Mapping Global Electricity: Identifying those that are left behind.

Table of Contents

1. Introduction5

2. Electricity Mapping7

3. High Resolution Electricity Access10

3.1. Input Data10

3.1.1. Nighttime Lights10

3.1.2. Settlements and Population10

3.1.3. Geographic Boundaries11

3.1.4. Land Cover11

3.1.5. Processed Data12

3.2. Data Processing13

3.2.1. Identifying Isolated Areas13

3.2.2. Outlier Removal13

3.2.3. Regression13

3.3. Results15

3.4. Validation16

4. Machine-learning Estimation of Electrification21

4.1. Input Data23

4.2. Data Processing24

4.3. Results28

4.4. Validation29

5. Cross Method Comparison32

6. Conclusion: Learning and Challenges36

6.1. Institutional Collaboration37

6.3. Lag39

6.3. Data Production, Storage and Visualization.39

6.4. Validation41

References43

Executive Summary

- What the problem is
- What we did
- What we learned: can provide where it is good to do where not to
-

1. Introduction

Electricity plays a key role in development as it fundamentally changes the development trajectory and range of possibilities for the poor. Evidence at the country level suggests substantial long run macroeconomic benefits of electrification (Stern et al 2019). There is, however, a need to enrich this analysis with the impact of electricity quality (reliability and infrastructure). Microeconomic analyses attest to the benefit of electrification in multiple domains, especially in the developing world (the following are taken from Lee et al 2017). These include positive impacts on health, poverty, and employment opportunities. Electricity access, for instance, is especially important for respiratory health as electrification can offset combustion-based heating and cooking indoors (Barron and Torero 2017). Electrification also has direct impacts on poverty and increases household welfare: multiple studies demonstrate gains in income and consumption especially for rural households in developing countries (Chakravorty et al. 2016, Khandker et al 2012, Khandker et al. 2014, Lipscomb et al 2013, Van de Walle et al 2017). Electrification is associated with increased employment, especially for women (Dinkelman 2011, Grogan and Sadanand 2013). nd, electricity access has been shown to improve education related outcomes (Hassan and Lucchino 2016).

The number of people worldwide with access to electricity has increased in recent decades: from an estimated 71% in the year 1990 to 87% with access in the year 2016 (Ritchie and Roser 2020). Globally, 759 million people lack access to electricity (IEA et al 2021). While the number of people without access to electricity has steadily declined, the rate is too slow to achieve SDG7 (Ritchie and Roser 2020).

It is critically important for social and economic development that initiatives to provide electricity access are ramped up. The UNDP has embarked on an ambitious plan to provide access to 500

million people by 2025: the UNDP's Energy Compact pledges to work with partners to provide access to clean and affordable energy to 500 million additional people, focusing on the most vulnerable communities. Many related policy efforts will result in reaching this milestone including building policy support, mobilizing funds and building supply. One key element to enable all of this will be targeting. In order to halve by 2030 those without access to electricity, it is important to know where they are – both locally (within countries) and globally (across countries and regions).

To help address this challenge, the UNDP has created an interactive global visualization of electricity access at the local level, that allows users to track progress on electrification over time. The visualization gives users an estimate of the level of electricity access instantaneously for any specified period and geography. Users can specify the level of spatial granularity i.e., the visualization provides data at the subnational and global levels. There is also a repository for electricity access data with which users can generate trends for electricity access over time, enabling them to track progress.

Mapping global electricity access is challenging and there are multiple ways to approach it. The objective of this report is to review different methodologies to map electricity access, describe global efforts to generate this data, and identify strengths and weaknesses of these approaches for policymaking. The report also describes in detail two approaches for mapping electricity access, and their results that will be important to target and monitor effectively the commitment to electrify 500 million people. Finally, it presents conclusions centered around lessons learnt from this work.

2. Electricity Mapping

Mapping global electricity access with high spatial and temporal resolution is a technically challenging task. In a world without constraints on resources and limits on feasibility of execution, a high frequency door-to-door census of all households in the world would provide current, accurate and precise estimates of electricity access, including nuances on source and reliability. However, this is not technically feasible given current technologies and methods. What is available are infrequent surveys and remote sensing data.

2.1. Survey-based Approaches to Mapping Electricity Access

In 2010, the United Nations Secretary General launched the Sustainable Energy for All (SEforAll) initiative, a multilateral partnership to support global efforts to ensure universal access to modern energy services. Its annual tracking reports provide the most comprehensive data on electricity access rates for countries around the world, relying on nationally representative household surveys including the Demographic and Health Surveys (DHS), Living Standards Measurement Surveys (LSMS), Multi-Indicator Cluster Surveys (MICS), the World Health Survey (WHS), and national censuses. Access to electricity is not a core subject of these surveys, but many include modules with questions regarding the availability of electricity in a home, or the power source used for lighting that are used to provide a binary indicator of access.

These surveys, given their primary aims at studying national welfare and health measurement, are low frequency (every few years) with limited sub-national and country-level coverage. This means that the data that is available from these is a snapshot of electrification for a specific country at a specific time and not an up-to-date harmonized global picture of electrification.

The Global Electrification Platform (GEP) is a publicly accessible source for country level data that systematically utilizes survey data on electrification. It was jointly developed by The World Bank Group, the Energy Sector Management Assistance Program (ESMAP), and the KTH Royal Institute of Technology in Stockholm. The results on electrification use country data constructed through a patchwork of surveys: from 1990 through 2019, 1,282 surveys have contributed to the estimates available in the GEP and only 28% of countries have updated these through annually (IEA et al 2021). The estimates on electrification and associated data in the WEP provides a useful snapshot of the general global picture on electrification and progress on SDG7. While, useful at the macro and global level, the estimates that policymakers need to create policy at more local levels is not possible from this. Additionally, the data contained in WEP have an assortment of vintages, where some country estimates are more current than others.

Survey-based estimation is sporadic, spatially and temporally limited, and, at least based on current methodologies, is typically not current. The closest that efforts have come to doing this in a systematic, regular and internationally harmonized manner is through the World Bank's Multi-Tier Framework (MTF) initiative. The MTF classifies access along a tiered spectrum, from Tier 0 (no access) to Tier 5 (highest level of access), relying on survey-based measurement of different modes of energy usage, including electricity. At the present time, surveys using the MTF cover 17 countries. The framework provides a comprehensive basis to assess progress toward SDG7 and detailed measurement of energy consumption at the household level. However, it is subject to the constraints of survey-based methods to measure electrification namely, given the logistical complexity of mounting a survey, low frequency.

2.2. Use of Satellite Data to Generate Electricity Access Mapping

At present, the methods that generate relatively accurate access information with high frequency and spatial resolution rely on satellite imagery of human settlements and night-time lighting (NTL). Satellite data is relatively low-cost, with high spatial and temporal resolution. The drawback of satellite data to estimate electrification is that it is second-best to a survey-based methods: with survey-based methods, there is direct measurement of electricity usage by a specific population, whereas satellite data infers both electricity usage (proxied through nighttime lighting information) and population.

Inferring electricity access from satellite night data is non-trivial. Among the technical challenges are acquiring and preparing data on NTL and human settlements, and correctly inferring electricity access from NTL given interference from other sources of light such as the moon, reflections, gas flares and other ambient light. A number of studies have been pushing the technical boundaries to use NTL for inferring electricity access (Doll and Pachauri 2010, Dugoua et al 2018, Elvidge et al 2011, Falchetta et al 2019, Min et al 2013). However, while overcoming technical challenges, these studies tend to be limited – they are not systematically sustained, often limited in geographical scope and are typically one-off productions. While the methods employed are replicable and can be updated, they are not typically updated unless the authors have some incentive to do so. Additionally, the work is not built for constant update and easy public access.

The value addition of UNDP's work on mapping electricity access is a system that provides public access to high quality electricity access data at a high spatial and temporal resolution. Two methods were used to generate electricity access estimates: High Resolution Electricity Access and machine-learning estimation of electrification.

3. High Resolution Electricity Access

To identify electricity access gaps across the world, the High Resolution Electricity Access (HREA) project leverages high resolution satellite data to develop estimates of electrification, energy use, and power supply reliability down to the settlement-level. Developed at the University of Michigan, the HREA project uses large-scale computational analysis of the complete 250-terabyte archive of every nighttime light image captured by the VIIRS sensor since 2012 to generate high-resolution temporal light signatures over every human-built settlement.

3.1. Input Data

3.1.1. Nighttime Lights

For daily nighttime imagery of the planet, Visible Infrared Imaging Radiometer Suite (VIIRS) data from the Suomi National Polar Partnership (SNPP) satellite in the form of 15 arc second GeoTIFFs spanning the globe each night was used. Each image strip has data from two 750 m sensors, as well as additional useful metadata. The key data comes from the Day/Night Band (DNB), which is the visible radiance, and which are in nanowatts/cm^2/sr. The second sensor data type is thermal infrared (TIR), which is in W/m^2/sr/μm. In addition, there is information on lunar illumination (LI) measured in lux, the sample position, and a quality flag bit which is necessary for subsetting only useful, "good" data (discarding data whenever the following are detected: clouds, fire, lightning, high energy particles, or stray light). Only data that are truly in the nighttime zone (solar zenith angle above 101°) are kept, and only data with a lunar illumination below .001 lux are kept.

3.1.2. Settlements and Population

Two primary pieces of information are required to generate locally relevant electricity access estimates: where people live, and how many people live there. The Facebook High Resolution

Settlement Layer (FBHRSL) and Global Human Settlement Layer (GHSL) are used. Both products use machine learning to identify built-up areas based on satellite imagery and fine-grained census data to interpolate population density in these areas.

