Open Energy Maps™: Electricity Access and Demand Layers

User Documentation V1.0

Updated March 26, 2024

Stephen J. Lee, PhD

leesj@mit.edu

1. Contributors, Sponsors, and Additional Resources

The Open Energy Maps™ (OEMaps) platform, the LItLDF model, the BEACON model, and the OEMaps electricity access and demand open datasets were made possible through key research collaborators, sponsors, and data partners. We would like to especially thank our East African utility and government data partners for enabling this research.

1.1. Contributors

Jay Taneja, John W. Fisher III, Ignacio J. Perez-Arriaga, Darlain Edeme, Joel Mugyenyi, Bob Muhwezi, Martin Kuppers, Christopher Dean, Daniele Bricca, Davide Fioriti, Rob Stoner, Andrea Micangeli.

1.2. Sponsors

International Energy Agency, Power Africa, the Energy for Growth Hub, Project InnerSpace.

1.2. Sponsor Disclaimer

This platform was made possible through support provided by the Power Africa initiative, led by the U.S. Government, and the U.S. Agency for International Development, under the terms of Award No. 720-674-20-IO-00001. The opinions expressed in this platform are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development.

1.3. Additional Resources

- *OEMaps Electricity Access and Demand Layers Web Platform*: https://www.openenergymaps.org/elec_demand
- *OEMaps Privacy Policy*: https://www.openenergymaps.org/privacy_policy
- *OEMaps Terms of Use*: https://www.openenergymaps.org/tos
- *Technical methodology*: Lee, S.J. Multimodal Data Fusion for Estimating Electricity Access and Demand (Doctoral Thesis). Massachusetts Institute of Technology, 2023. Retrieved from https://stephenjl.com/pdf/sjl-phd-thesis-final_20230831.pdf
- GitHub Repo for OEMaps Electricity Access and Demand Data: https://github.com/stephenjlee/oemaps-access-and-demand-data
- GitHub Repo for BEACON (access model): https://github.com/stephenjlee/beacon
- GitHub Repo for LItLDF (demand model): https://github.com/stephenjlee/litldf

2. Introduction

Electricity is critical to modern society, fueling everything from homes to industries. Yet energy poverty remains a significant barrier to economic growth in many low- and middle-income countries (LMICs). At the heart of this challenge is the need for accurate information on electricity access and demand, without which planners are navigating in the dark. Too often, this crucial data is fragmented or entirely absent due to various factors, including uncollected or non-aggregated meter information and restricted data accessibility. This gap hinders effective electrification planning and the judicious allocation of resources.

Understanding building-level access to electricity (electrification status) and accurately estimating electricity demand is paramount. Insufficient data can lead to the planning of redundant infrastructure or overlook areas where investment could yield significant benefits. Similarly, inaccuracies in demand estimation can lead to either over-planning, where infrastructure unnecessarily inflates costs and diverts resources from more needy areas, or under-planning, which misses the benefits of economies of scale and network utilization.

Fortunately, the landscape of data availability and analysis is transforming. Recent advancements in remote sensing and machine learning have enabled new capabilities for high-resolution, large-scale data analysis. These technologies allow us to characterize nearly every building across diverse geographical and national contexts. Leveraging widely available datasets, such as building footprints, nighttime lights, internet speed data, and land use data, alongside high-resolution satellite imagery, enables scalable machine learning inference to provide insights into sparsely measured or opaque phenomena like electricity access and demand.

In this context, we introduce Open Energy Maps™, a data platform providing (to our knowledge) the first building-level electricity access and demand estimates. These estimates are derived from a robust aggregation of remote sensing features, application-tailored machine learning techniques, and utility infrastructure and metered consumption data. While these results represent a significant advancement, they are preliminary and should be viewed as a stepping stone toward more refined analysis. Estimating building-level electricity consumption using only remotely sensed data is inherently challenging, as details about the internal dynamics of buildings are not directly visible from remote sensing features. Therefore, our approach yields probabilistic assessments, inferring probable characteristics based on correlations within the data we collect and analyze.

Although our models have shown promising results in East Africa, a limitation is their regional specificity. To enhance the generalizability of our findings and extend their

applicability across the African continent and globally, we are looking for partnerships to avail high-resolution datasets. We see Open Energy Maps™ as not just a tool but a potential platform to mediate data sharing and collaboration, aiming to illuminate the path to more informed and effective infrastructure planning worldwide.

