated to electricity access or even electricity consumption (Lee et al., 2020), the nighttime lights were included for further analysis. In the bottom right, a distribution of the identified clusters of households per county and province is presented, while in Fig. 13, the average shares of lighting in the total electricity consumption of a typical household per province and county are identified (see Fig. A.8 in the appendix for reference on the counties' location). For further discussion on the correlation of the type of lamp used and the households' income, Fig. 14 shows the average ownership of each of the analyzed type of lamps, against the average households' monthly income per province. The households' monthly income was computed by aggregating each household member's income, averaging it by province. From this figure, the correlation between the average households' monthly income and the average ownership of incandescent and LED lamps has the strongest negative and positive trends, respectively. However, this correlation is not evident for fluorescent lamps. For the CFL lamps, a slight negative trend is found.

An average household in Central, Eastern, and Rift Valley provinces consumes more electricity and lighting, and has the highest ownership of incandescent lights compared to other provinces. These provinces are among the ones with the largest share of households belonging to cluster 4. According to the real measured data from Table 11, households in cluster 4 are the largest lighting consumers, with an average value of 658.33 Wh/day. Lighting covers 27.76%, 40.11%, and 47.36% of the total electricity consumption of households in these provinces (Fig. 12). Some of the highest values of average monthly radiance are also found there (Fig. 12, bottom left), interestingly, the average monthly income of households included in this study from Central and Eastern provinces is the lowest (Fig. 14); therefore, no evident correlation was found between the monthly average radiance and the average household income.

Western province has the largest share of households belonging to cluster 3 compared to the others. Fig. 14 shows that the lighting share of the total electricity consumption of households in this province is 79%, although in Fig. 14 they presented to have one of the largest share of ownership of LED lamps with 59.52%, following the Coast and North-Eastern provinces. All sampled households in the North-Eastern province own LED; lamps, however, it is important to note that they only represent one county (Wajir). Analyzing the results at a county level, it is interesting to note the high shares of the total electricity consumption that is allocated to lighting in Kakamega, Busia and Vihiga in the Western province. According to the results, these three counties along with Embu (Eastern province) and West Pokot, Elgeyo-Marakwet and Baringo (Rift Valley) are the ones in which most rural households mainly use electricity for lighting purposes.

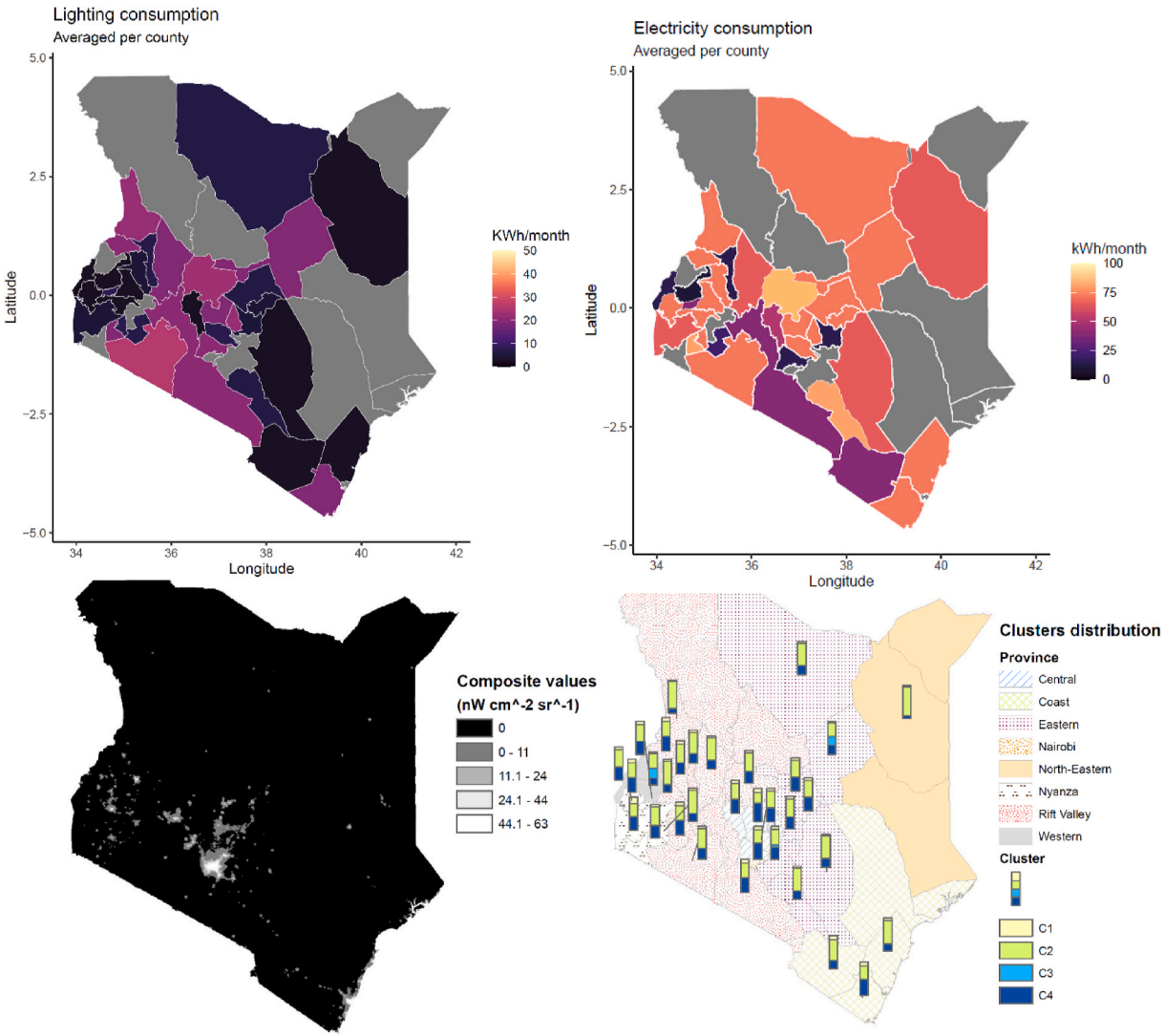


**Fig. 12.** Top left, geospatial representation of the model's results with predictions of the monthly lighting consumption, averaged per county. Top right, geospatial representation of the actual electricity consumption documented in Dataset 2, averaged per county. Bottom left, nighttime lights extract for Kenya, Dataset 5. Bottom right, distribution of households belonging to each cluster per province. The sub-county representation for the top figures is found in Fig. A7 in the Appendix.