FBHRSL has higher spatial resolution than the GHSL product. The resolution of FBHRSL is 1 arc second (about 30 meters), while that of GHSL is 250 meters (about 8 arc seconds). While a 30 m version of GHSL exists, the most recent estimates are not available at that level. Since HREA has focused on recent years, the 2015 GHSL estimates are used. Building and population data from Facebook are for around the same time period.

In order to generate absolute estimates of the numbers of electrified vs. unelectrified people in a country, the population data is rescaled to more accurate estimates for the time period of interest using United Nations Department of Economic and Social Affairs "World Population Prospects" dataset.

3.1.3. Geographic Boundaries

The database of Global Administrative Areas (GADM) was used for national and subnational-level administrative boundaries.

3.1.4. Land Cover

Variations in land cover and their albedo (surface reflectance) are important in the observed brightness of a given area of the planet. The interaction effects of albedo and lunar illumination are especially significant. For example, both snow and desert sand reflect much more light and thus appear brighter than forested areas. For this reason, Moderate Resolution Imaging

Spectroradiometer (MODIS) land cover classification data developed by NASA was used. Specifically, the MCD12Q1 product in the IGBP Land Cover Type Classification was used.[1]

3.1.5. Processed Data

Because the VIIRS data have a resolution of 15 arc seconds, a 15 arc second grid that encompasses the country boundary was generated and intersected with settlement data. In the case of the FBHRSL, 225 of the 1 arc second settlement cells can fit into each 15 arc second VIIRS cell. While the FBHRSL, VIIRS, and MODIS data use the WGS 1984 projection, the original GHSL data uses a Mollweide projection. To more cleanly match the GHSL cells to the 15 as grid, the 250 m GHSL is reprojected to 7.5 arc second WGS84 using bilinear interpolation. While a smaller resolution could be used, the 7.5 arc second projection is close to the original 250 m resolution (approximately 8-9 arc seconds), and this reduces increasing the total population count through the interpolation process. GHSL cells with population values less than two are recoded as "non-settlement" cells, and dropped. This is for two reasons: 1) the projection process creates many small noise artifacts that are likely not good indicators of populated areas; and 2) the GHSL, unlike FBHRSL, has occasional large contiguous tracts of barely populated areas that do not appear to correspond well to actual built-up areas.

Similar to how the settlement cells are matched to the VIIRS cells, VIIRS cells are matched to MODIS cells. This is done by intersecting the centroids of the 15 arc second grid to the 300 arc second MODIS data. Thus each 15 arcsecond grid cell has a corresponding land cover type, and each 1 or 7.5 as settlement cell has a corresponding VIIRS cell.

[1] 2012 global mosaics that were reprojected to approximately 300 arc second resolution WGS 1984 GeoTIFFs by the Global Land Cover Facility at the University of Maryland.

3.2. Data Processing

With data readied, the next step is to process it to generate electricity access estimates.

3.2.1. Identifying Isolated Areas

To determine expected natural background radiance, 15 arc second grid cells which contain zero settlement cells, and all of whose eight adjacent neighbors contain zero settlement cells, are selected.

3.2.2. Outlier Removal

Next, outliers among non-settlement areas are removed, i.e., brightly lit unpopulated areas are dropped as they should be dark. These places might be unexpectedly light because they are roads or gas flares, for example. This is done in two steps. First, the means and standard deviation of the DNB radiance is calculated for each candidate cell. Then, by land cover type, the quantiles of the means and standard deviations are calculated. Cells are only kept if their means and standard deviations are between the 1st and the sum of the 50th percentile plus the difference between the 50th and 1st percentile (this is more robust than simply using the 99th percentile as a cutoff, as the goal is to omit bright outliers). After taking a random sample of each land cover type, individual outlier observations are dropped. Any nightly observation that is above the median of the logged radiance plus four times the standard deviation for a given land cover type is removed. Thus, a representative sample of actually dark places remains.

3.2.3. Regression

Next, for each calendar year, a linear mixed effects model on light output for all pixels in areas with no settlements is run,

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{Z}\mathbf{u}+\boldsymbol{\epsilon}$$

where

- **y** is a vector of observed radiance
- **X** is a matrix of observed covariates (lunar illumination, time, month, land cover, and land cover crossed with lunar illumination)
- **Z** is a vector of observed dates
- **β** is an unknown vector of fixed effects
- **u** is an unknown vector of random effects with mean 0 and variance G
- **ϵ** is an unknown vector of random errors with mean 0 and variance R

Below is the formula used for the mixed-effects model for the *i*th isolated non-settlement cell observation of the *j*th date:

$$Light_{ij} = \beta_{0j} + \beta_1 Lunar\ Illumination_{ij} + \beta_2 Time_{ij} + \beta_3 Month_{ij} + \beta_4 Land\ Cover_{ij} + \beta_5 Land\ Cover \times Lunar\ Illumination_{ij} + e_{ij}$$

The model includes observations from a selection of isolated non-settlement pixels from all good quality nights, and includes fixed controls for month, land type, lunar illumination, local time, and the interaction between land cover type and lunar illumination, as well as a date random effect. Using these statistical parameters learned from data on non-settlement areas, the expected level of light output for all areas with settlements is calculated. These predicted values represent a

counterfactual estimate of how much light would be expected on that specific day on that type of land, if the only sources of light were from background noise and other exogenous factors. Areas with higher observed light output than expected light output will be assumed to have electricity access.

3.3. Results

For each settlement, the difference between the observed and predicted counterfactual DNB radiance is calculated for each night. This is divided by the model sigma to create a standardized residual. These residuals, which are in effect z-scores, are then compared against a standard normal distribution to calculate the likelihood of the area being electrified in different ways. One method is to calculate the variability of the light output. To do this, each observation of each cell is coded as being lit or not based on different thresholds (85, 90, or 95% confidence), and then a proportion of the nights that are observed to be lit is generated. A second method is to take the mean of all residuals and compare this value against the standard normal distribution. If this mean value is above some confidence threshold (85, 90, or 95%), it is considered electrified.

These electrification scores can then be used to determine the electrification rates by multiplying them by the population. The electrification rate for a given area is the number of people living in electrified settlements divided by the total population of that area. Absolute numbers of people with or without access can similarly be generated by adjusting the populations by the estimated total population for that region in the given year.

A sample image has been rendered to showcase the output of this process (figure 1). It shows changes in electrification rates within African countries from 2012 to 2020 in the graphs on the left,

as well as a snapshot of the geographical distribution of electrification on the continent as of 2020 on the right. Blue areas are those that have reliable access to electricity, while areas in red represent populated places that lack high quality electricity access. The time series graphs show that, for the most part, there has been a steady upward trend in the percentage of the population with access to electricity in most countries. As of 2020, it is estimated that around 645 million people lack reliable electricity access in Sub-Saharan Africa.

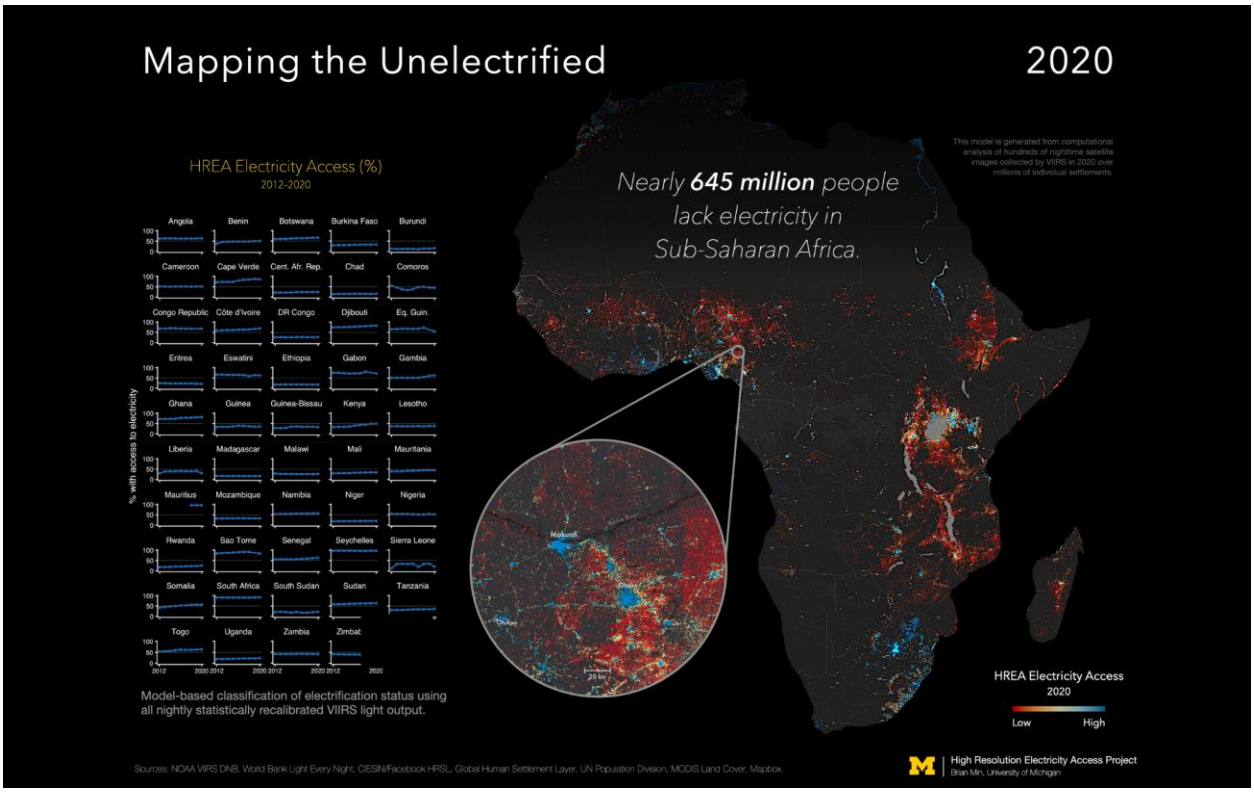

Figure 1. Sample output from HREA showing electricity access for Sub-Saharan Africa.