3. Input Data

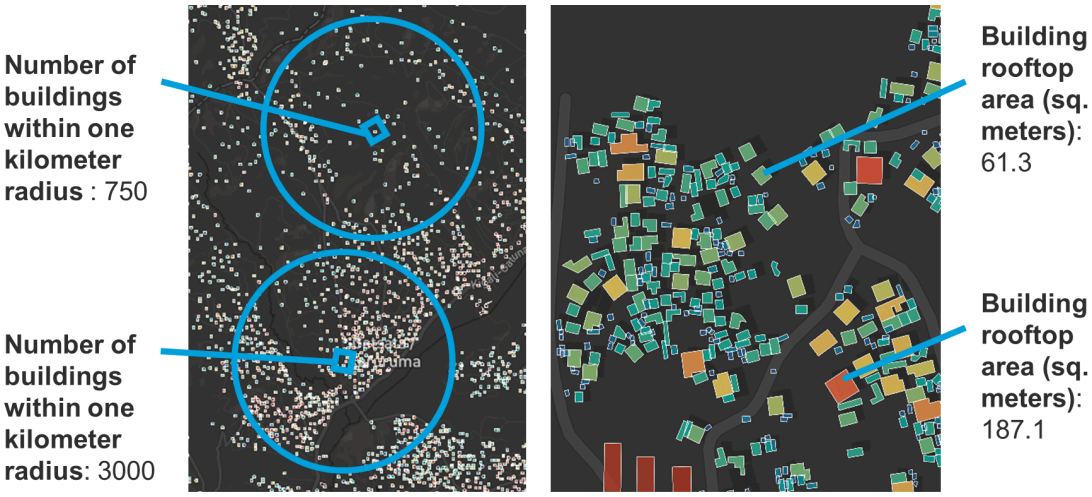
In the development of Open Energy Maps, we undertake a multi-step process of aggregating, processing, and integrating geospatial datasets to delineate and characterize individual buildings in Rwanda. This process encompasses several key steps outlined below:

3.1. Data Aggregation: We start by sourcing building footprints from two primary repositories: the Google Open Buildings[1] and Microsoft Global ML Building Footprints[2]. These datasets present individual structures as polygons marked by geospatial coordinates.

3.2. Data Merging: When building polygons from Google Open Buildings and Microsoft Global ML Building Footprints overlap, we prioritize the use of the polygon from the Google Open Buildings dataset. This merged dataset forms the base for further analysis.

3.3. Feature Extraction: Utilizing the consolidated building data, we compute essential features such as the area of each building footprint and local building density, quantified as the number of buildings within a one-kilometer radius of each building.

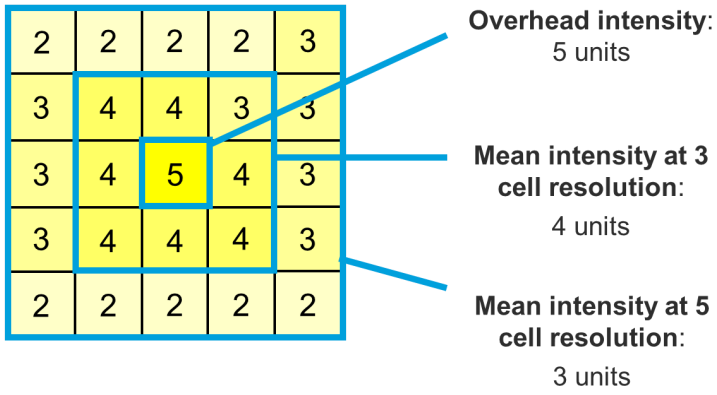
[1] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y.S.E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, J.A. Quinn. Continental-scale building detection from high resolution satellite imagery. arXiv:2107.12283, 2021.
[2] Microsoft. Global ML Building Footprints. GitHub repository, 2024. Available at https://github.com/microsoft/GlobalMLBuildingFootprints



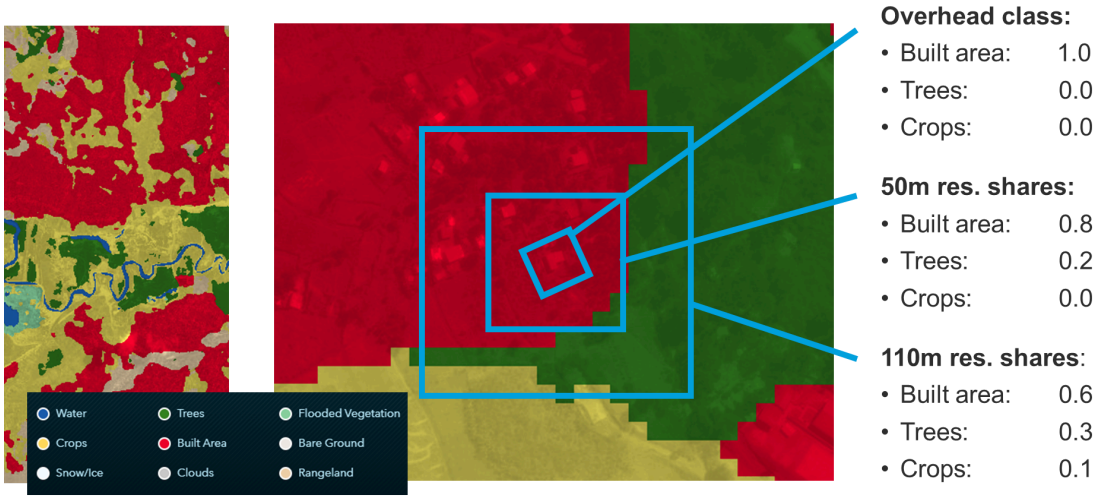
Building density and rooftop areas are calculated using building footprint datasets. Building density gives a metric corresponding to urbanization, and building footprint area is used to approximate the size of the building.