## 6. Study boundaries and limitations

The overall limitations of this study include that, as a data-driven methodology, it is highly dependent on the quality and amount of data used for creating the models. Even if the average accuracy of all predictive models is 80%, further developments for increasing their robustness by adding more samples in the training dataset are required. The power ratings' definition for each type of lamp was made based on the assumptions specified in Table 5. However, the luminous efficacy and the power rating for each lamp type may vary from wider ranges than the ones proposed in this paper, and the arbitrary selection of a final value from the proposed ranges may act only as an approximation. For future work, the methodology can be improved by developing a calibration method to account for the activity variation during the weekdays and weekends, as well as for the seasonal variation. In addition, more measured data should be collected for validation.

## 7. Conclusion

Creating data-driven methods to support the planning of energy access solutions in developing countries is of great importance mainly for optimizing the investment of resources for future electrification projects. The proposed methodology was created using publicly available data from two of the most representative countries of East Africa, Kenya and Tanzania, applying machine learning approaches for determining occupant behavior patterns, lighting requirements, lamps type and ownership to compute hourly lighting profiles for rural households. Its application was validated by modeling the profiles of 13 measured households in rural Kenya, obtaining an average normalized root mean squared error of 0.7%, which is 0.6% less compared to existing simulation approaches based on on-site data collection. The model generally underestimated other indicators, such as the peak power and the aggregate consumption; however, the average errors were relatively low compared to existing approaches. The inconsistency on the information provided by households through surveys and their real lighting consumption measurements can be perceived, affecting specially the performance of the existing approaches, as they depend on these data as input. The latter approaches are suitable when field data are collected from specific sites. Furthermore, data collection implies a high amount of resource investment for project developers. To facilitate this task, the proposed methodology uses only publicly available data as input, thus it can be applied in places where specific information is not available. However, it is important to note that even if the simulated approximations are made considering different household-level attributes as

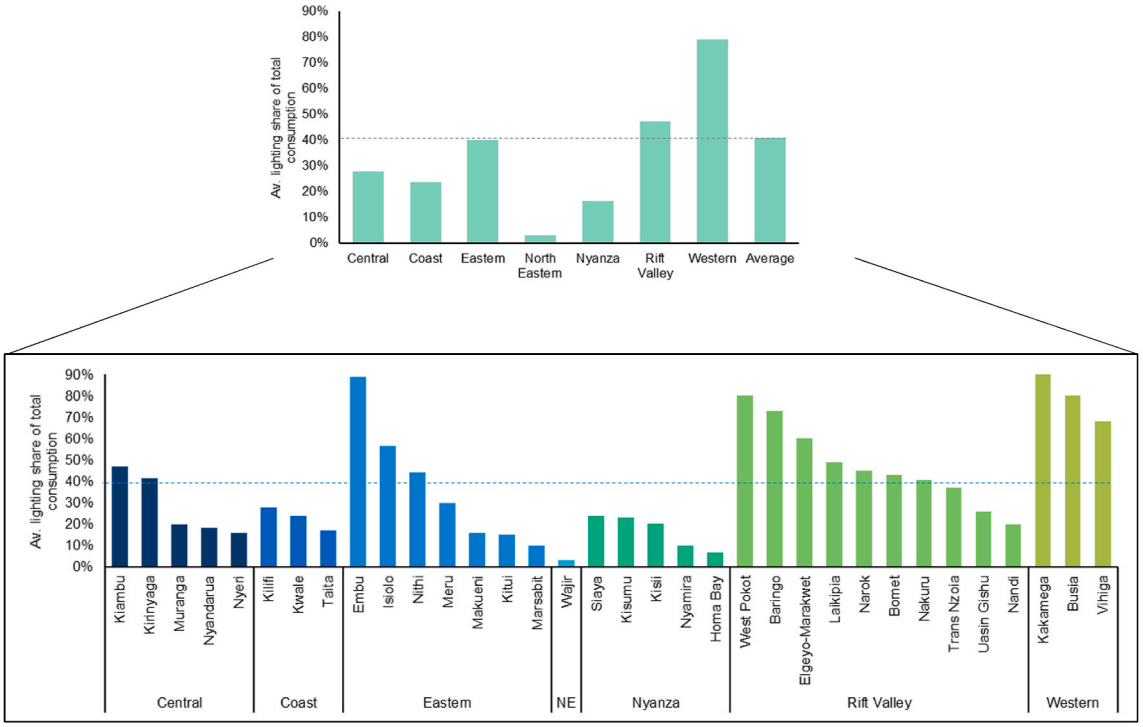


**Fig. 13.** Average lighting share of the total electricity consumption (%) per province.

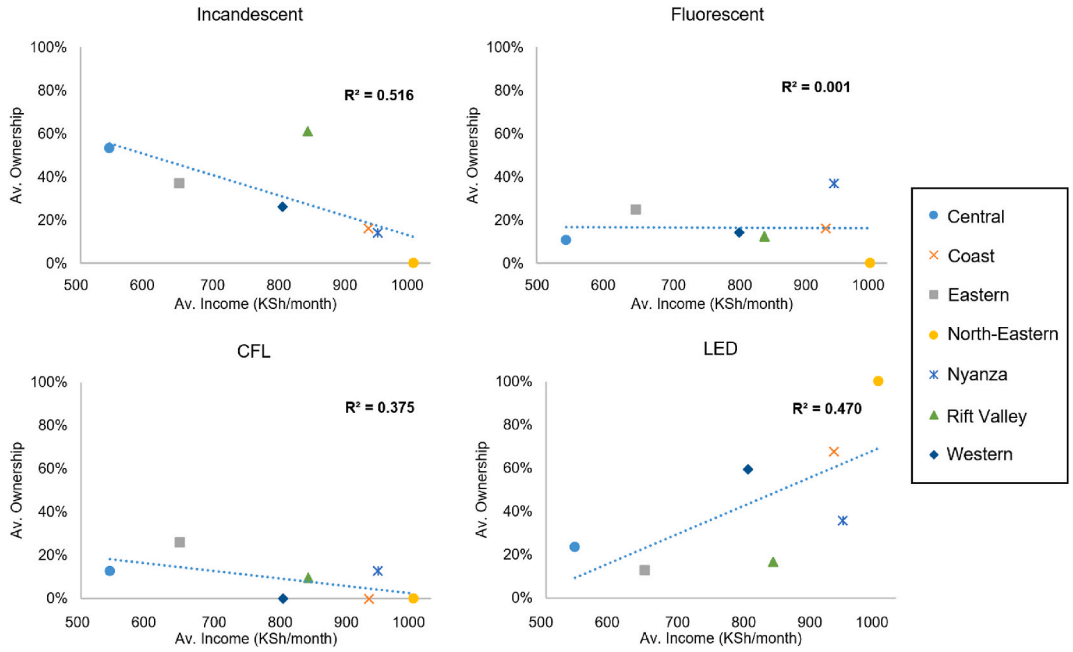


**Fig. 14.** Average ownership of lamps (% of households) against the average monthly income (KSh), for each type of lamp and each province.

indicators, this does not mean that computer-generated data will replace the qualitative data that can be collected on-site.