3.4. Validation

HREA electricity access estimates were validated against survey-based metrics of electricity access in a given year. HREA estimates of the populations of the settlement cells that are coded as electrified are summed up and divided by the total population of settlement cells within a given area. Survey data on access is used to generate electrification rate estimates for larger geographic units. The

intercept and slope of the corresponding regression line between HREA estimates and survey-based estimates, as well as the correlation coefficient, provides a strong preliminary test of correspondence.

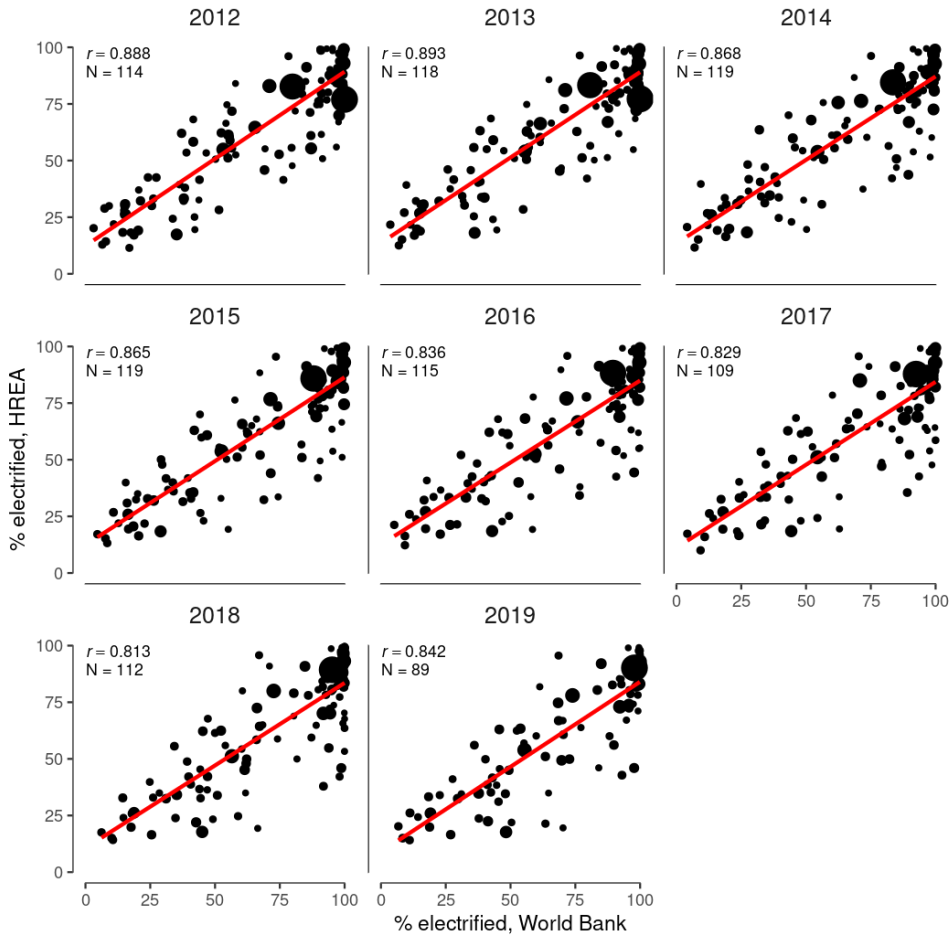

Figure 2. Each black dot represents a country-year observation, where size corresponds to population. The red line is the line of best fit. Pearson's correlation coefficient and number of observations reported in the top-left.

Figure 2 compares HREA estimates at the country-year level with those produced by the World Bank. Each black dot represents a country-year observation, of which there are 804 from 104 countries. World Bank estimates are on the x-axis, and HREA estimates are on the y-axis. The

diagonal blue line is the line of best fit through the data. While there are some large disagreements the correlation is strong (Pearson's correlation coefficient is .87).

Next, sub-national (administrative unit 1, or ADM 1) survey estimates of electricity access from eleven Demographic and Health Survey (DHS) microdata on electricity access and Living Standards Surveys were compared to HREA estimates for multiple countries across several years (figure 3).[2] First, HREA grid cells were spatially matched to sub-national regions and the proportion of population with access to electricity was calculated. Population figures are adjusted using yearly country-level estimates from the United Nations. This is done by weighting each settlement population by the appropriate weight such that the country's population in that year would equal the UN estimate, assuming a constant rate of change across the country.

[2] The units used are the administrative level 1 (ADM1) units of those countries, the specific name of which (state, region, province, or department) varies by country. Admin level 1 units are the first subnational geographic political unit below the country.

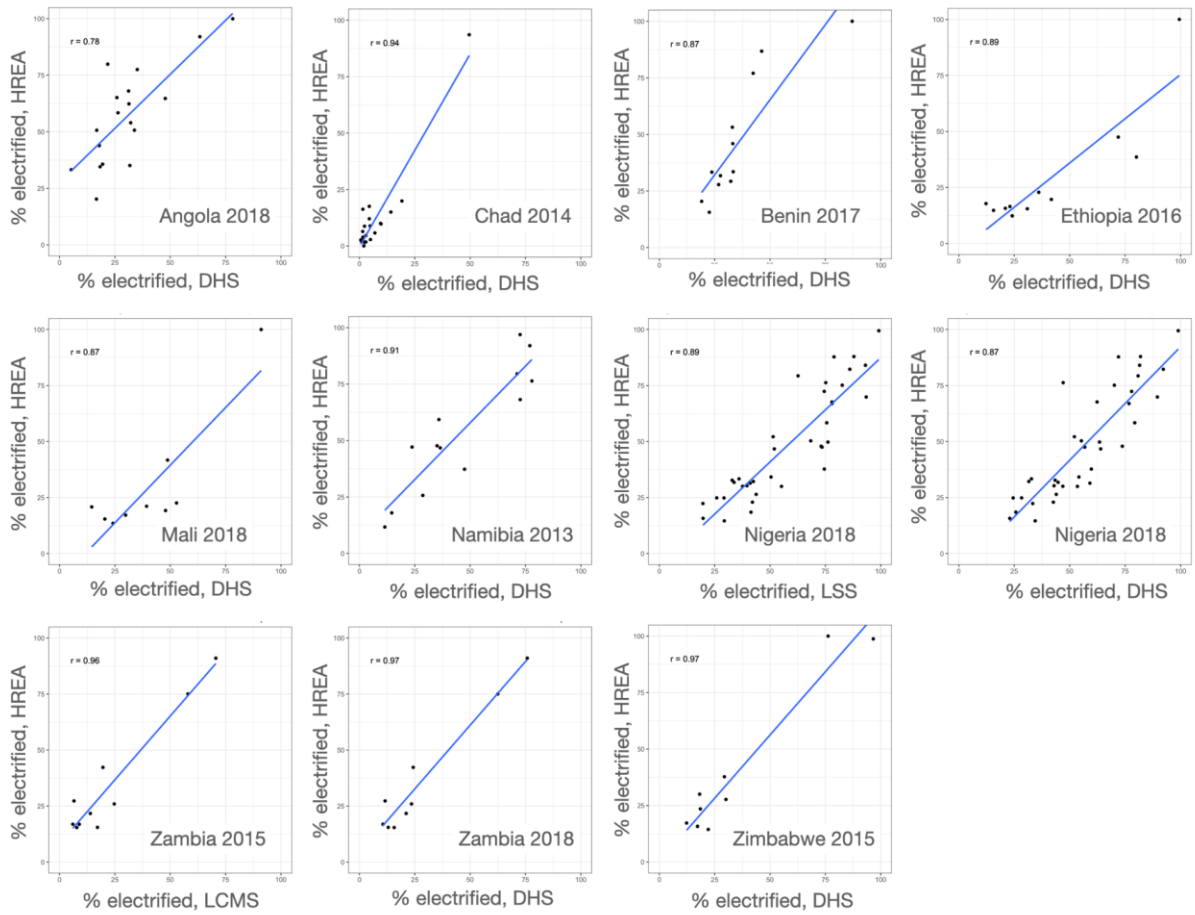

Figure 3. HREA electricity access estimates compared to survey based estimates.

The generally consistent methodology of the DHS makes it possible for more aggregated comparisons. In figure 4, all DHS estimates at the subnational level are plotted against HREA estimates. The size of each data marker is proportional to the population of the unit, and colors correspond to different countries. The diagonal black line represents the line of best fit through the data. As can be seen, it closely approximates a 45-degree line, which would represent perfect correspondence. Pearson's product moment correlation is .82, again indicating excellent fit between the remote-sensing based HREA measure with the survey-based DHS measures.