3.4. Spatial Joining and Enhancement: Building on the consolidated building dataset, we execute spatial joins with additional geospatial features using spatial joins. These enhancements include nighttime light illumination values from the VIIRS sensor[3], categorized land use types from Microsoft's Land Use Land Cover dataset[4], high-resolution satellite imagery from Esri[5], and internet connectivity metrics from Ookla[6]. Each feature is integrated, ensuring a detailed and comprehensive geospatial profile for each building.

[3] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In 2021 *IEEE international geoscience and remote sensing symposium IGARSS*, pages 4704–4707. IEEE, 2021.
[4] Christopher D Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, 13(5):922, 2021.
[5] Esri. World Imagery Map, 2022. Available at: https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9.
[6] Ookla. Speedtest by Ookla Global Fixed and Mobile Network Performance Map Tiles, 2022. Available at: https://github.com/teamookla/ookla-open-data

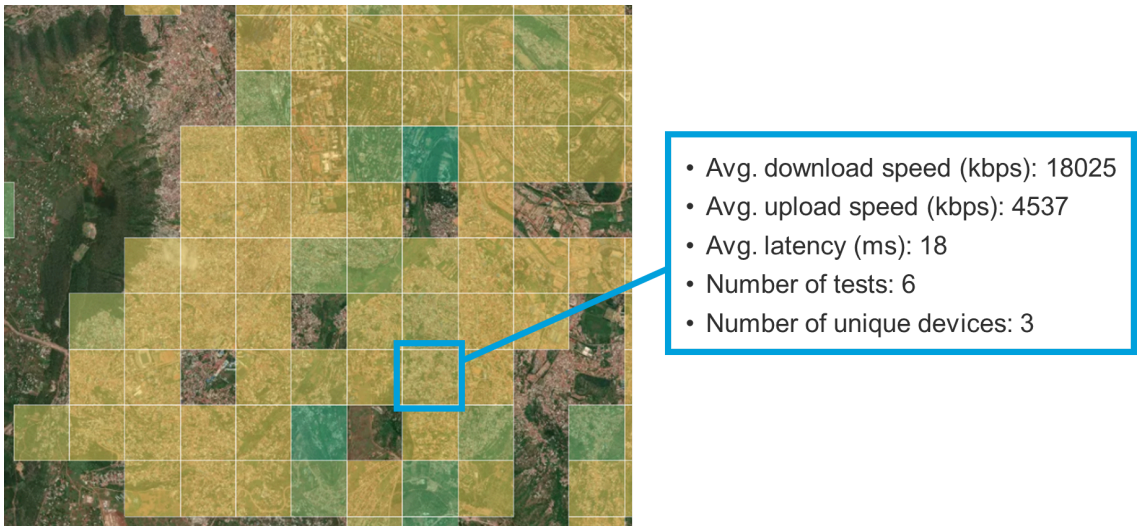


VIIRS nighttime lights reflect the distribution and brightness of artificial lighting at night. These data are globally accessible as yearly composite images with a spatial resolution of approximately 500 meters at the Equator (15 arc-seconds). Additionally, we compute average light intensity values for larger areas covering 11 and 51 grid cells.



We utilize land use and land cover (LULC) data from ESRI and Microsoft derived using computer vision techniques on Sentinel-2 imagery at a 10m resolution. We also compute LULC share statistics at resolutions of 110m and 510m.

High-resolution satellite images from ESRI, with spatial resolutions ranging from 50 cm to 1 meter, are used. These images are assembled and trimmed to ensure that each building aligns with a central tile positioned over its centroid.