The main findings of the study include that rural households in East Africa can be characterized into four different occupant behavior patterns: the ones that mostly perform farming activities, those that perform domestic work, those that own a business at home, and finally, those that have salaried jobs. These are defined considering not only the household member's activity diaries, but also different socioeconomic and demographic attributes. Predictive models for classifying households based on their behavioral patterns, the type of lamps they own, and the number of lamps used for indoor and outdoor purposes were created, with average accuracy of prediction of 80% on the testing sets. The overall input variables for the proposed methodology range from demographic, socioeconomic and geographic attributes at household and village/sub-county levels, including the household size, housing materials, livestock and other assets ownership, nighttime lights of specific locations, village access to grid-electricity, among others. Interestingly, the nighttime lights extracted from satellite imagery showed having an impact for defining the type of lamps that households own. Thus, stronger average monthly radiance has a positive

contribution on the use of fluorescent and CFL lamps; while it affects negatively the use of incandescent and LED lamps. The village and sub-county level variables were found having more impact for defining the number of indoor and outdoor lights owned by households, rather than for the type of lamps they own. These variables are the sub-county access to streetlights and population density, and village access to grid-electricity. A geospatial characterization for Kenya is also presented as an example of the application of this methodology for geographically identifying sites with the highest lighting consumption using publicly available data. From this, it is identified that households in the provinces of Western, Rift Valley and Eastern use an average of 79%, 47% and 40%, respectively, of their total electricity consumption to meet their lighting needs.

Finally, this paper introduces a methodology for characterizing rural households based on their occupant behavior and predicting their detailed lighting devices ownership in terms of type of lamp used, and amount owned for indoor and outdoor purposes. From this, their potential energy usage dedicated only to lighting can be identified. The results obtained in this study can serve as reference to project developers or solar home system distributors of the amount of the electricity consumption that rural households allocate only to meet their lighting needs, which bridges a knowledge gap in the literature that contributes to a better understanding of their electricity consumption habits. The methodology can be potentially applied for performing pre-feasibility studies, as most existing load simulation approaches require specific on-site collected input data in order to get estimations of what would be the latent energy demand (energy that households may consume when given electricity access). The required inputs for this methodology can be easily accessed from public databases, which increases the possibilities of applying and validating the methodology in other countries in the region.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the ETH Grant [ETH-10 16-2], under the project: “Forecasting rural electricity usage profiles for developing regions”. We would like to thank the anonymous reviewers for their valuable comments.

## Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.deveng.2021.100073.

## Appendix A

**Table A.1**
Variables tested in linear discriminant analysis (LDA) for selecting the most significant variables for the clusters definition. The criterion used for selection was a relative significant t-value and a p-value equal of less than 0.001. The selected variables are in bold.

| Variable | t-value | | | p-value | | |
|---|---|---|---|---|---|---|
| | Pseudo F | Pseudo R2 | Levene | Pseudo F | Pseudo R2 | Levene |
| Demographics/economic activities | | | | | | |
| Household Size | 1.2E+00 | 3.3E-02 | 3.7E+00 | 1.6E-02 | 1.6E-02 | 4.8E-03 |
| Respondent gender* | 2.0E+01 | 4.2E-02 | 2.7E+01 | 2.0E-04 | 2.0E-04 | 2.0E-04 |
| Age | 1.1E+00 | 1.3E-01 | 4.2E+00 | 3.6E-02 | 3.5E-02 | 2.0E-04 |
| Relationship to household head* | 5.3E+00 | 3.4E-02 | 1.7E+01 | 2.0E-04 | 2.0E-04 | 4.0E-04 |
| Food farming | 8.2E+00 | 1.8E-02 | 3.1E+00 | 2.0E-04 | 2.0E-04 | 7.5E-02 |
| Cash farming | 5.3E+00 | 1.1E-02 | 1.4E+01 | 2.0E-04 | 2.0E-04 | 4.0E-04 |
| Livestock raising | 6.2E+00 | 1.3E-02 | 1.3E+01 | 2.0E-04 | 2.0E-04 | 4.0E-04 |
| Non farming activities | 5.0E+00 | 1.1E-02 | 1.2E-01 | 2.0E-04 | 2.0E-04 | 7.4E-01 |
| Salaried job | 4.6E+00 | 9.9E-03 | 6.7E-03 | 2.0E-04 | 2.0E-04 | 9.3E-01 |
| Fishing | 3.7E+00 | 7.9E-03 | 4.3E+00 | 2.0E-04 | 2.0E-04 | 4.0E-02 |
| Housing | | | | | | |
| Number of rooms | 1.1E+00 | 1.2E-02 | 5.3E+00 | 2.3E-01 | 2.1E-01 | 3.2E-03 |
| Walls material | 1.4E+00 | 1.8E-02 | 3.8E+00 | 1.4E-03 | 8.0E-04 | 7.6E-03 |
| Roof material | 1.0E+00 | 9.1E-03 | 8.4E+00 | 2.9E-01 | 3.0E-01 | 3.2E-03 |
| Floor material | 1.8E+00 | 4.0E-03 | 5.0E+00 | 9.4E-03 | 8.2E-03 | 2.6E-02 |
| Cooking fuel | 1.3E+00 | 1.2E-02 | 7.6E+00 | 1.1E-02 | 1.2E-02 | 1.3E-02 |
| Ownership of productive capital | | | | | | |
| Land | 6.3E+00 | 1.4E-02 | 5.5E+00 | 2.0E-04 | 2.0E-04 | 2.5E-02 |
| Other land | 1.3E+00 | 2.9E-03 | 3.6E-01 | 9.3E-02 | 1.0E-01 | 5.6E-01 |
| Large livestock[a] | 4.6E+00 | 1.0E-02 | 7.6E+00 | 2.0E-04 | 2.0E-04 | 6.0E-03 |
| Small livestock[b] | 4.7E+00 | 1.0E-02 | 1.9E+01 | 2.0E-04 | 2.0E-04 | 2.0E-04 |
| Poultry | 4.1E+00 | 8.9E-03 | 1.7E+01 | 2.0E-04 | 2.0E-04 | 2.0E-04 |
| Fish | 2.3E+00 | 4.9E-03 | 2.8E-01 | 2.0E-03 | 1.0E-03 | 5.9E-01 |
| Non-mechanized farm equipment[c] | 6.3E+00 | 1.3E-02 | 4.0E+00 | 2.0E-04 | 2.0E-04 | 4.3E-02 |
| Mechanized farm equipment[d] | 1.3E+00 | 2.8E-03 | 2.3E+00 | 1.4E-01 | 1.2E-01 | 1.2E-01 |
| Business equipment[e] | 1.9E+00 | 4.0E-03 | 3.9E+00 | 6.8E-03 | 8.0E-03 | 4.9E-02 |
| House | 2.5E+00 | 5.4E-03 | 2.5E-01 | 4.0E-04 | 2.0E-04 | 6.1E-01 |
| Large appliances[f] | 3.8E+00 | 8.2E-03 | 2.1E+01 | 2.0E-04 | 2.0E-04 | 2.0E-04 |
| Small appliances[g] | 1.0E+00 | 2.2E-03 | 1.9E-01 | 4.0E-01 | 4.1E-01 | 6.6E-01 |
| Mobile phone | 3.1E+00 | 6.6E-03 | 1.1E+01 | 2.0E-04 | 2.0E-04 | 1.4E-03 |
| Transport means[h] | 1.5E+00 | 3.2E-03 | 1.3E+00 | 5.6E-02 | 5.7E-02 | 2.7E-01 |

