CROSSBOUNDARY

Innovation Insight: Machine Learning to Predict Mini-Grid Consumption
December 2019

CrossBoundary LLC
ABC Place, Waiyaki Way
Nairobi, Kenya

www.crossboundary.com
contact@crossboundary.com

Energy 4 Impact
Mbaruk Road, Opposite Panorama Court
Off Muchai Drive, Off Ngong Road
Nairobi, Kenya

www.energy4impact.org
energy4impact_lab@energy4impact.org

Table of Contents

I. Introduction: the Mini-Grid Innovation Lab publishes the *Innovation Insight* series to share actionable business intelligence on innovations to the mini-grid business model

The Innovation Lab tests innovations that improve the mini-grid business model

CrossBoundary Advisory launched the Mini-Grid Innovation Lab in 2018 with support from The Rockefeller Foundation. The Lab is supported by Energy 4 Impact, who designed and conducted the surveys, collected and cleaned the data used for this analysis, and provided advice on the analysis of results, and by the University of Massachusetts Amherst, Carnegie Mellon University, and Duke University, who provide advice and support on experiment design, survey design, and analysis of results. The Lab prototypes and tests innovations that help mini-grids in Africa provide more power, to more people, at lower cost.

The Lab shares evidence with developers, governments, and funders so they can scale the successful innovation prototypes

The Lab works closely with mini-grid developers to test and identify innovative prototypes that improve the business model, and our work and the results presented here are strongly endorsed by the African Mini-Grid Developer Association (AMDA). Once proven, the Lab works with partners – developers, government, and funders – to scale the prototypes across other developers and markets. The Lab shares evidence on successful prototypes' impact on the business model to inform how partners can best support it to scale.

The Lab launched an *Innovation Insight* series to provide early, actionable business intelligence on initial results from its prototypes

The Lab's *Innovation Insight* series provides ongoing, early insights on the prototypes so mini-grid developers, governments, and funders can act on the results as they emerge. All results and analysis in these series is therefore shared as *actionable business intelligence* rather than scientific evidence.

While these series are not intended to meet the standards of an academic paper, the Lab will publish more complete reports at the end of each prototype, and has partnered with University of Massachusetts Amherst, Duke University, and Carnegie Mellon University to publish academic papers on certain prototypes. Due to differing analytical methods and degrees of academic rigour, it is possible conclusions and interpretations between the two publications may somewhat vary.

II. Executive Summary: matching a mini-grid's generating capacity to electricity consumption is critical to profitability. However, customer surveys do not accurately forecast individual customer demand.

Mini-grids are self-sufficient electricity grids that serve households and businesses isolated from or integrated with the main grid. The Mini-Grid Innovation Lab estimates they are the cheapest way to deliver power to *at least* 100 million people living off-grid in Africa.

One of the fundamental challenges to the profitability of mini-grids is the difficulty of predicting customer demand, and the high costs associated with building a grid with either too much or too little generating capacity. For example, if a 30 kW mini-grid achieves a 12.5% internal rate of return (IRR) when capacity perfectly matches demand, then when *oversized* by 50% IRR drops to 9.2%, and when *oversized* by 100% IRR drops to just 6.1%. Also, when grids are *undersized* and built with too little generating capacity to meet customer demand, developers either miss out on potential revenue, or incur high costs by retroactively expanding the grid or using expensive diesel to supplement renewable electricity generation. One of the Lab's developers recently described accurately forecasting demand at new sites as "the single biggest challenge developers face."

What makes it so difficult to accurately predict consumption at a site? Electricity demand is influenced by numerous characteristics of a community; it is challenging to isolate the relative importance of each of these and how they interact with each other to impact demand. A site's local economy affects its residents' need for electricity. For example, a village on Lake Victoria may have a high demand for charging fishing lights that allow people to fish at night. Demographic and socioeconomic characteristics also affect customers' willingness and ability to pay. Older customers may see less need for electricity, since they've lived so long without it, and may have less disposable income to pay for it.

However, as ever more consumption and payment data is combined with increasingly granular information on site characteristics, machine learning techniques should enable past experience to predict the future. With over 550 million data points on customer behavior across 62 sites in 4 countries, the Lab is uniquely placed to help mini-grid developers use historic patterns to generate accurate estimates of energy demand at new sites.