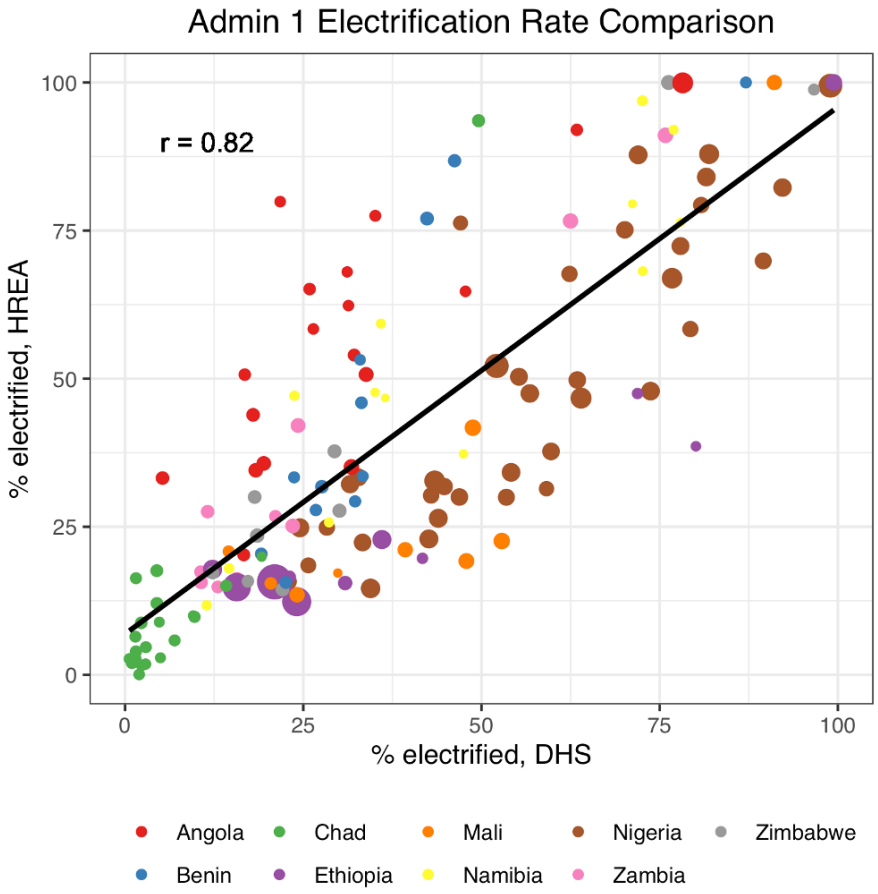

Figure 4. Pooled DHS and HREA electricity access comparison.

Finally, figure 5 shows a comparison of HREA with the MTF estimates for ADM1 level units in Zambia for 2018. MTF is designed to measure electricity access, therefore it is an ideal and high-quality external comparison benchmark. However, availability of the data at the time of writing was limited, so only one comparison – Zambia 2018 – was possible. The agreement between HREA and MTF estimates are once again quite consistent.

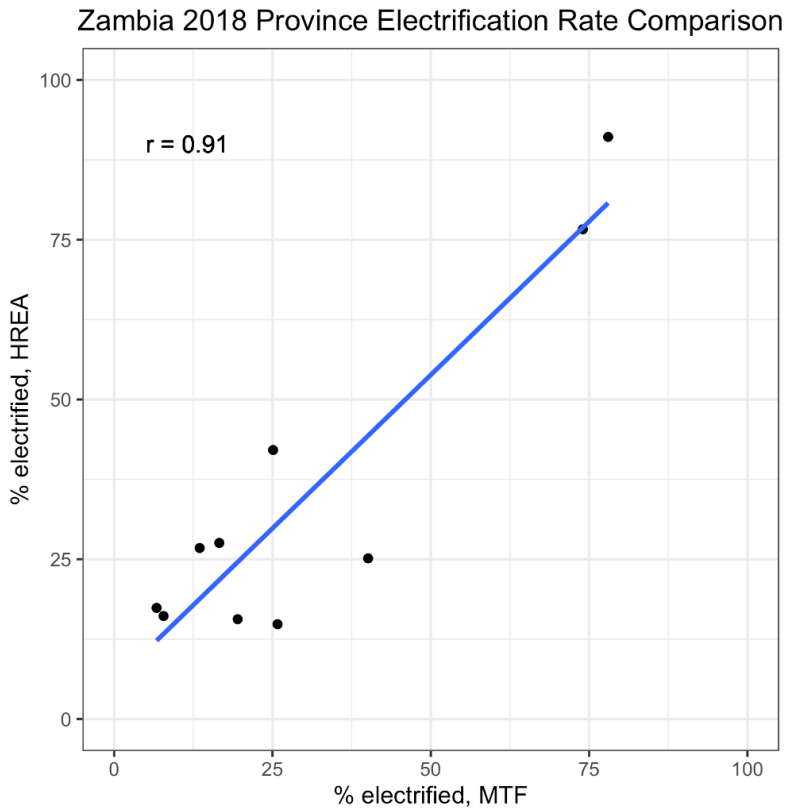

Figure 5. HREA estimates compared to MTF estimates for Zambia.

Overall, the comparisons of HREA estimates and survey-based estimates across a sample of different countries, years, and survey methodologies reveals consistently strong correlations. While there are some outliers and disagreements in the estimates, these appear to be about the same magnitude as the disagreements between different surveys. For example, the average absolute percentage point disagreement between DHS and MTF is 7.7 percentage points, while the average disagreement with HREA ranges from 8.4-10.6 percentage points; the average absolute percentage point disagreement between DHS and LSS is 8.3, while it is 11.3-11.8 for HREA compared to the other two.

4. Machine-learning Estimation of Electrification

To set up machine-learning estimation of electrification (MLEE), two critical pieces of information were needed: (a) detection and extraction of settlement information from satellite images for population density level; and (b) accumulation of information on the level of light pollution at night, enabling comparison of this information with expected illumination level at night. The first piece of information has been made available by a number of organisations: there is detailed information on population density and distribution for many countries from 2000 to 2020, with a resolution ranging from $100m^2$ to $1km^2$. Extracting information on the level of light pollution from night images required development of suitable transformation techniques to convert existing population levels for specific areas into the expected level of illumination at night. Electricity access was estimated with these two pieces of information in hand.

The machine learning approach created a model that can be used worldwide without any customization for specific countries. This means that MLEE can be applied universally across all countries and does not need to be customized for each country by changing the parameters of the model. The model is trained to use satellite images, given that they provide sufficient coverage and the level of detail needed for accurate outputs. It automatically updates electricity access estimates on a monthly basis, thus requiring minimal human input. This is a useful characteristic in resource and capacity constrained settings, as the application will be able to function everywhere with no trained staff.

Electricity access estimates can be produced daily, since the model works with standardised inputs and the data collected is comprised of multiple daily layers, meaning that one can be separated to produce daily estimates if needed. The only prerequisite for daily updates is weather on a particular day, as high cloud coverage (6 or more oktas) will undermine data extraction. Thus, if weather

conditions are favourable, the model can be consistently updated with a high frequency. Overall, the primary advantage of the model is the fact that machine learning is utilised to analyse data more effectively and discover hidden relationships between different layers. This allows for more accurate results in a shorter period, maximising the value of the final output i.e., the system produces an interactive map highlighting electricity access on a national, regional and local scale, whilst also constructing a table of values for easier data extraction and manipulation.

Moreover, the population layer is not a strict requirement for MLEE's operation, albeit the accuracy of the model will be enhanced if it is included after initial training. Reduced data may include a reduction in the number of optical and non-optical layers, as well as a reduction in accumulated data (the number of days) and the potential removal of the population layer. Due to the sophisticated architecture of the system, it can work with all types of reduced data although the accuracy may be lower as a result. Also, the application can work using limited or 'noisy' data, as the system is trained to filter noise out to provide the best results. Moreover, the MLEE system can use a wide variety of source data, meaning that all satellites operating in all resolutions can be used with no or extremely limited model changes. Therefore, the MLEE system supports both optical and non-optical bands, which makes is an extremely versatile system that can extract data from infrared, radar and optical bands. Notably, the system can also utilise optical and non-optical bands simultaneously, whereby the data is gathered from distinct layers to allow for greater accuracy; greater accuracy is ensured by verifying values in different layers and looking for discrepancies, which are subsequently marked as unsuitable if present. A final property of MLEE is that over time, with more data to train with, the accuracy of the predicted electricity access estimates will improve.

4.1. Input Data

WorldPop was identified as the best data source for the population layer, as it provides access to data with a resolution of $100m^2$ to $1km^2$ (WorldPop). Initially, data from NASA's Black Marble was considered for NTL. However, this data is limited to a narrow time range (annually accumulated data for 2012 and 2016). To overcome these limitations, Light Every Night - World Bank Nigh Time Light Data was used instead, with a large repository of monthly data at a high spatial resolution (World Bank). Finally, country border information as well as their sub-national divisions, were acquired from the Database of Global Administrative Areas (GADM).

The MLEE model's performance and accuracy will need to be monitored, with model re-training using new data only occurring in case of performance degradation. Currently, the model does not require re-training, as it simply finds the most suitable transformation coefficient (from light and population to an expected electricity access level); these coefficients will only be changed if there are significant alterations in population structure (such as towns growing and expanding to become large cities), which will be automatically captured by MLEE

4.2. Data Processing

Source data can have certain limitations including different band coding (layer characteristics), as some data is byte-encoded while other data uses their own specific data format. Therefore, a process was devised to avoid source data limitations, with the following steps:

- Acquiring population and night illumination data from any data sources with different resolutions, as all the data are re-scaled in the next steps.
- Extracting areas of interest using GADM (country and region information), which enables scaling-up and speeding-up the process (enabling use of a computational cluster-based calculation, instead of single-machine for this calculation).