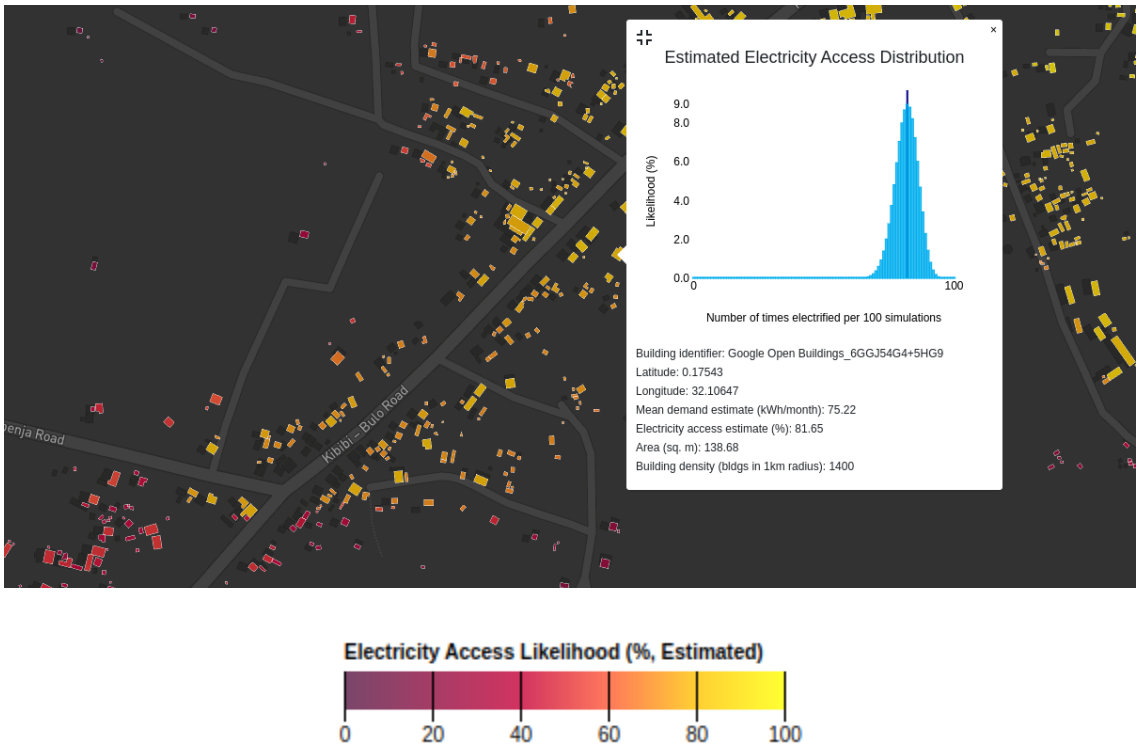


The Ookla internet speeds dataset provides information on internet performance metrics, including average download and upload speeds, average latency, the total number of speed tests conducted, and the number of devices participating in these tests. This data is geospatially consolidated at the Web Mercator zoom level 16, corresponding to a spatial resolution of approximately 610.8 meters at the Equator.

3.5. Ground Truth and Assumptions: We finally employ ground truth metered electricity consumption and location data sourced from East African utilities. We use the assumption that buildings within 40 meters of a meter connection are electrified, guided by local electrical infrastructure standards.

4. The BEACON Model for Building-level Electricity Access Estimation

The Bayesian Electricity Access ClassificatiON (BEACON) model for building-level electricity access estimation provides a key information layer for effective infrastructure planning in underserved regions. Please refer to Lee et al. for more detailed descriptions.[7] This model harnesses a lightweight data fusion (LDF) framework to deliver probabilistic insights into the electrification status of individual buildings, thus allowing for targeted and efficient infrastructure investment and aiding in the critical decision-making processes regarding the deployment of various technology choices and designs.