In 2019, the Lab initiated this project by working with DataKind, a nonprofit which conducts data science projects with mission-driven organizations, to produce a basic model to predict mini-grid customers' consumption from survey responses. To learn which customer characteristics most inform consumption, and how, we combined almost two years of mini-grid consumption and payment data from 31 sites across East Africa with customer survey

responses. The goal was to produce a model which would allow developers to forecast electricity demand using streamlined customer surveys or census data.[1]

The resulting analysis shows each customer's survey responses are *not* predictive of their individual consumption. The Lab's best performing model, using random forest regression, predicted each customer's demand with a 65% error rate – i.e. if a customer's true consumption was 10 kilowatt hours (kWh)/month, the model might predict their consumption to be 16.5 kWh/month or 3.5 kWh/month.

Analysis by the Lab's partners suggests combining all customer-level predictions at a site improves the accuracy of site-level consumption estimates, even if each individual prediction is inaccurate. An overestimate of one customer's consumption could, for instance, be balanced by an underestimate of another's. Given the importance of site-level estimates, the Lab is conducting ongoing analysis to robustly quantify this prediction error.

The minimal predictive power of survey data on customer-level consumption suggests that customers' responses provide either inaccurate or irrelevant data on their individual demand. Calculating each survey response's predictive power using machine learning techniques, the results show the survey variable most predictive of actual electricity consumption is self-reported electricity consumption, with a predictive power score of 12%. The other 300 variables provided negligible predictive power– less than 0.3% on average. To forecast demand with better precision, the model needs accurate data which *is* relevant to consumption.

These findings are in line with developers' own experiences. The Lab's developers have shared anecdotally that they've found predictions for demand based on customer survey data at new sites do not match reality. At one site in Kenya, consumption forecasts based on customer survey responses led to a developer building a 30 kW mini-grid. However, based on actual consumption, the developer estimates they should have built a mini-grid almost three times as large to meet demand. Our findings are also in line with recent research by academics such as Blodgett et al. (2017), Hartvigsson et al. (2018), and Louie and Dauenhauer (2016), which found consumption prediction errors of up to 305% based on survey responses. It is also corroborated by anecdotal evidence from energy access companies outside the mini-grid sector like the solar home system company Mobisol. Mobisol found that customers' responses to questions on household energy usage varied significantly even within the same household. *Asking customers about a product – electricity – they've never, or only recently, had access to, does not appear to be an effective method to accurately forecast their individual demand for that product.*

The Lab's findings hold significance for the entire sector, not just for developers. Surveys are the most commonly used tool by donors and governments to make consumption forecasts

[1]Promising results from preliminary analysis on the predictive power of survey data conducted by the Lab in 2018 motivated fuller analysis on a larger dataset.

which define mini-grid tenders and subsidy programs. Setting tariffs or subsidy levels with poorly predictive survey data may result in uneconomic mini-grid projects.

We acknowledge that conducting site surveys as currently designed may be worthwhile in the absence of an alternative approach. The average cost per customer for the Lab's survey was approximately $8, so running the survey increases the capex per connection by $8, a small figure relative to the average overall capex per connection of approximately $1,000. However, the Lab's financial modelling shows that every 10% in sizing error reduces project IRR by 0.6% – 0.8%. This means oversizing by 50% a mini-grid with a 12.5% IRR when perfectly sized reduces its IRR to 8.5-9.5%. Developers should not be satisfied with relying on any means of predicting customer consumption that reduce IRRs so significantly, particularly as mini-grid returns are already so challenging.

It's also possible that data predictive of site-level consumption could be collected through surveys, but we must better understand what data is most valuable. The Lab's survey collected a catalogue of information on household demographics, income generation, expenditure, asset ownership, energy usage, access to finance, and life satisfaction. Targeting data collection to verifiable variables that have a strong link to demand could lead to better predictions. For example, conducting in-depth interviews with a community's primary business owners on their ability and willingness to spend money on appliances and electricity might be more valuable than surveying all households. A grain mill operator who sells flour to the community typically consumes up to 100x as much electricity each month as a residential customer who watches TV for 3 hours a day. Surveys focused on identifying the number and needs of such 'anchor' customers at a site may be more effective at predicting overall demand than lighter-touch surveys conducted with all households. Analysis conducted by the Lab shows that 75% of electricity at a site is consumed by 25% of customers. The Lab will continue to engage survey practitioners to understand how and if different survey approaches can improve their predictive power.