Note: *Information from respondent, the rest of information addresses the household as a whole.

[a] Oxen, cattle.
[b] Goats, pigs, sheep.
[c] Hand tools.
[d] Tractor plough, power tiller, treadle pump.
[e] Solar panels used for recharging, sewing machine, brewing equipment, fryers.
[f] Televisions, refrigerators, DVD.
[g] Radios, mobile phones.
[h] Bicycle, motorcycle, car.

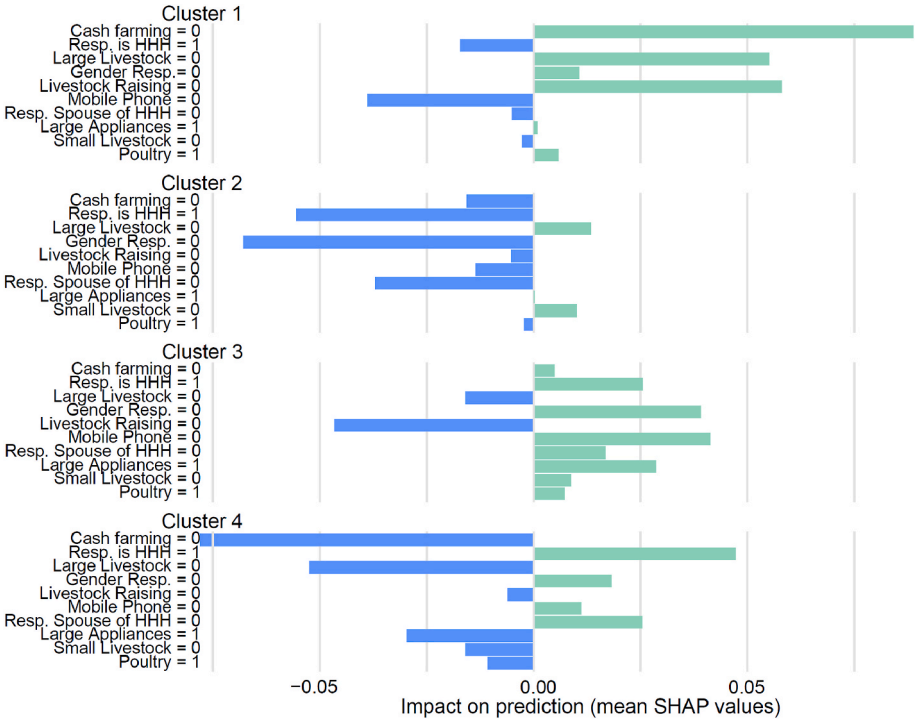


**Fig. A.1.** Impact of each variable on clusters classification prediction, expressed in mean SHAP values. The negative values represent a negative impact of the variable value. A value of '0' indicates negation, while '1' affirmation, e.g. owning poultry has a negative impact for clusters 1 and 3, and positive for clusters 2 and 4.

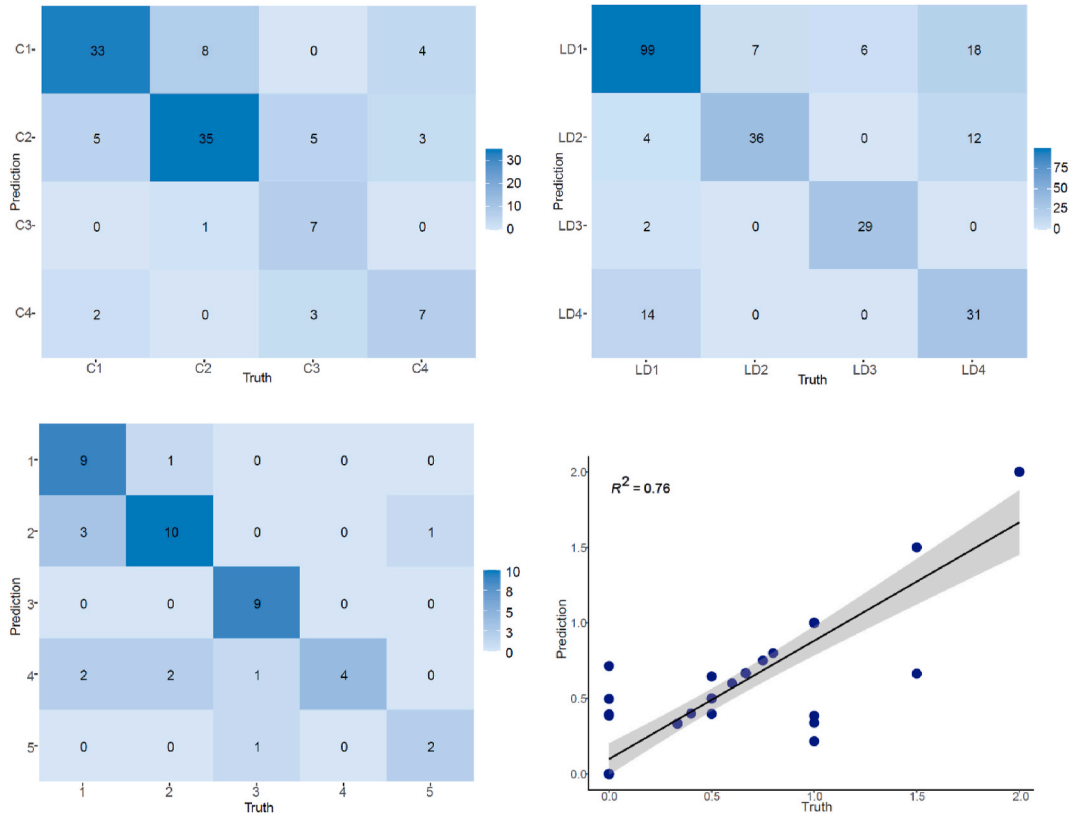


**Fig. A.2.** Confusion matrices and linear relation. The matrices' columns represent the true class households belonging to each cluster, while the rows show the households that were predicted to belong to each cluster. The sum of the diagonal values are the households that were predicted correctly.