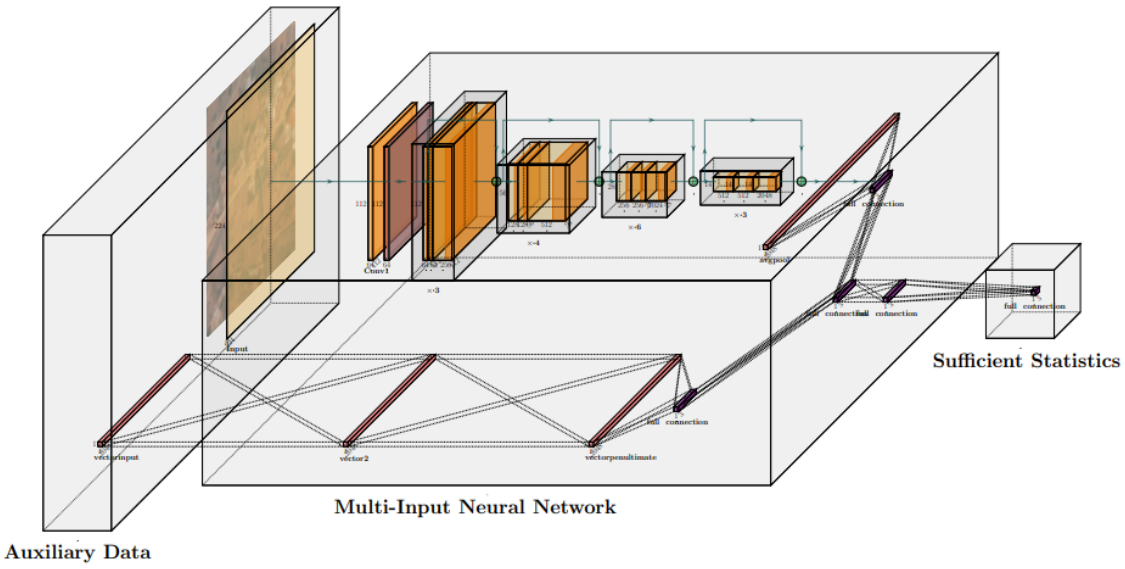


Sample outputs from the BEACON model, where point estimates for electricity access likelihood correspond to heatmap color values for each building footprint modeled. Any individual building also has an associated probability distribution, here represented as the likelihood of the number of times a building is electrified in 100 of our model's probabilistic 'simulations.'

[7] Lee, S.J. Multimodal Data Fusion for Estimating Electricity Access and Demand (Doctoral Thesis). Massachusetts Institute of Technology, 2023.

Related work pertaining to electricity access estimation, such as studies by Doll et al.[8] and Min et al.[9], including Min et al.'s High-Resolution Energy Access (HREA) dataset[10], primarily rely on correlations between nighttime lights and electrification; they provide electricity access estimates at settlement-level granularity or coarser but have not been successfully employed in higher-resolution planning contexts. Falchetta et al.'s GDESSA model VIIRS nighttime lights data, LandScan gridded population data, and MODIS land cover data for sub-Saharan Africa at 30 m and 1 km resolution but do not provide sub-provide-level validation[11].

Our approach via the BEACON model integrates diverse and high-resolution datasets and joins them to individual building footprints as described in Section 3. BEACON employs a mutli-modal neural network architecture embedded within a simple probabilistic graphical model, exploiting statistical conjugacy to efficiently generate electricity access estimates while characterizing the inherent uncertainty in these predictions. This application-tailored methodology not only boosts the model's accuracy but also its interpretability, providing a robust foundation for planning and investment decisions.



[8] Christopher N.H. Doll and Shonali Pachauri. Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy*, 38(10):5661–5670, 2010. ISSN 03014215. doi: 10.1016/j.enpol.2010.05.014.
[9] Brian Min, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou. Detection of rural electrification in Africa using DMSP-OLS night lights imagery. *International Journal of Remote Sensing*, 34(22):8118–8141, 2013. ISSN 0143- 1161. doi: 10.1080/01431161.2013.833358.
[10] Brian Min and Zachary O'Keeffe. High resolution energy access indicators. Github repository, 2020. Available at: https://github.com/zachokeeffe/nightlight_electrification.
[11] Giacomo Falchetta, Shonali Pachauri, Simon Parkinson, and Edward Byers. A high-resolution gridded dataset to assess electrification in sub-saharan africa. *Scientific Data*, 6(1):110, 2019.

The BEACON model employs a multi-input neural network designed to combine image data with vector-based representations of building properties. Images are currently exclusively high-resolution building-centered satellite images. Vector descriptions include information about nighttime lights values, nearby land use characteristics, etc.

The comparative analysis in Lee et al. reveals BEACON's superior performance against HREA and GDESSA models in terms of accuracy, F1 score, and calibration, underscoring its state-of-the-art capabilities in building-level electricity access estimation for at least the validation dataset employed based in East Africa.

Method	Accuracy (%)	F1 Score	AUC	Precision	Recall
LDF (ours)	**80.7**	**0.748**	**0.859**	0.810	0.696
HREA	70.9	0.539	0.691	0.778	0.412
GDESSA	48.3	0.580	N/A	0.436	0.866
Naive classifier	58.7	0.585	N/A	0.413	1.000