The Lab's analysis additionally shows substantial variation in electricity usage across sites. The Lab's next step is to dig deeper to identify the differing characteristics of these sites that drive the varying consumption levels and establish how best these can be measured at new mini-grid sites.

We can make two principal observations at this stage:

1. Conducting customer surveys at a prospective new mini-grid site has not to date reliably resulted in accurate estimates of future customer demand. More work on survey design and methodology needs to be done to improve the accuracy of forecasts.
2. Sites do systematically differ from one another in energy consumption. Datasets on site-level characteristics such as density of buildings, distance from main roads, and types of local economic activity can be used to identify other ways in which sites differ and to test the extent to which those characteristics impact consumption.

It is important to emphasize that these initial results represent actionable intelligence rather than scientific evidence. These are preliminary results and may change with more data over time, more data from additional sites and other markets, or alternative analytical approaches. An important caveat is that these surveys were conducted at sites where customers have had electricity for at least six months: results could be different for surveys conducted at sites previously unconnected to a mini-grid. However, we expect surveys at such 'prospective' sites to result in *less* accurate predictions of electricity usage, because customers have even less experience to base their responses on.

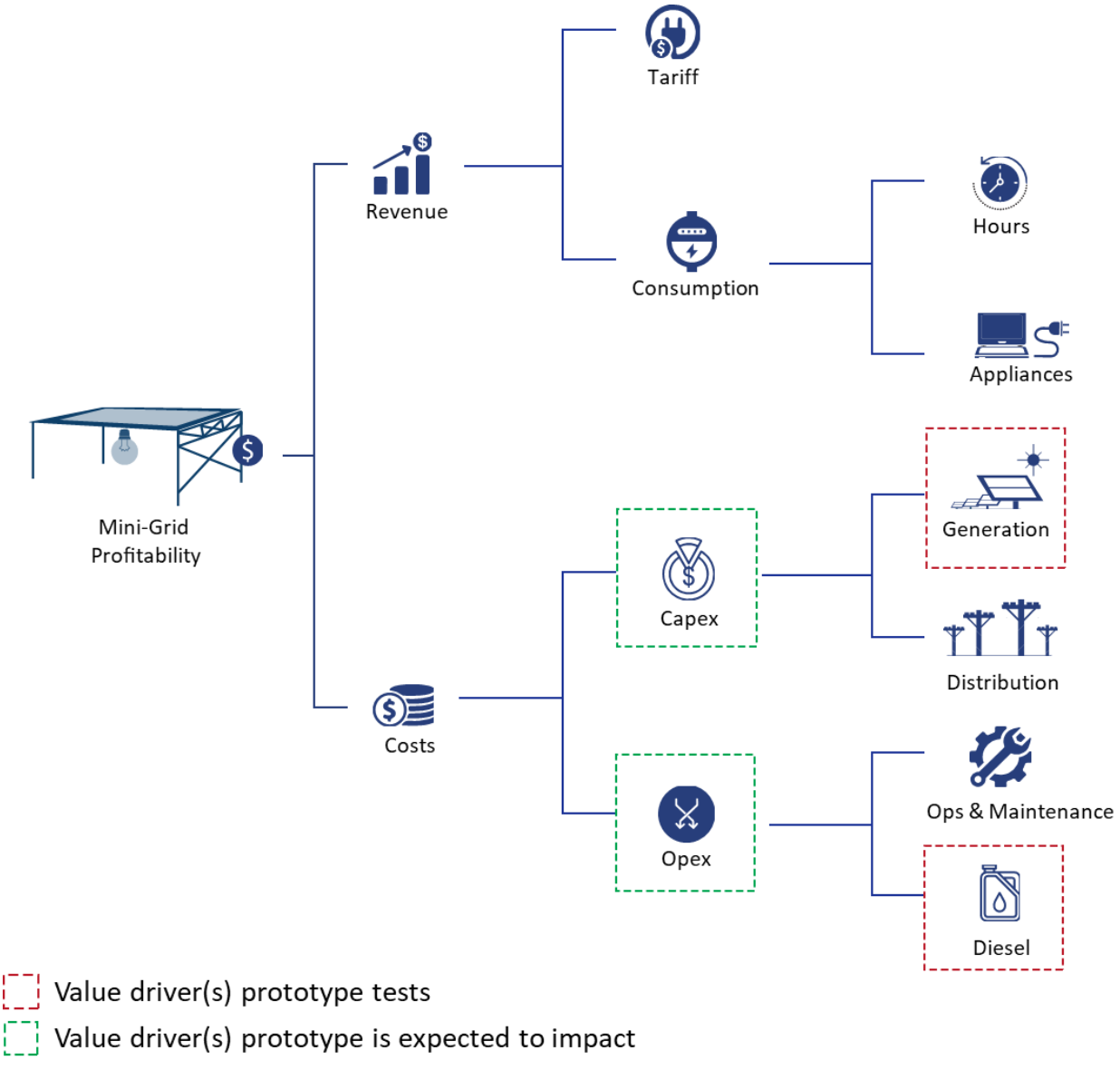
The Lab's goal is to produce a robust model for developers to use that *can* accurately predict mini-grid consumption, trained and validated by combining the Lab's database of consumer behavior with other geospatial datasets, such as maps of roads, grid lines, types of economic activity, and weather patterns.