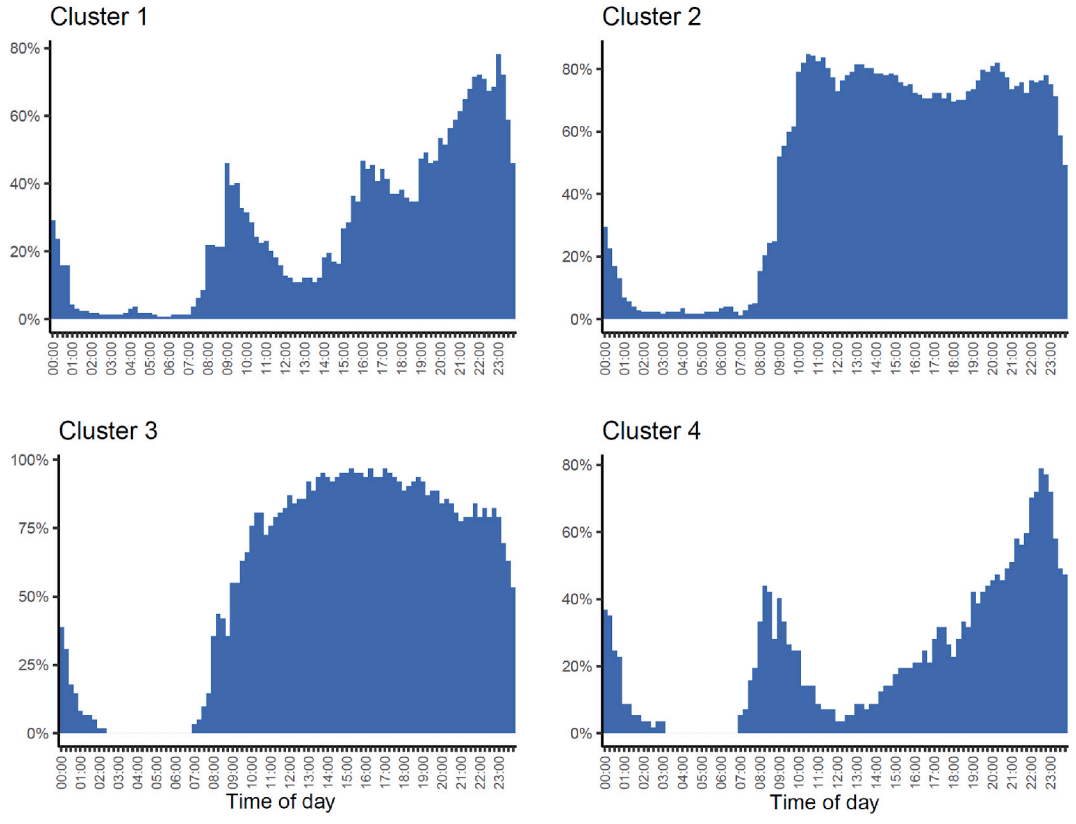


**Fig. A.3.** Activity profiles, the y-axis represents the probability of activity at home that requires lighting, the x-axis represents the time of the day.

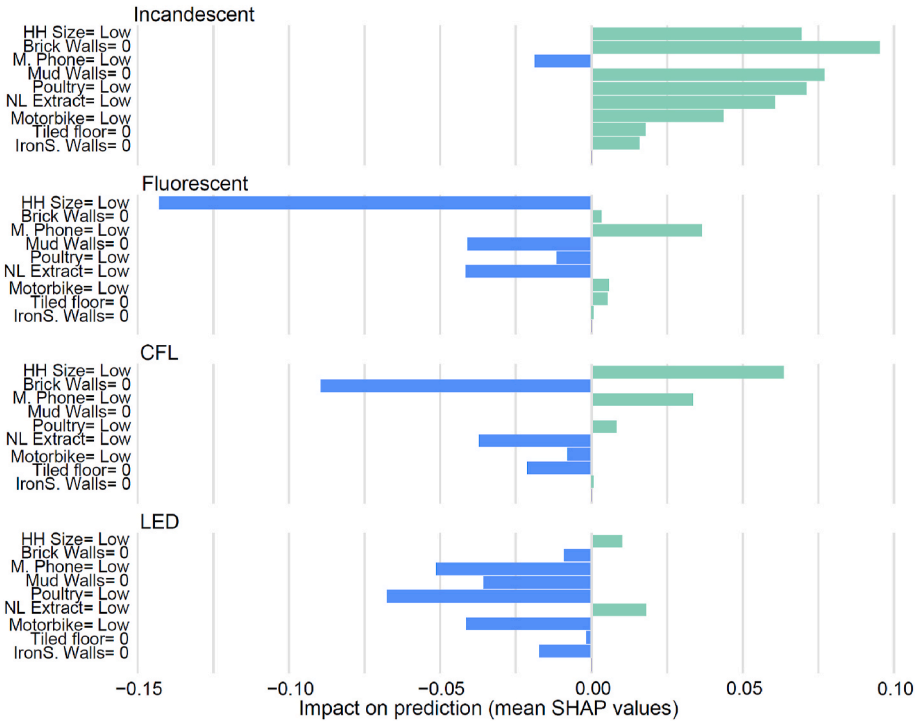


**Fig. A.4.** Impact of each variable on type of lamp prediction, expressed in mean SHAP values. The values indicated as "Low" or "High" represent values that are located either lower on higher than the mean of the respective variable in the dataset. For walls and floor materials, '0' indicates negation, while '1' affirmation.