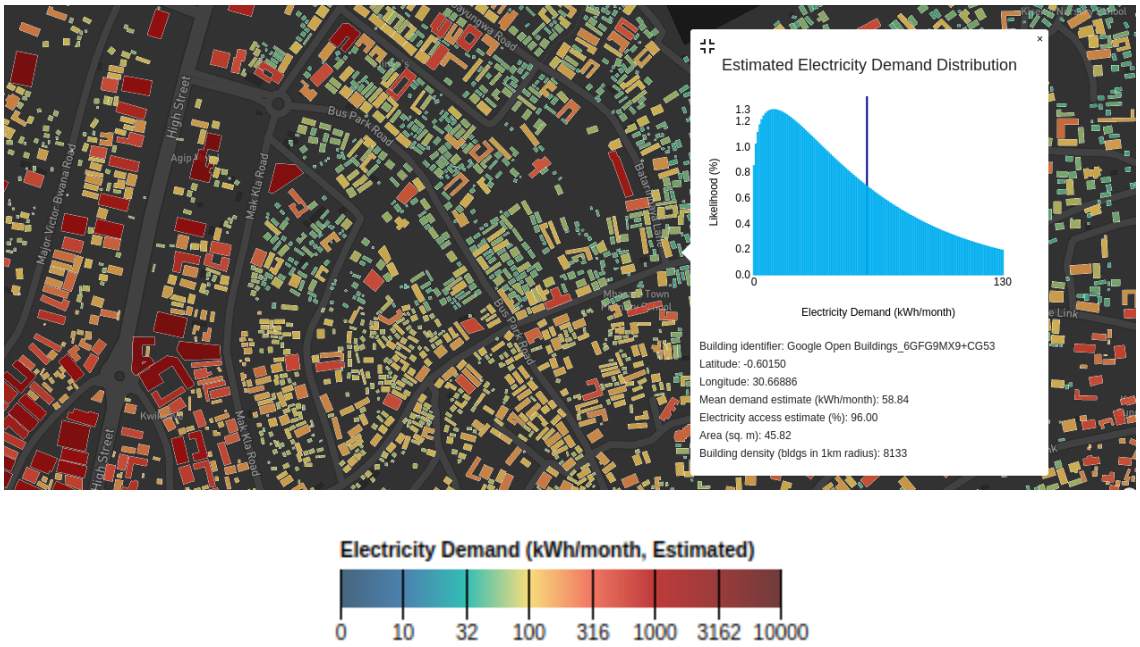
The BEACON model reflects higher accuracy, F1 Score, and AUC relative to baselines for our East African validation dataset.

BEACON represents a novel approach in the electrification planning domain, offering nuanced, probabilistic estimates of building-level electricity access that can significantly improve the allocation of infrastructure investments. Future advancements are expected to extend the model's applicability, further enhancing its utility in global electrification efforts.

5. The LItLDF Model for Building-level Electricity Demand Estimation

The Load Inference through Lightweight Data Fusion (LItLDF) model introduces an application-tailored approach for estimating building-level electricity demand. Please refer to Lee et al. for a more detailed description.[12] This methodology bridges gaps in existing demand estimation, particularly in low-access regions, by providing building-level estimates and addressing significant ambiguity from uncertain meter-to-building mappings. Ground truth data for metered consumption in these areas often comes with imprecise geolocation data, leading to complex building-meter relationship dynamics.

[12] Lee, S.J. Multimodal Data Fusion for Estimating Electricity Access and Demand (Doctoral Thesis). Massachusetts Institute of Technology, 2023.