We are building on the work of mini-grid developers themselves and others in the energy access sector to identify high-quality datasets and the best techniques for demand prediction. The Lab is collaborating with members of the e-GUIDE Initiative, which is led by two of the Lab's academic partners, Jay Taneja of the University of Massachusetts Amherst and Nathan Williams of Carnegie Mellon University, and funded by Rockefeller Foundation. E-GUIDE draws together numerous datasets on historic electricity usage and satellite imagery to make energy consumption predictions, estimate grid reliability, and identify areas where electricity supply could unlock growth in the agricultural sector. The World Resources Institute (WRI)'s Energy Access Explorer has curated a database of quality datasets to map energy access and identify areas with high predicted electricity demand for certain African countries. The World Bank's Energy Sector Management Assistance Program (ESMAP) provides guidance and geospatial tools for electricity demand mapping, and partnered with Facebook, KTH Royal Institute of Technology, the WRI, and University of Massachusetts Amherst to predict the location of the main grid in six African countries. The Lab is building on, referencing, or partnering with these organizations for the tools and analysis they continue to make publicly available.

III. Why we're doing this: right-sizing a new grid to match consumption could be what makes mini-grids investable

The Predicting Consumption prototype tests the impact of accurately predicting customer demand on the mini-grid business model. The Lab expects accurately predicting customer demand will optimize **capex** and **opex**, because developers can avoid over-sizing the grid (**generation**) and using expensive diesel to meet demand in the case of under-sizing the grid (**diesel**).

Accurately predicting consumption is expected to optimize capex and opex.



IV. How we're doing it: combining consumption and revenue data with customer survey responses to see which characteristics most predict consumption

To test this prototype, the Lab worked with DataKind, a nonprofit which conducts data science projects with mission-driven organizations, to use survey data to identify which customer characteristics are most predictive of consumption. The goal was to produce a predictive model developers could use to forecast demand by running surveys at prospective mini-grid sites.

The survey data was collected by the Lab in 2018 through baseline and midline surveys for the Lab's Tariff Reduction and Appliance Financing prototypes. These surveys were conducted in-person with 1,599 households at 31 mini-grid sites operated by 5 developers in Kenya and Tanzania. It took approximately 2 hours to complete each survey. The surveys offered insight into 300 variables on household demographics, income generation, expenditure, asset ownership, energy usage, access to finance, and life satisfaction. The consumption and revenue data was drawn from the same customers' smart meters from early 2017 until January 2019.