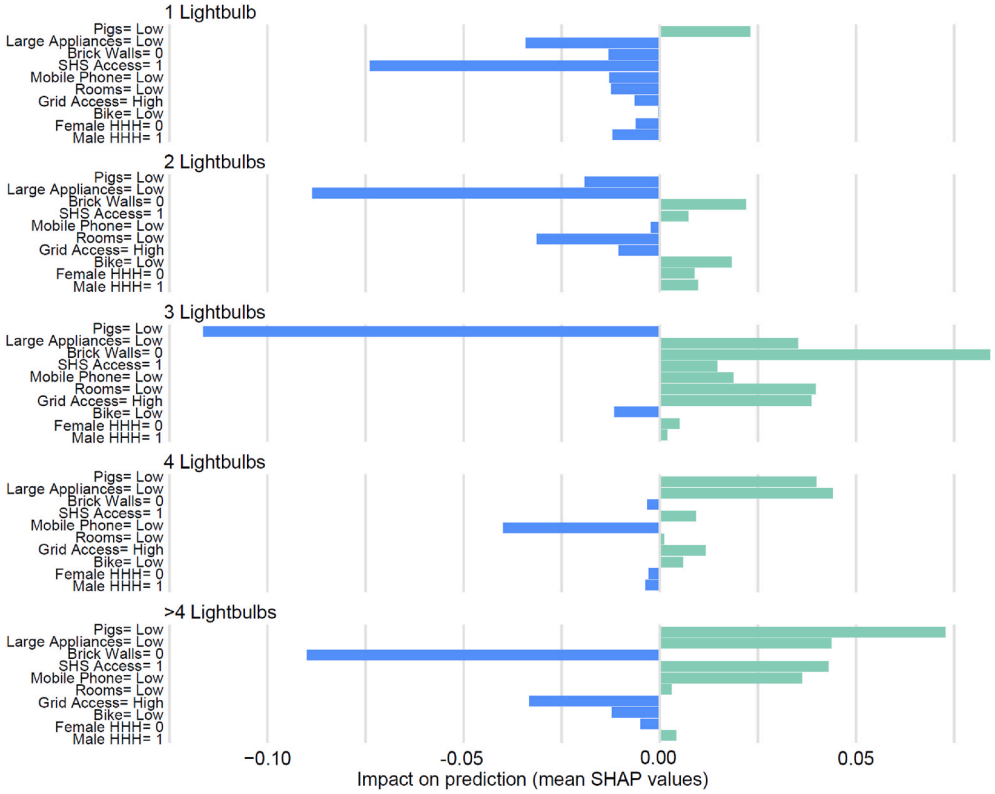


**Fig. A.5.** Impact of each variable on type of lamp prediction, expressed in mean SHAP values. The values indicated as "Low" or "High" represent values that are located either lower on higher than the mean of the respective variable in the dataset. For walls materials, SHS Access, Female and Male household head (HHH), '0' indicates negation, while '1' affirmation.

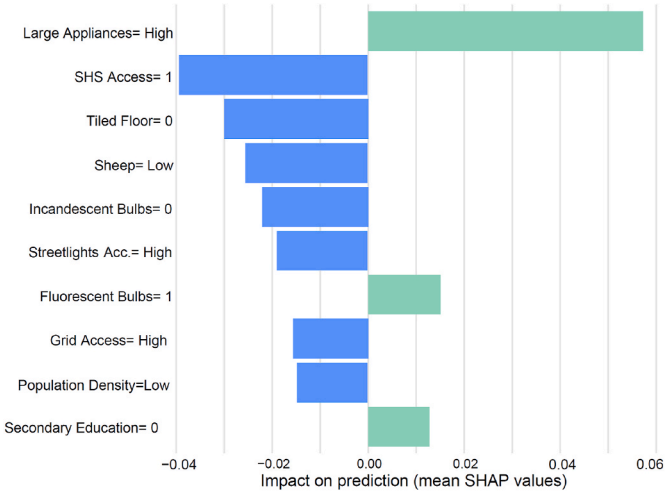


**Fig. A.6.** Impact of each variable on outdoor lamps ownership prediction, expressed in mean SHAP values. The values indicated as "Low" or "High" represent values that are located either lower on higher than the mean of the respective variable in the dataset. For floor material, SHS Access, Incandescent, Fluorescent, and Secondary Education, '0' indicates negation, while '1' affirmation.

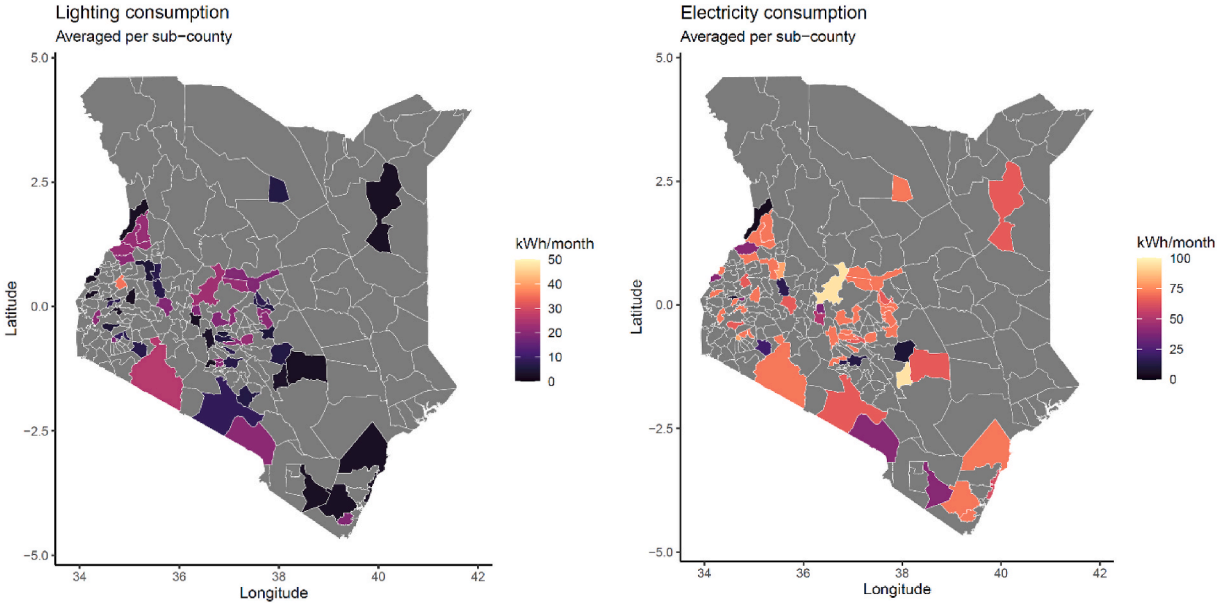


**Fig. A.7.** Left, geospatial representation of the model's results with predictions of the monthly lighting consumption, averaged per sub-county. Right, geospatial representation of the actual electricity consumption documented in Dataset 2, averaged per sub-county.