- Joining night light bands into a single layer. Usually, satellites provide multi-band data, but subsequent steps of the process require one layer. There are a number of methodologies to transform RGB-based (or any other) layers into a single layer, from simple averaging to more complex calculations. Simple data averaging for visible (RGB) layers was chosen.[3]
- Normalising population and night illumination layers using non-linear interpolation. This means that the population and night light data ranges were split into bands. Then, using a machine learning (ML) optimisation algorithm called dual annealing (Xiang and Gong 2000, Xiang et al 2013), both layers were transformed into normalised datasets. A non-linear transformation was used due to an extremely wide range of population densities in comparison to the illumination level that can be produced. For example, an area with a population of a thousand people living in $1km^2$ can produce almost the same level of illumination as several thousand living in $1km^2$. Examples of transformation functions are shown in figure 6. This algorithm minimises the mean squared error between predicted electricity access and a baseline figure from a valid source. For this algorithm, the World Bank's country level electricity access percentage was used. Finally, another Machine Learning algorithm named Multi-layer Perceptron regressor (Glorot and Bengio 2010, He et al. 2015, Hinton 1989, Kingma and Ba 2014) was created and trained to transform normalised layers into an electricity access percentage number for a specific area.

[3] It worth noting that memory optimisation algorithms, such as GDAL Virtual File Systems, have been used to significantly speed-up calculations by more than 99%.

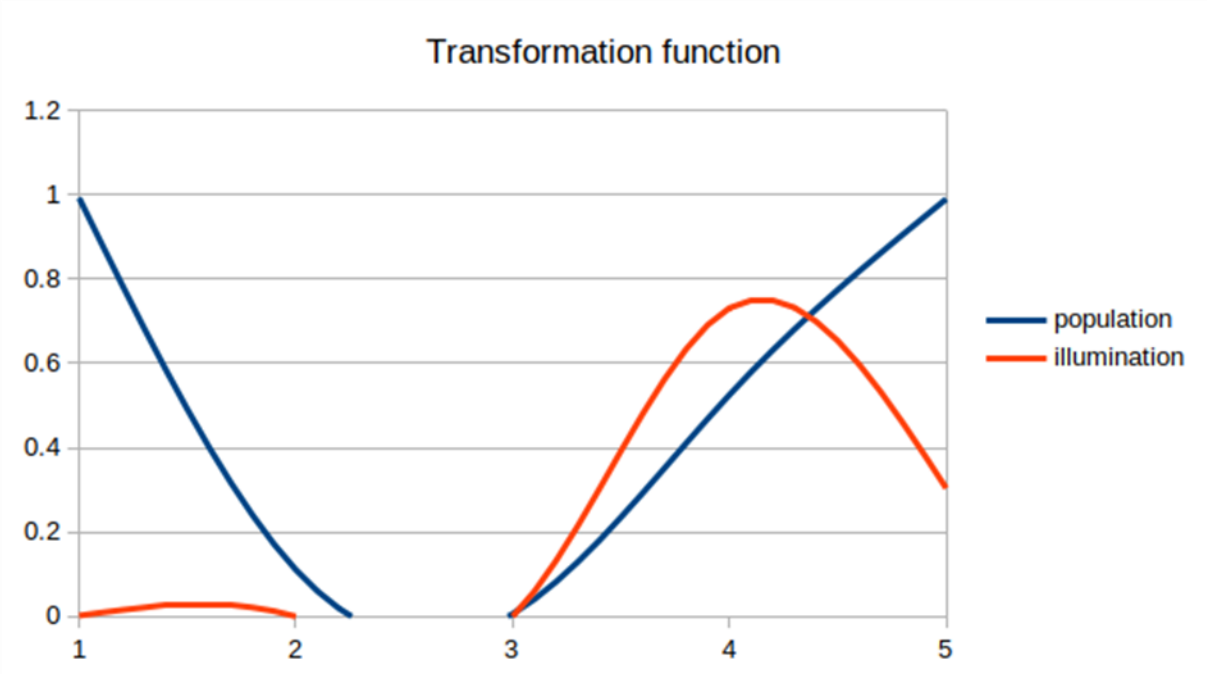

Figure 6. Transformation function representation.

- Next, the normalised 'real night light' layer was subtracted from the normalised population layer using cell-by-cell manipulation (figure 7). This allows filtering out any kind of 'non-settlement' lights, such as those emitted by gas towers and other industrial sites. The result can then be interpreted as the proportion of the population that has no electricity access.

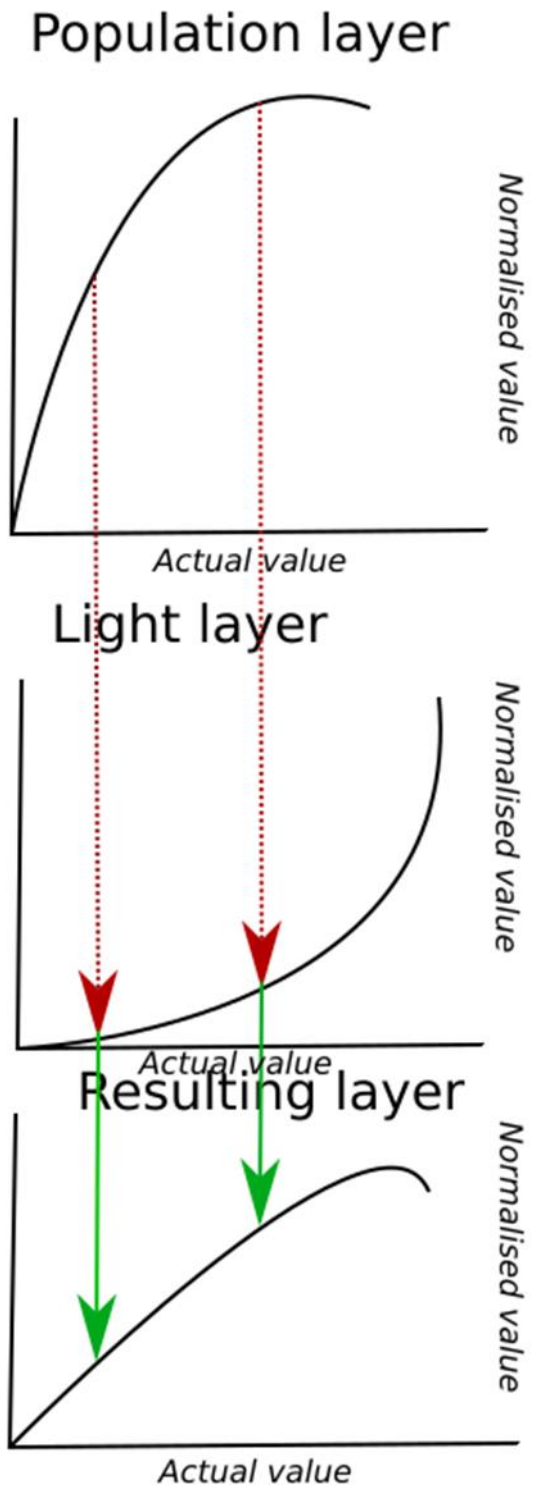

Figure 7. Principle of data transformation

- Finally, a set of GeoTIFF and Database tables was produced which visualise findings in both a static and interactive way.

4.3. Results

An example image that MLEE produces is shown in figure 8. The image shows access to electricity in each 1 1km^2 grid cell for the city of Ibadan, Nigeria in the year 2018. Darker areas show that few people have access to electricity, while lighter shades show higher access. The colour range reflects the proportion of population without electricity access ranging from zero to a hundred per cent. Some white points reflect areas where almost all persons have access to the electricity. To calculate the number of people without access, the proportion specified in each grid cell needs to be multiplied with the population (or population density) in that cell. This can be extended to the whole population (at the country level) to generate country-level electricity access level.