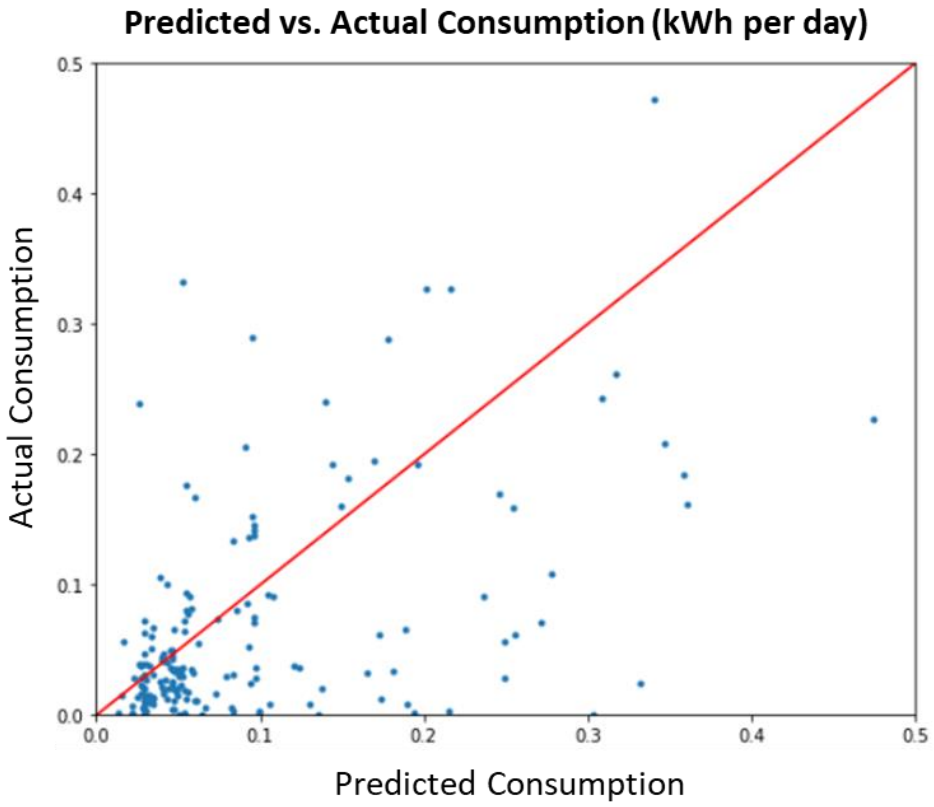
Machine learning algorithms find patterns in data produced in the past and use them to make predictions about the future. DataKind tested a range of machine learning techniques[2] to identify a reliably predictive model of electricity consumption. By splitting the Lab's data into a 'training set' and a 'testing set', DataKind developed the model by training it on one section of the Lab's data, and then tested its ability to make predictions by applying it to the second section. The model's error rate was generated by observing how close the model's prediction of consumption for a customer was to their actual consumption. In doing so, DataKind identified the survey responses with the most predictive power and ranked them in order of importance. Using regression and classification methods, they tested the models' accuracy at predicting an individual's consumption and classifying that individual as a high-, medium-, or low-consuming customer.

[2] The models tested included feature selection, random forest regression, and linear regression. Their performance was assessed using confusion matrices, grid search, and cross-validation techniques.

V. What we're seeing: the model does not generate accurate predictions because customers' survey responses provide little predictive power