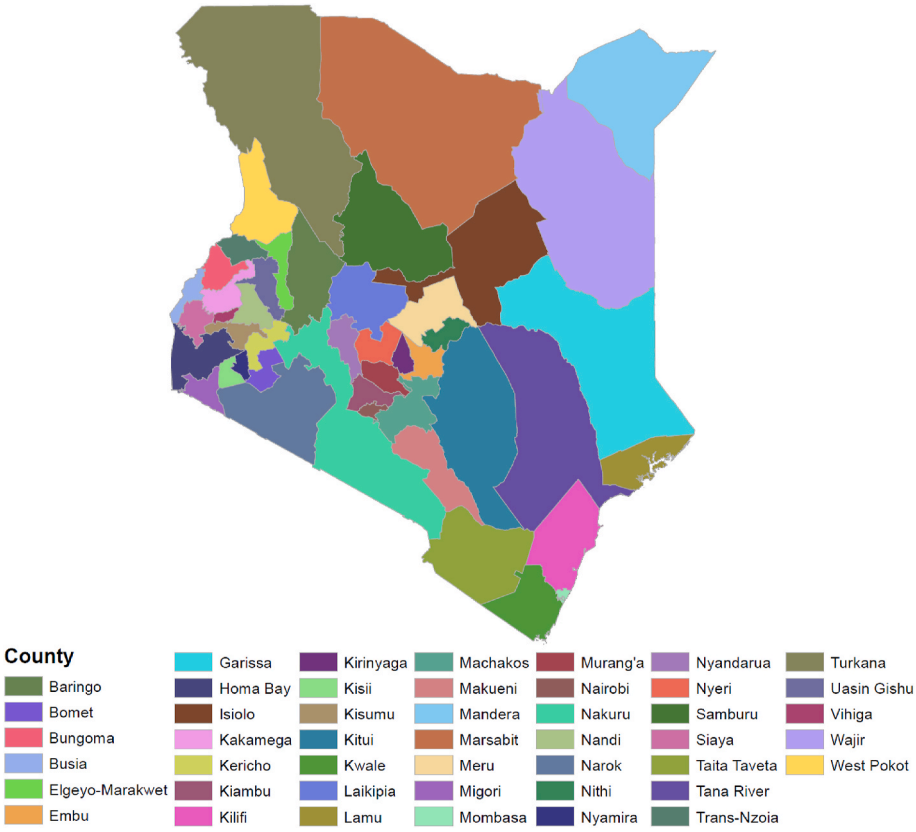


**Fig. A.8.** Administrative areas of Kenyan counties.

## References


Adeoti, O., B, O., 2011. Solar photovoltaic-based home electrification system for rural development in Nigeria: domestic load assessment. Renew. Energy 24 (1), 155–161. https://doi.org/10.1016/S0960-1481(00)00188-9.

Ahlborg, H., Sjostedt, M., 2015. Small-scale hydropower in Africa: socio-technical designs for renewable energy in Tanzanian villages. Energy Research & Social Science 5, 20–33. https://doi.org/10.1016/j.erss.2014.12.017.

Blodgett, C., Dauenhauer, P., Louie, H., Kickmam, L., 2017. Accuracy of energy-use surveys in predicting rural mini-grid user consumption. Energy Sustain. Develop. 41, 88–105. https://doi.org/10.1016/j.esd.2017.08.002.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3 (1), 1–27. https://doi.org/10.1080/03610927408827101.

Deshmukh, R., J, P., 2013. Sustainable Development of Renewable Energy Mini-Grids for Energy Access: a Framework for Policy Design. Lawrence Berkeley National Laboratory, Berkeley.

Dobler, G., Ghandehari, M., Koonin, S.E., Sharma, M.S., 2016. A hyperspectral survey of New York city lighting technology, 2047 Sensors 16 (12), 2047. https://doi.org/10.3390/S16122047, 16, 2016.

Dominguez, C., Orehounig, K., Carmeliet, J., 2018. One step further for electrifying rural households in Africa: a novel electricity demand modelling approach. In: Proceedings of the Building Simulation and Optimization Conference (BSO), Cambridge University, United Kingdom.

Dominguez, C., Orehounig, K., Carmeliet, J., 2020. Energy Consumption Patterns of Rural Households in Kenya. https://doi.org/10.3929/ethz-b-000455732.

Dominguez, C., Orehounig, K., Carmeliet, J., 2021. Understanding the path towards a clean energy transition and post-electrification patterns of rural households. Energy Sustain. Develop. 61, 46–64. https://doi.org/10.1016/j.esd.2021.01.002.

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. VIIRS night-time lights. Int. J. Rem. Sens. 38, 5860–5879. https://doi.org/10.1080/01431161.2017.1342050.

Elvidge, C.D., Zhizhin, M., Ghosh, T., Hsu, F.-C., Taneja, J., 2021. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. Rem. Sens. 13 (5) https://doi.org/10.3390/rs13050922.

Energy Sector Management Assistance Program (ESMAP), 2018. World Bank. Multi-tier framework for measuring energy access - Kenya household survey. Ref. KEN_2016-2018_MTF_v02_M. https://microdata.worldbank.org/index.php/catalog/3703/study-description.

Falchetta, G., Stevanato, N., Moner-Girona, M., Mazzoni, D., Colombo, E., Hafner, M., 2020. M-LED: multi-sectoral latent electricity demand assessment for energy access planning. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3700858.

Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M., 2011. Analyzing and visualizing state sequences in R with TraMineR. J. Stat. Software 40 (4), 1–37. https://doi.org/10.18637/jss.v040.i04.

Gastwirdth, J., Gel, Y., Miao, W., 2009. The impact of Levene's test of equality of variances on statistical theory and practice. Stat. Sci. 24 (3), 343–360. https://doi.org/10.1214/09-STS301.

Goldblatt, R., Jones, M., Bottoms, B., 2019. Geospatial data for research on economic development. Develop. Eng. 4, 100041. https://doi.org/10.1016/j.deveng.2019.100041.

Hartvigsson, E., Ahlgren, E., 2018. Comparison of load profiles in a mini-grid: assessment of performance metrics using measured and interview-based data. Energy Sustain. Develop. 43, 186–195. https://doi.org/10.1016/j.esd.2018.01.009.

IEA, IRENA, UNSD, World Bank, WHO, 2021. Tracking SDG 7: the Energy Progress Report. World Bank, Washington DC.

IEA. SDG7. Data and projections. https://www.iea.org/reports/sdg7-data-and-projections/access-to-electricity. (Accessed 7 September 2021).

Ishizawa, O., Miranda, J., Zhang, H., 2017. Understanding the Impact of Windstorms on Economic Activity from Night Lights in Central America. Policy Research Working Paper. Social, Urban, Rural and Resilience Global Practice Group. World Bank. http://documents1.worldbank.org/curated/en/338881498577038860/pdf/WPS8124.pdf.

Kumar, A., Choudhury, R., 2014. Characteristics of Light Sources. Object Appearance, Colour Perception and Instrumental Measurement, Principles of Colour and Appearance Measurement. Elsevier, pp. 1–52. https://doi.org/10.1533/9780857099242.1.