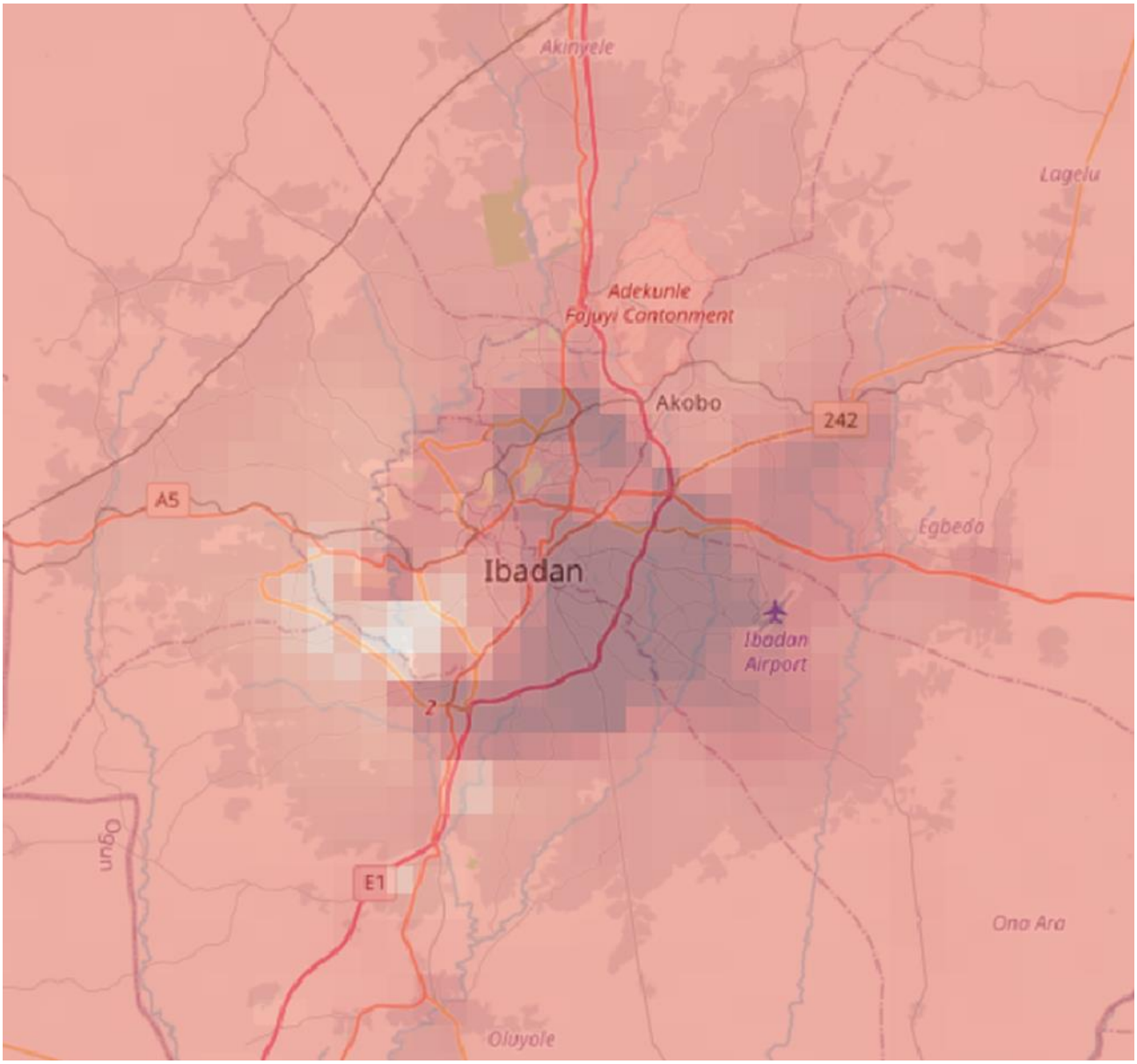

Figure 8. Example of a visualization showing the electricity access level for Ibadan, Nigeria.

4.4. Validation

The goal of MLEE was to create a universal approach to identify electricity access levels, regardless of the country or continent in question. The model training process finds the best coefficients to transform population and night light layers into new layers with a distribution of values in the range from zero to one. When the coefficients have been identified, they are applied at the relevant level (subnational or national level) and the results are then compared to external estimates. The first set

of estimates at the country-year level are compared to data from the World Bank (figure 9). Each dot represents a country-year observation. There are a total of 804 observations (red dots) from 104 countries. World Bank estimates are on the x-axis, and MLEE estimates are on the y-axis. The diagonal blue line is the line of best fit through the data. The results are promising, with a high correlation coefficient of 0.85.

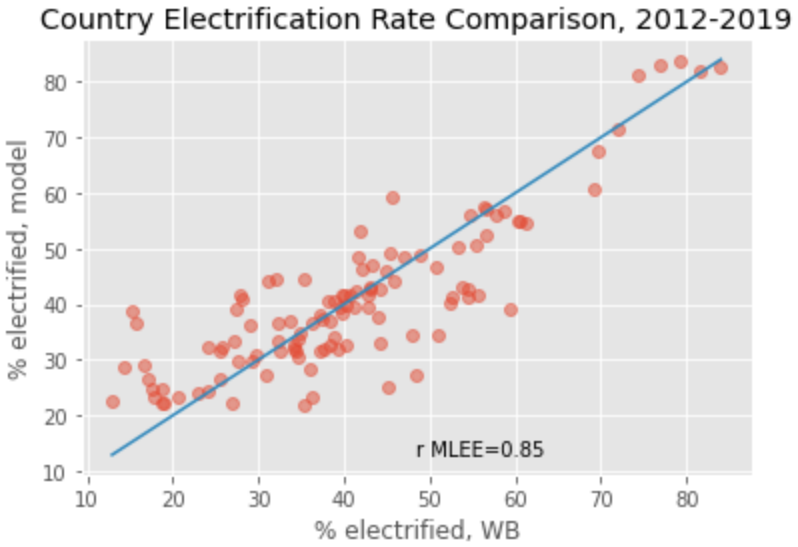

Figure 9. Each dot represents a country-year observation.

Next, using electricity access data for sub-national levels for specific years from DHS, the model was trained to then validate model generated estimates. The results of this are shown in figure 10. The 45-degree line (blue) is perfect correlation; red dots show actual subnational estimates from MLEE and DHS. Broadly, the results are promising, since the correlation coefficients are high, with the exception of Chad. High correlation coefficients suggest that MLEE is able to predict electricity access estimates that are close to DHS estimates.

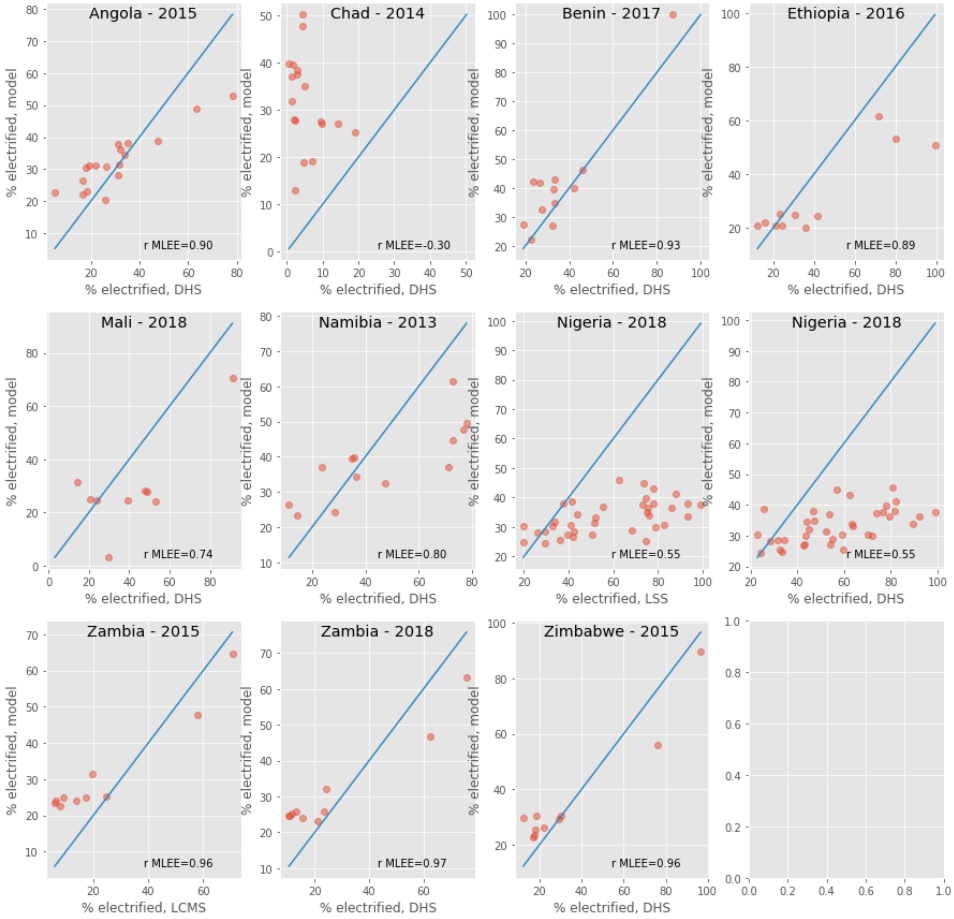

Figure 10. Validation results for MLEE using subnational estimates from DHS

These subnational comparisons were also conducted using a pooled sample is shown below. The overall correlation coefficient is a bit lower at the administrative level 1 likely due to variations in the approaches to data collection and processing.

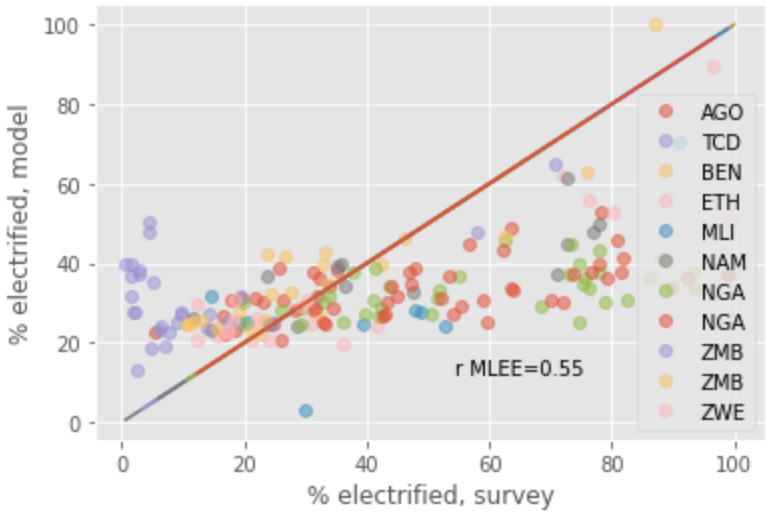

Figure 11. Pooled DHS and MLEE electricity access comparison.

A final comparison was done with subnational data from Zambia MTF 2018 (figure 12). The result is encouraging with a high correlation coefficient (0.97). MTF is likely a more reliable benchmark, thus a high correlation to it is an encouraging sign of the methodological validity of MLEE.

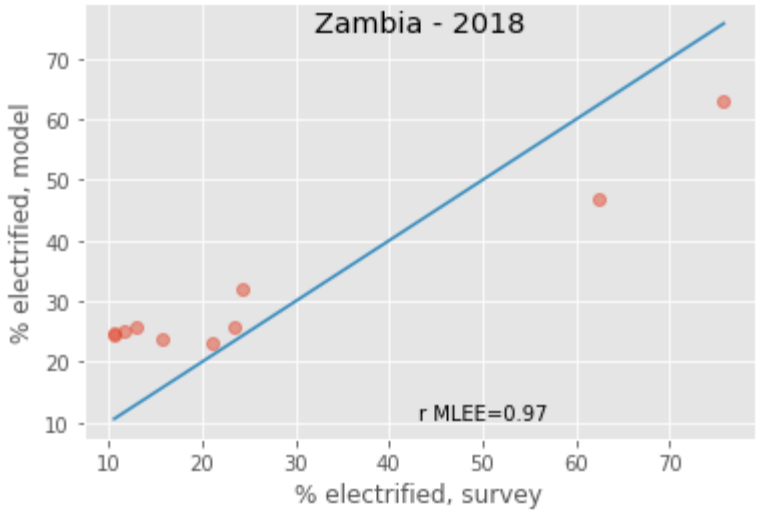

Figure 12. MLEE estimates compared to MTF estimates for Zambia.