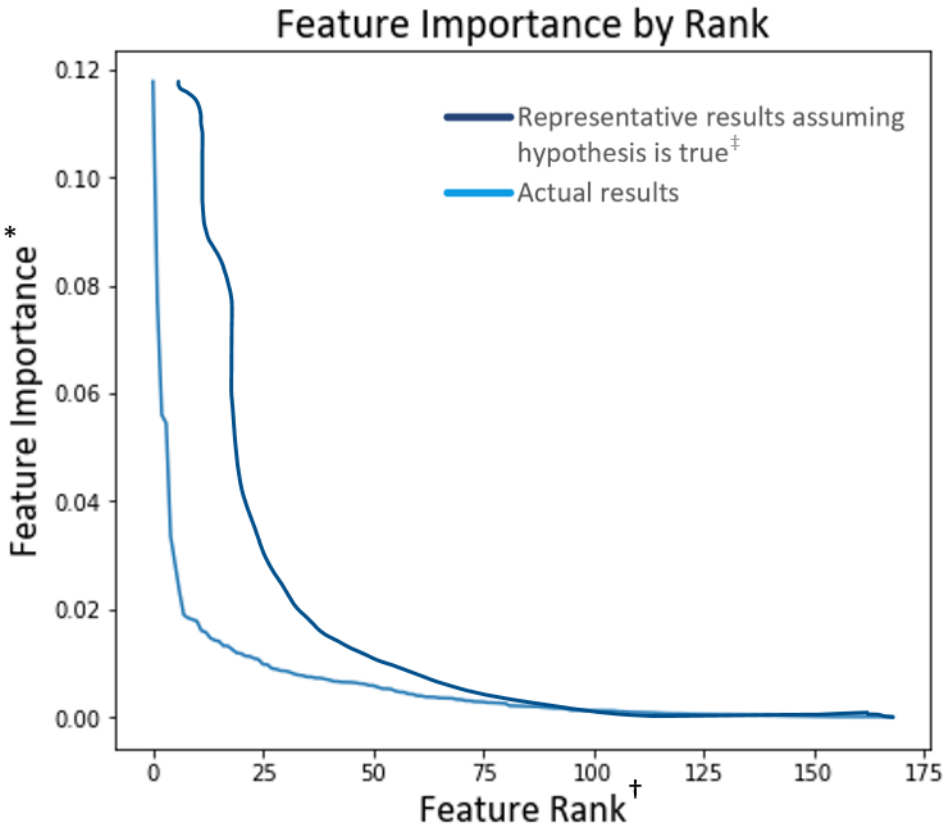
Hypothesis: Developers can predict customer consumption, and therefore right-size new grids, by surveying customers on their spending habits, energy usage, and household demographics.

The Lab's model predicts a customer's consumption with an error rate of 65% when the prediction is compared to their actual consumption.



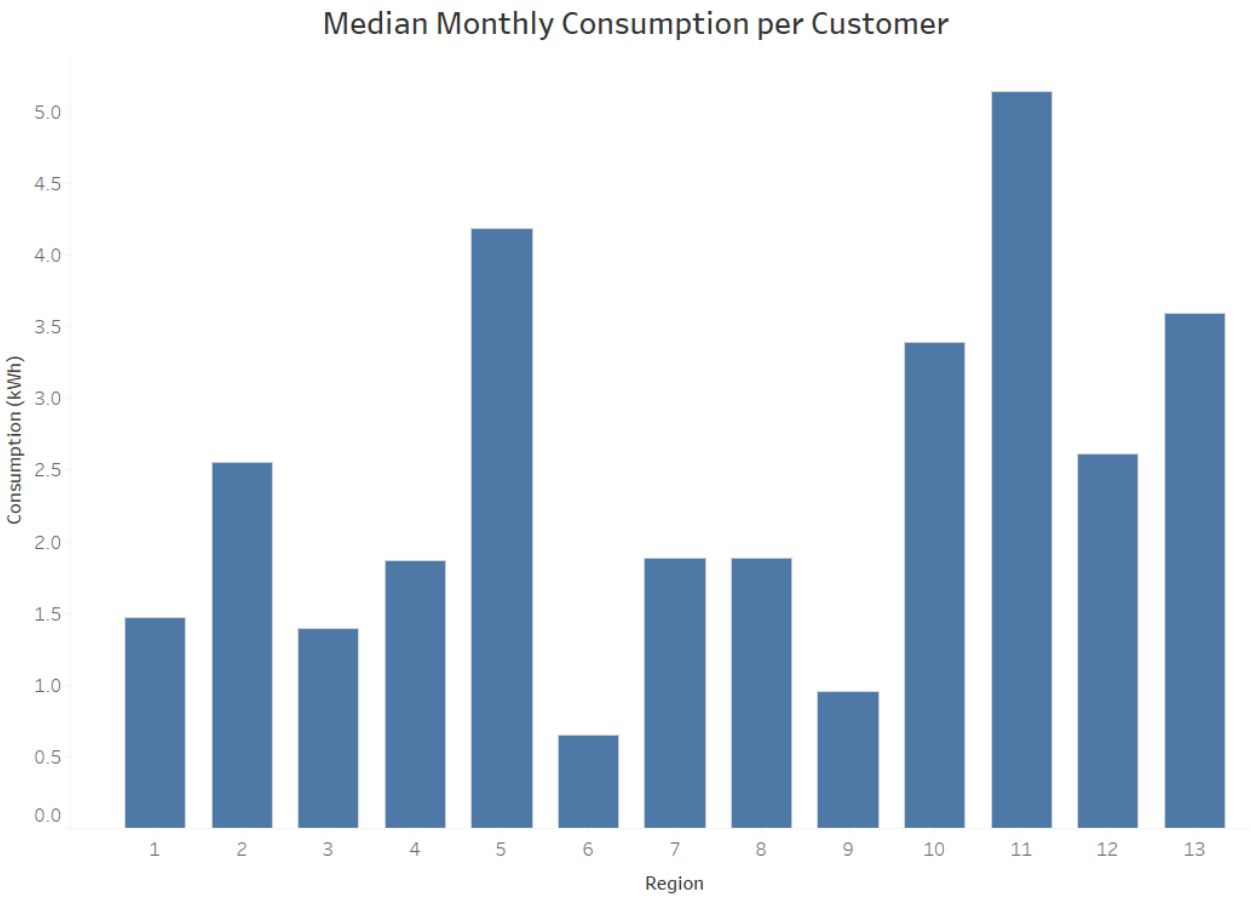
Note: Each blue dot represents a customer. It indicates what the model predicts their consumption to be, relative to what their actual consumption was. The red line indicates perfect predictions: dots falling on this line are customers whose predicted values are identical to their actual values.