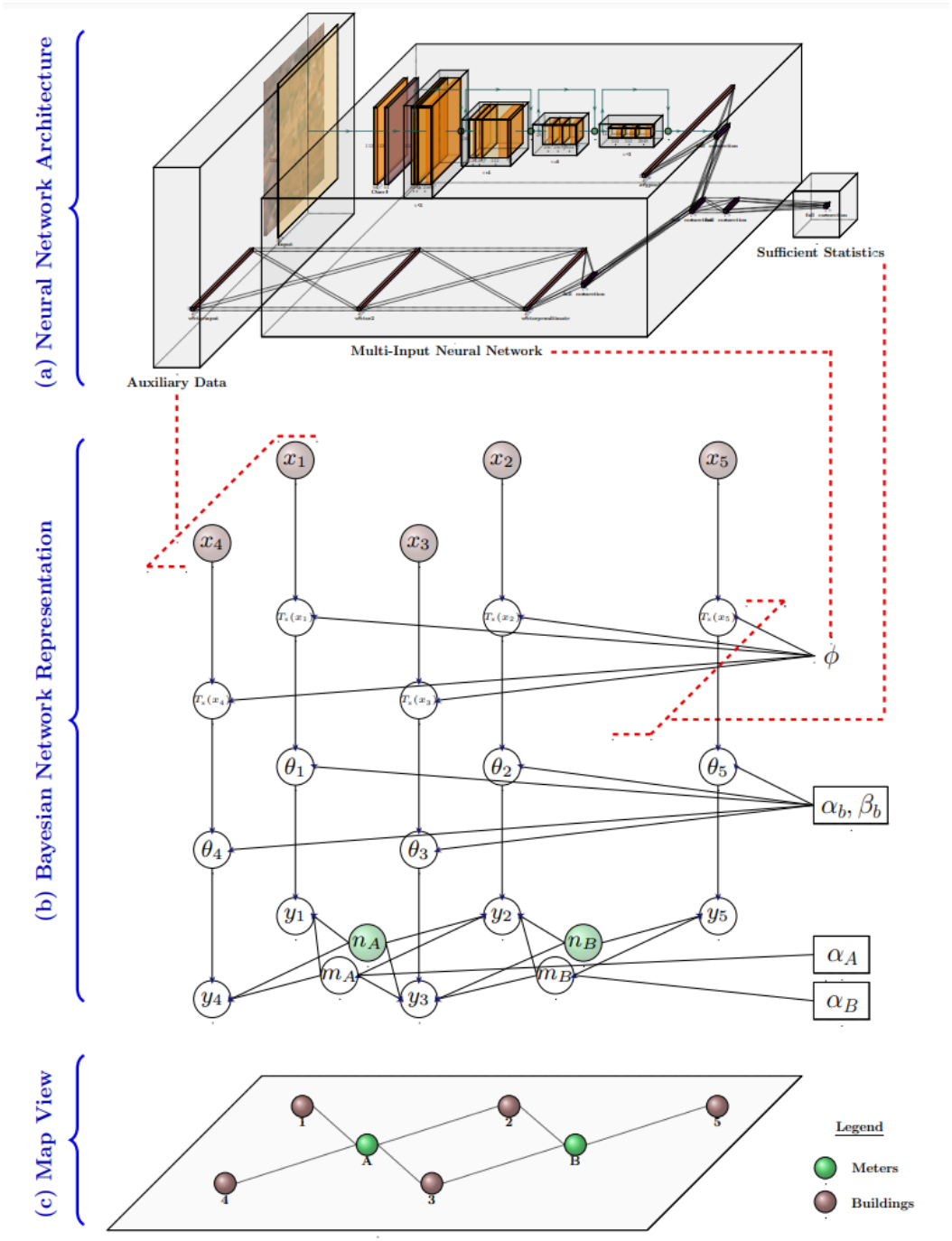
Sample outputs from the LItLDF model, where point estimates for electricity demand correspond to heatmap color values for each building footprint modeled. Any individual building also has an associated probability distribution, here represented as the likelihood of electricity demand at specific discrete values.

The related work in high-spatial resolution electricity demand estimation reveals that while machine learning methods are promising, their application has been limited due to data scarcity and region-specific focus. Studies like those by Heunis and Dekenah[13] in South Africa, and Fobi et al.[14] in Kenya, provide valuable insights but are constrained to characterizing residential load and specific geographic locales.

The LItLDF model is comprised of a multi-input neural network embedded within a Bayesian network for probabilistic data fusion. This dual-component architecture exploits statistical conjugacy for more efficient inference within a Metropolis Hastings inference algorithm, facilitating efficient estimation of electricity demand and uncertainty characterization. Connected subgraphs within the data help delineate complex many-to-many relationships between buildings and meters, enhancing the model's predictive accuracy and relevance for infrastructure planning.

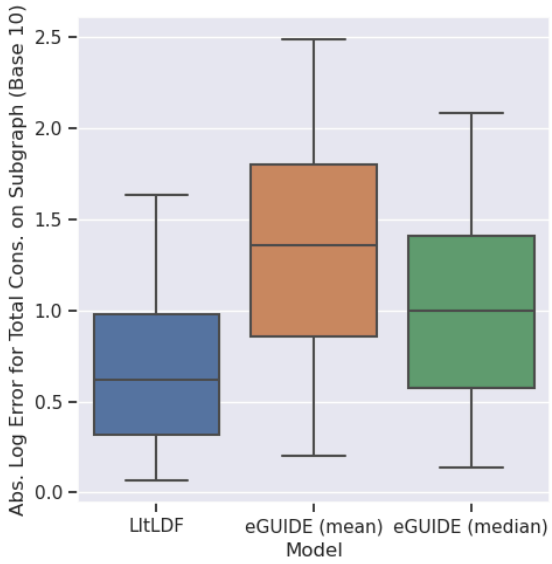
---

[13] Schalk Heunis and Marcus Dekenah. A load profile prediction model for residential consumers in South Africa. In *Twenty-Second Domestic Use of Energy*, pages 1–6. IEEE, 2014.
[14] Simone Fobi, Joel Mugyenyi, Nathaniel J Williams, Vijay Modi, and Jay Taneja. Predicting levels of household electricity consumption in low-access settings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3902–3911, 2022.



Components of the LItLDF model include (a) a multi-input neural network, analogous to that employed in the BEACON model, embedded within (b) a more complex Bayesian network that models potential building-to-meter connections. These statistical relationships reflect modeled proximity values between noisy meter locations and buildings, as depicted in the map view in (c).