Lee, K., Miguel, E., Wolfram, C., 2020. Does household electrification supercharge economic development? J. Econ. Perspect. 34 (1), 122–144. https://doi.org/10.1257/jep.34.1.122.

Levin, N., Kyba, C.C.M., Zhang, Q., Sánchez de Miguel, A., Román, M.O., Li, X., Portnov, B.A., Molthan, A.L., Jechow, A., Miller, S.D., Wang, Z., Shrestha, R.M., Elvidge, C.D., 2020. Remote sensing of night lights: a review and an outlook for the future. Remote Sensing Environ. 237, 111443. https://doi.org/10.1016/J.RSE.2019.111443.

EnergyPlus. 2020. Weather data. Solar and Wind Energy Resource Assessment (SWERA). https://energyplus.net/weather, 3 September 2020.

Last Mile Connectivity Project (LMCP). African Development Bank Group. [https://www.afdb.org/] (accessed 2 February 2020).

Lombardi, F., Balderrama, S., Quoilin, S., Colombo, E., 2019. Generating high-resolution multi-energy load profiles for remote areas with an open-source stochastic model. Energy 177, 433–444. https://doi.org/10.1016/j.energy.2019.04.097.

Mahapatra, S., Chanakya, H.N., Dasappa, S., 2009. Evaluation of various energy devices for domestic lighting in India: technology, economics and CO2 emissions. Energy Sustain. Develop. 13 (4), 271–279. https://doi.org/10.1016/j.esd.2009.10.005.

Mandelli, S., Brivio, C., Colombo, E., Merlo, M., 2016a. Effect of load profile uncertainty on the optimum sizing of off-grid PV systems for rural electrification. Sustain. Energy Technol. Assessment. 18, 34–47. https://doi.org/10.1016/j.seta.2016.09.010.

Mandelli, S., Merlo, M., Colombo, E., 2016b. Novel procedure to formulate load profiles for off-grid rural areas. Energy Sustain. Develop. 31, 130–142. https://doi.org/10.1016/j.esd.2016.01.005.

McNeil, M.A., V, E., 2010. Modeling diffusion of electrical appliances in the residential sector. Energy Build. 42 (6), 783–790. https://doi.org/10.1016/j.enbuild.2009.11.015.

Mellander, C., Lobo, J., Stolarick, K., Matheson, Z., 2015. Night-time light data: a good proxy measure for economic activity? PLoS One 10 (10), e0139779. https://doi.org/10.1371/journal.pone.0139779.

Moner-Girona, M., Bódis, K., Morrissey, J., Kougias, I., Hankins, M., Huld, T., Szabó, S., 2019. Decentralized rural electrification in Kenya: speeding up universal energy access. Energy Sustain. Develop. 52, 128–146. https://doi.org/10.1016/j.esd.2019.07.009.

Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31, 274–295. https://doi.org/10.1007/s00357-014-9161-z.

Nieuwenhout, F.D.J., van de Rijt, P.J.N.M., Wiggelinkhuizen, E.J., van der Plas, R.J., 1998. Rural Lighting Services: a Comparison of Lamps for Domestic Lighting in Developing Countries. World Bank. http://documents1.worldbank.org/curated/en/136981468779966908/pdf/268550EnergyIssues112.pdf.

Philips. Product Catalog, 2021. https://www.lighting.philips.com/main/prof. (Accessed 7 February 2021).

Rural Electrification and Renewable Energy Corporation (REREC). https://www.rerec.co.ke/. (Accessed 7 February 2021).

Rom, A., Günther, I., Borofsky, Y., 2020. Using sensors to measure technology adoption in the social sciences. Develop. Eng. 5, 100056 doi: 1016/j.deveng.2020.100056.

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Sánchez de Miguel, A., Kyba, C.C.M., Aubé, M., Zamorano, J., Cardiel, N., Tapia, C., Bennie, J., Gaston, K.J., 2019. Colour remote sensing of the impact of artificial light at night (I): the potential of the International Space Station and other DSLR-based platforms. Remote Sensing Environ. 224, 92–103. https://doi.org/10.1016/J.RSE.2019.01.035.

Sebitosi, A.B., Pillay, P., 2007. New technologies for rural lighting in developing countries: white LEDs. IEEE Trans. Energy Convers. 22 (3), 674–679. https://doi.org/10.1109/TEC.2006.888024.

Smith, T., McKenna, C., 2013. A comparison of logistic regression Pseudo R2 indices. Multiple Linear Regression Viewpoints 39 (2), 17–26.

Stokes, M., Rylatt, M., Lomas, K., 2004. A simple model of domestic lighting demand. Energy Build. 36 (2), 103–116. https://doi.org/10.1016/j.enbuild.2003.10.007.

Studer, M., Ritschard, 2016. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J. Roy. Stat. Soc. 179 (2), 481–511.

United Nations Development Programme (UNDP), 2020. The 2020 Global Multidimensional Poverty Index (MPI). http://hdr.undp.org/en/2020-MPI. (Accessed 9 February 2021).

U.S. Department of Energy, 2021. Office of Energy Efficiency & Renewable Energy. https://www.energy.gov/energysaver/save-electricity-and-fuel/lighting-choices-save-you-money/how-energy-efficient-light. (Accessed 2 February 2021).

Van Ruijven, Bas, J., Schers, J., van Vuuren, D., 2011. Model-based scenarios for rural electrification in developing countries. Energy 386, 397. https://doi.org/10.1016/j.energy.2011.11.037.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Widén, J., Wäckelgård, E., 2010. A high-resolution stochastic model of domestic activity patterns and electricity demand. Appl. Energy 87 (6), 1880–1892. https://doi.org/10.1016/j.apenergy.2009.11.006.

Widén, J., Nilsson, A.M., Wäckelgård, E., 2009. A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. Energy Build. 41 (10), 1001–1012. https://doi.org/10.1016/j.enbuild.2009.05.002.

Williams, N., Jaramillo, P., Campbell, K., Musanga, B., Lyons-Galante, I., 2018. Electricity Consumption and Load Profile Segmentation Analysis for Rural Microgrid Customers in Tanzania. Proceedings of the IEEE PES/IAS PowerAfrica Conference, Cape Town, South Africa. https://doi.org/10.1109/PowerAfrica.2018.8521099.

World Bank, 2016. Tanzania National Bureau of Statistics. National Household Survey Panel (NHSP), 2015. http://econ.worldbank.org/.

World Bank. Country indicators. https://data.worldbank.org/indicator/. (Accessed 7 February 2020).

World Resources Institute (WRI). Accelerating mini-grid deployment in sub-saharan Africa: lessons from Tanzania. https://www.wri.org/publication/tanzania-mini-grids. (Accessed 10 September 2020).