5. Cross Method Comparison

How close are estimates generated by HREA and MLEE? To process these kind of data, it is often combined with high resolution vector data that represents, for example in this case, the administrative level 1 boundaries. One of the common operations that combine big vector and the raster data generated by the two methods of electricity access estimation is zonal statistics. Zonal statistics is a fundamental operation for processing the combination of raster and vector data to compute aggregate values for each subnational level using the values provided by the HREA and MLEA data. The output is the value of the aggregate function when applied to all pixels that overlap with each subnational level separately. The aggregate function used in this instance is the mean. Using results at the same subnational level of aggregation (administrative level 1), the two electricity access estimation methods were compared. Figure 13 shows a summary of the Pearson Correlation coefficient for the HREA and MLEE against the DHS data.

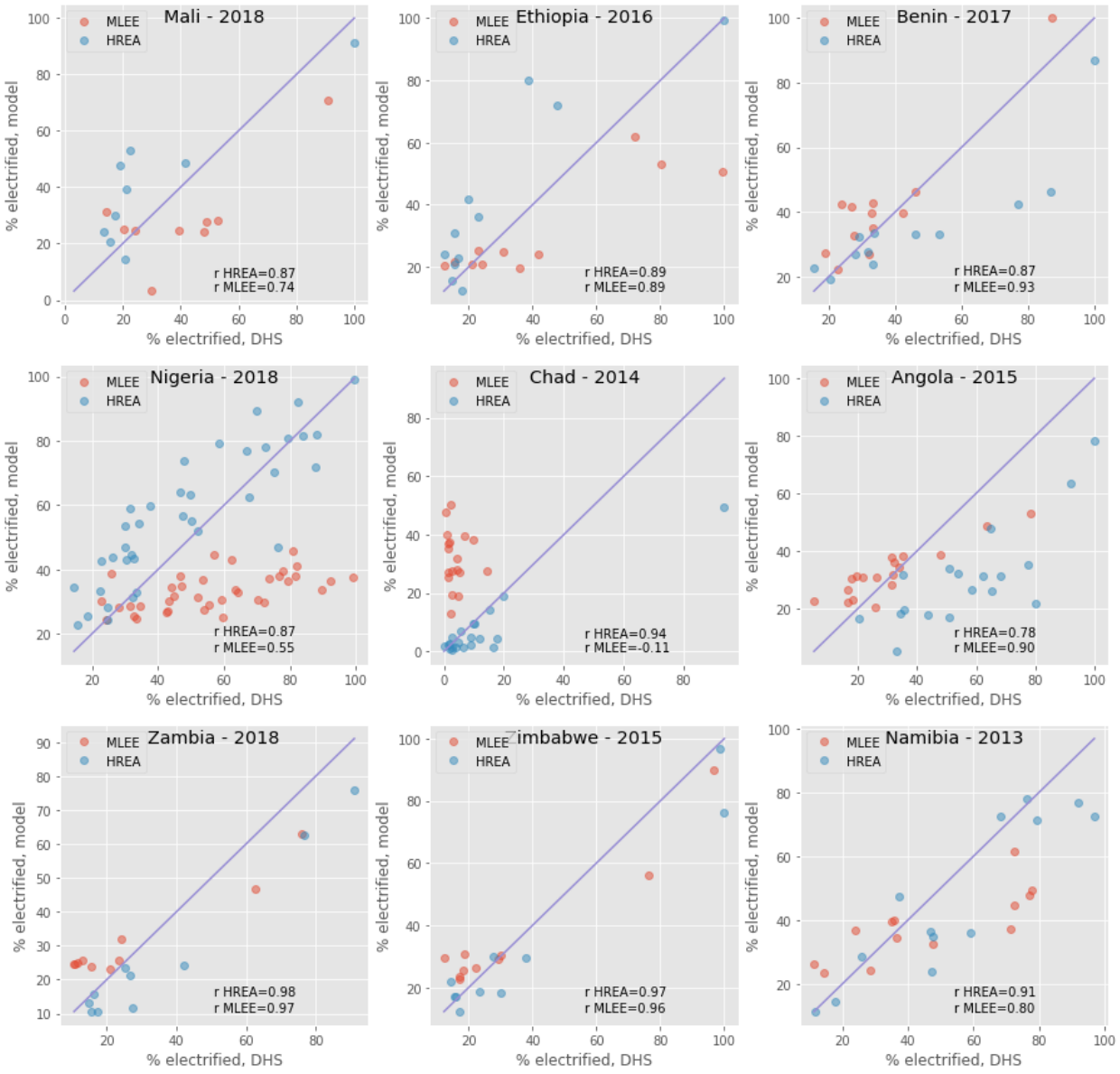

Figure 13. Pearson Correlation coefficient summary for HREA and MLEE vs. DHS

Table 1 summarizes the Pearson Correlation coefficient for HREA and MLEE with DHS in addition the correlation between HREA and MLEE. The HREA comparison to the DHS shows a high degree of correlation in the nine country-year data sampled. The MLEE comparison to the benchmark shows a high degree of correlation in all but one country, Chad, where the degree of correlation is very low. Comparing the HREA with the MLEE reveals a high degree of correlation in all countries but Angola, where there is a moderate degree of correlation.

Country-Year	HREA-DHS correlation coefficient	MLEE-DHS correlation coefficient	HREA-MLEE correlation coefficient
Angola 2018	0.78	0.98	0.41
Benin 2017	0.87	0.93	0.76
Chad 2014	0.94	0.11	0.60
Ethiopia 2016	0.89	0.89	0.87
Mali 2018	0.87	0.74	0.92
Namibia 2013	0.91	0.80	-0.80
Nigeria 2018	0.87	0.55	-0.71
Zambia 2018	0.97	0.97	0.89
Zimbabwe 2015	0.97	0.96	0.95

Table 1: Summary of Pearson Correlation Coefficients with the benchmark

The comparisons generated show that the two methods of estimating electricity access produce similar results with each other and with the external benchmarks, even with their different methodologies. While the HREA relies on a statistical technique, MLEE relies on machine learning based optimization. Both methods provide credible options for estimating global electricity access. With the exception of Chad for MLEE, both MLEE and HREA provide high correlations with external validation data from the DHS. The correlation between HREA and MLEE has a lot more variation; and, in two countries it is a negative relationship. The implication of this is that, while both the HREA and MLEE produce relatively accurate estimates overall (as suggested when compared with an external benchmark), the exact subnational estimates of electricity access can differ. In the most extreme cases (Namibia and Nigeria) subnational estimates of electricity access differ substantially so that the correlation coefficients are negative.

6. Conclusion: Learning and Challenges

Mapping global electricity access, while challenging, is possible to do. This paper described the different existing methods to measure access. It highlighted the fact that existing survey-based methods tend to be sporadic, logistically complex and resource intensive. The MTF effort offers a way to systematize survey-based approaches to measuring electricity access. On the other hand, are methods based on remote sensing data from satellites, which present their own unique technical challenges in terms of data processing and inference, while offering higher spatial and temporal resolution. The work done by the UNDP on electricity mapping has resulted in real progress on electricity mapping with high resolution. Following are key lessons learnt during this process.

6.1. Global versus Local Estimates

One key implication from this work is the level at which estimates for electricity access are available to policymakers. Estimates based on satellite imagery, such as HREA and MLEE, are able to produce relatively up-to-date and high-granularity estimates. This is useful to national and regional level policymakers. Both methods also produce aggregate, global level estimates, which is important at the global policymaking level to ensure tracking and that adequate funding is sustained. The methods also generate estimates that can be relied on, as the validation exercise demonstrates, both locally and globally. Thus, tracking progress is an outcome of this work that can be reliably had.

What is important to note is that, despite high levels of agreement between the methods and validation data, different methodologies produce different sets of specific results on electricity access. Thus, HREA, MLEE and DHS, while broadly in agreement on subnational estimates of

electricity access, are different in terms of their exact estimates for subnational units (figure 13). This is due to a number of factors. Satellite images have limitations such as cloud cover and local geophysical conditions which introduces variation by country and over time for measures of both NTL and local settlement information. Next, existing survey efforts, with the exemption of MTS, are not designed to capture estimates of access as their focus is on other categories of consumption. Finally, the estimation methods themselves will improve as more data is made available and underlying algorithms, technologies and methods evolve.

Consequently, the choice of model that the policymaker uses will determine the estimates they have available for actual policy decisions. The way to address this challenge may require relying on a combination of methods. A combination of local knowledge (grid placement, electricity generation, local projects), surveys and satellite-based techniques will help local level policymakers focus on areas that most require attention.

6.2. Institutional Collaboration

As documented in this report, several institutions are currently engaged in mapping estimates of global electricity access. These institutions use different methods and technologies, and have different scopes of work, though there is considerable overlap. Institutional collaborations offer a cost-effective way to leverage existing efforts. Institutional collaboration can add value by augmenting existing efforts. This can be achieved by playing a coordinating and resourcing role, systematizing existing work, in three ways.