In comparative analysis, LItLDF demonstrates superior performance against existing models like those developed by Fobi et al., achieving state-of-the-art accuracy for the application of building-level demand estimation. This advantage is amplified in larger aggregations of buildings, illustrating how stochastic variations in demand tend to cancel out over larger datasets. One innovative aspect of LItLDF lies in its ability to estimate building-level consumption without direct observation of such data. Its use of principles from Bayesian inference offer nuanced demand estimates that outperform less complex implementations.



LItLDF achieves state-of-the-art performance denoted by its lower error metrics relative to baselines.

The LItLDF model represents an advancement in building-level electricity demand estimation. Our hope is that it can be used to enhance decision-making processes in infrastructure planning and resource allocation, ultimately contributing to more efficient and informed infrastructure planning strategies.

6. Output Data

Output data is currently available as geojson files corresponding to grid cells for each available country. These geojson files can be opened using common GIS software and programming language libraries. There are a few notable property names in these geojson files:

- **elec access (%)**: The estimated mean likelihood that a building of interest has grid access to electricity.
- **cons (kWh/month)**: The mean point-estimate for electricity demand
- **std cons (kWh/month)**: The standard deviation for the electricity demand estimate, characterizing estimation uncertainty.

Geojson files also have electricity access distribution parameters: “a_alls_elec” and “b_alls_elec” reflecting the the ***α*** and ***β*** values in the following Beta distribution equation describing building-level access rates:

$$p(x;\alpha,\beta) = \mathrm{Beta}(x;\alpha,\beta) \triangleq \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Lastly, geojson files also have electricity demand distribution parameters: “a_alls_dem” and “b_alls_dem” reflecting the the ***α*** and ***β*** values in the following probability Gamma-Poisson compound distribution describing building electricity demand:

$$p(y \mid \alpha,\beta) = \Pr\{Y=y\} = \frac{\Gamma(\alpha+y)}{y!\,\Gamma(\alpha)} \left(\frac{1}{\beta+1}\right)^{y} \left(1-\frac{1}{\beta+1}\right)^{\alpha}$$

Caveats, Limitations, and Future Work

Inherent lack of information: The main caveat behind our methods is that it’s likely impossible to actually know precisely what building electricity access status’ are and electricity demand solely based on remote sensing data alone. Distribution lines are often too small to see from the resolutions of satellite imagery that are available in our LMIC contexts of interest. That said, it should be noted that any estimate that we provide is just that: an estimate. We aim to provide more utility than simple estimates by characterizing probability distributions and estimation uncertainty and in fact, our relative error decreases as larger and larger clusters of buildings are aggregated during analyses. Nevertheless, estimating building-level access and demand are fundamentally difficult tasks and our methods and pipelines are always going to be subject to error.

Challenging to generalize. Another limitation is that our ground truth metered data is currently largely concentrated in the East Africa region. This leaves model generalization issues open as economies across LMICs vary; thus, it is likely that relationships between remote sensing features and electricity access rates and electricity demand also vary by region. For future work, we are focusing on improving these generalization issues by both working with collaborators to train our models on more metered consumption data and by

gathering and calibrating our models to aggregated ground truth at national and sub-national levels of aggregation.

Class imbalance in training sets. We employ ground truth metered electricity consumption data from different regions reflecting different customer segments in our analysis. Oftentimes, this data is not comprehensive of every electricity connection in a region. It is then a challenge to construct a representative training data set for our machine learning algorithms that are balanced for all customer segments in our regions of interest. Compounding the challenge is that we have very few large commercial and industrial consumers that comprise a major share of demand. Tuning models to not under-assign or over-assign importance on these effects is an ongoing challenge. Additional data and further analysis will help us to improve potential issues from class imbalance.

FAQs and Troubleshooting

- **My Conda environment takes a very long time to solve**. Please try using Mamba to solve the environment. We have found that it works much better than Conda for our specific configurations.
- **Why are some buildings' "std cons (kWh/month)" values higher than their "cons (kWh/month)"?** This can be the case when phenomena are characterized by long-tailed distributions, as is often the case with our demand estimates.
- **Where can I download the data?** Please visit our GitHub data page here: https://github.com/stephenjlee/oemaps-access-and-demand-data
- **How often will you make updates?** Periodically, whenever we make an improvement to the underlying remote sensing features or in the ground truth data that we collect.
- **I can download all of your remote sensing input features except for high-resolution satellite imagery. What's going on?** All of our features are freely available with the exception of satellite imagery. Because of the policies around such data, you are left to procure satellite imagery on your own. Note that the models can be run based on other features; satellite imagery is optional.
- **Something specific about the data doesn't make sense to me**. Please submit a GitHub issue on the data repo and I will try to provide an answer!
- **Something specific about the code doesn't make sense to me**. Please submit a GitHub issue on the appropriate code repo and I will try to provide an answer!