First, with the right institutional collaboration, the effort to map global electricity access can be made far more systematic. This means that resources are brought that enable the production of

regular, high resolution and globally harmonized estimates of electricity access. Some of the cutting-edge efforts are ad hoc and sporadic, often led by small teams within academia and research. These can be leveraged by providing more resources and enabling them to grow.

Second, there is the real potential for growing and refining results with greater collaboration between upstream and downstream producers. Upstream producers include those that execute measurements and make datasets available, while downstream producers process that data to generate products centered around estimates of electricity access. For instance, Facebook produces settlement and population datasets that are crucial to estimates of electricity access. Upstream collaboration with agencies that generate remote sensing data. At the same time, working with downstream users, these datasets and map layers can be optimized to fit the purpose of estimating electricity access better.

Third, there is a distinct opportunity to bridge the two broad methods of electricity access estimation, i.e., survey-based and remote sensing based methods. Both methods have their advantages and there is space to integrate the two systematically so that they effectively complement each other to produce a higher quality of results estimated than either method would achieve individually. Therefore, institutional collaboration to bridge the two methods can serve to improve data quality.

As an example, the institutional partnership between the UNDP and collaborators from the University of Michigan meant that a high-quality electricity access dataset with the means to visualize it was produced. With resources mobilized by the UNDP, HREA was extended to global coverage, updated to the present time and thoroughly validated. This has meant that a systematic approach is now in place to continually generate electricity access estimates. Additionally, collaboration enabled

building in-house capability in the domain of satellite-based electricity access estimation that used modern machine learning tools.

6.3. Lag

No current electricity access estimation procedure is real-time, including the work undertaken and described here. That means that no existing method provides local level electricity access estimates for the entire world in real time. However, the work described here comes very close, providing electricity access estimates for the last calendar year. This lag is a function of the frequency with which NTL data is made available and how quickly it can be processed to share. Thus, how quickly the requisite VIIRS data are updated along with computational capacity limit the ability to achieve real time estimates.

6.4. Data Production, Storage and Visualization.

There was a need to ensure clarity in the output produced by the analytic work that generated local high resolution electricity access data and the infrastructure that eventually absorbed this data (storage and visualization). The problem was somewhat simultaneous as both the infrastructure and the input data (electricity access estimates) were evolving side-by-side. Good coordination was needed between the parties generating the data and those developing the storage and visualization infrastructure to ensure that the pieces integrated well. This essentially meant regular contact between the teams working on this project and agreement on data formats. Moreover, the volumes of data involved are very large, therefore substantial computational infrastructure was required to ensure adequate storage, processing and accessibility of all data by production staff and end-users.

Next, ML procedures – as used for MLEE for instance – require considerable time to train and optimise, although the time taken is dependent on the quality and resolution of source data. For example, when switching from a $1km^2$ cell to $100m^2$ cell, the number of cells to be analysed increases by 10,000, which causes data storage size to also increase.

Finally, visualizing electricity access data is nontrivial. It requires thoughtful choices on what users can do without overwhelming them. At the same time, visualization technologies are resource constrained and can produce a limited set of usable visualizations at a given point in time (i.e., at the time of user demand). The sheer difficulty of visualizing electricity mapping data stems from the remarkably high resolution of the datasets (30 meters). This poses challenges on both sides, the client side where the web browser can be overwhelmed by the number of pixels/geometries that need to be displayed at a certain scale and on the server side where significant amounts of resources are required to fetch, aggregate, and render the data. As a result, specific approaches are necessary to provide the users with consistent and effective visualizations, like scale-based aggregation of data in the form of vector tiles or advanced clustering scale dependent visualization algorithms like heat maps. Finally rendering the data on the server side as images and serving them through web mapping is also a viable alternative. Usually, all the above enumerated methods need to be combined to produce satisfactory user experiences. However, to make the most of the data, the visualizations need to be dynamic, in the sense that users need to be able to change various parameters and get almost instant feedback and results and this is where the difficulties lie.

These types of requirements are hard to implement and usually require time, expensive commercial off the shelf (COTS) software boxes, specific dedicated hardware infrastructure and highly skilled personnel. This combination of factors can result in prohibitive costs for low-income countries.

Additionally, in some areas of the world, the lack of available highly skilled and specialized human resources can pose insurmountable problems and can lead to inefficient implementations.

6.5. Validation

The core challenge with remote sensing data is validation. Both the HREA and MLEE use satellite-based settlement (population) and NTL data to generate electrification estimates, which creates challenges for validation. With survey-based methods, both the number of people and their electrification status are verified by direct observation. Satellite data require ground truthing to verify them. Both the HREA and MLEE used existing micro and national level electricity access figures to validate estimates that they generate. This goes beyond what any existing efforts have done and goes a long way to ensuring that the estimates generated by HREA and MLEE can be relied on. However, in the long run, systematic global and national level ground truthing will need to be undertaken to ensure continued validation. As we noted in section 5, HREA and MLEE estimates can differ at the subnational level. This is significant for a policymaker at the national level – what estimation tool they use may provide different local level estimates. The way to resolve this is coordinated data collection to enable systematic ground truthing and an opportunity for satellite-based methods to learn as more data comes in. At the same time, globally aggregated values do not differ, thus global level policymaking can use results from satellite-based methods to track progress.

References

Barron, M., & Torero, M. (2014). "Electrification and Time Allocation: Experimental Evidence from Northern El Salvador," MPRA Paper 63782, University Library of Munich, Germany.

Barron, M., & Torero, M. (2017). Household electrification and indoor air pollution. *Journal of Environmental Economics and Management*, *86*, 81-92.

Bhatia, M. and Angelou, N. (2015). Beyond Connections: Energy Access Redefined. ESMAP Technical Report; 008/15. World Bank, Washington, DC.

Chakravorty, U., Emerick, K., & Ravago, M. L. (2016). Lighting up the last mile: The benefits and costs of extending electricity to the rural poor. *Resources for the Future Discussion Paper*, 16-22.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review*, *101*(7), 3078-3108.

Doll, C. N. & Pachauri, S. (2010). Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy* 38, 5661–5670.

Dugoua, E., Kennedy, R., & Urpelainen, J. (2018). Satellite data for the social sciences: measuring rural electrification with night-time lights. *International journal of remote sensing*, *39*(9), 2690-2701.

Elvidge, C. D., Baugh, K. E., Sutton, P. C., Bhaduri, B., Tuttle, B. T., Ghosh, T., ... & Erwin, E. H. (2011). Who's in the dark—satellite based estimates of electrification rates. *Urban remote sensing: Monitoring, synthesis and modeling in the urban environment*, *250*, 211-224.

Falchetta, G., Pachauri, S., Parkinson, S., & Byers, E. (2019). A high-resolution gridded dataset to assess electrification in sub-Saharan Africa. *Scientific data*, *6*(1), 1-9.

GADM. Retrieved February 6, 2022, from https://www.gadm.org/

Glorot, Xavier, and Yoshua Bengio. (2010). "Understanding the difficulty of training deep feedforward neural networks." *International Conference on Artificial Intelligence and Statistics*.

Grogan, L., & Sadanand, A. (2013). Rural electrification and employment in poor countries: Evidence from Nicaragua. *World Development*, *43*, 252-265.

Hassan, F. & Lucchino, P. (2016). "Powering Education," CEP Discussion Papers dp1438, Centre for Economic Performance, LSE.

He, Kaiming, et al. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." arXiv preprint arXiv:1502.01852 (2015).

Hinton, Geoffrey E. (1989). "Connectionist learning procedures." *Artificial intelligence* 40.1. 185-234.

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

Kingma, Diederik, and Jimmy Ba. (2014). "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980.

Khandker, S. R., Barnes, D. F., & Samad, H. A. (2012). The welfare impacts of rural electrification in Bangladesh. *The Energy Journal*, *33*(1).

Khandker, S. R., Samad, H. A., Ali, R., & Barnes, D. F. (2014). Who benefits most from rural electrification? Evidence in India. *The Energy Journal*, *35*(2).

Lee, K., Miguel, E., & Wolfram, C. (2017). Electrification and economic development: a microeconomic perspective.

Lipscomb, M., Mobarak, A. M., & Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil. *American Economic Journal: Applied Economics*, *5*(2), 200-231.

Min, B., Gaba, K. M., Sarr, O. F., & Agalassou, A. (2013). Detection of rural electrification in Africa using DMSP-OLS night lights imagery. *International journal of remote sensing*, *34*(22), 8118-8141.

NASA. NASA's Black Marble. Retrieved February 6, 2022, from https://blackmarble.gsfc.nasa.gov

Ritchie, Hannah and Max Roser. (2020). "Energy". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/energy' [Online Resource]

Stern, D. I., Burke, P. J., & Bruns, S. B. (2019). The impact of electricity on economic development: a macroeconomic perspective.

Van de Walle, D., Ravallion, M., Mendiratta, V., & Koolwal, G. (2017). Long-term gains from electrification in rural India. *The World Bank Economic Review*, *31*(2), 385-411.

World Bank. 2021. World Bank Global Electrification Database. World Bank, Washington, DC. https://databank.worldbank.org/source/world-development-indicators.

World Bank. *Light Every Night - World Bank Nightime Light Data*. World Bank. Retrieved February 6, 2022, from https://registry.opendata.aws/wb-light-every-night/

Worldpop. *Open spatial demographic data and Research*. WorldPop. Retrieved February 6, 2022, from https://www.worldpop.org/

Xiang Y and Gong XG. (2000). Efficiency of Generalized Simulated Annealing. Physical Review E, 62, 4473.

Xiang Y, Gubian S, Suomela B, and Hoeng J. (2013). Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R. The R Journal, Volume 5/1.