The Lab's model is weakly predictive because no survey response variable has meaningful predictive power. The most predictive variable is self-reported electricity consumption, but this accounts for just 12% of the model's predictive power.



*"Feature Importance" is a measure of the predictive power of the variable
† "Feature Rank" ranks each variable in order of predictive power
‡ The representative results line demonstrates the results if 25 variables had meaningful predictive power and could therefore be used to accurately predict consumption.

The Lab's data indicates substantial variation in electricity demand across mini-grid sites and regions, which remains unexplained by variation in survey responses.



What we expected: A variety of customer characteristics, such as household expenditure and income-generating activities, are correlated with electricity consumption. Combining datasets of customer characteristics and energy usage will result in a reliably predictive model. Developers can conduct brief surveys with residents at prospective sites to collect the most important variables and use this information to decide how much capacity to build at a new grid.

What we're seeing: The customer survey data supports only mildly predictive models. Based on 300 survey response variables, the Lab's model predicts a customer's consumption with an error rate of 65%. Calculating each survey response's predictive power using machine learning techniques, the results show the survey variable most predictive of actual electricity consumption is self-reported electricity consumption, with a predictive power score of 12%. However, this can't be provided by customers at prospective sites. The

other 300 variables provide negligible predictive power– less than 0.3% on average.

Electricity consumption varies enormously between regions. The Lab's data shows the median monthly consumption per customer in some regions is up to 6 times higher than other regions *within the same country*. These differences must be driven by some characteristics the model does not account for, such as geographic features and developers' sales and marketing efforts. For example, a fisherman on an island in Lake Victoria is likely to have much more use for electricity than a customer living in a rural location whose primary occupation is growing maize, even if the customers otherwise share a very similar profile in terms of family-size, monthly income, and current spending on firewood. Without the inclusion of additional relevant data, the model cannot produce accurate consumption predictions for new sites.

It is worth noting that tariffs do have meaningful predictive power – when included in the model, they account for 36% of predictive power: lower tariffs are associated with higher consumption. This supports the findings from the Lab's Tariff Reduction prototype which found lowering mini-grid tariffs has a strong impact on overall site consumption.

What it means: Surveys are unlikely to be a reliable way to predict individual customer demand at a new site, and result in prediction errors that have a substantial negative impact on the mini-grid's IRR. Aggregating customer-level predictions to estimate site-level demand or surveying 'anchor' customers to understand their needs may make surveying more effective. To enable mini-grid developers to right-size new grids, thereby optimizing unit economics and serving all interested members of a community, other datasets should also be explored to identify site-level variables that *are* predictive of customer consumption.

Questions we'll answer in a future *Innovation Insight*:

1. What datasets *are* useful for predicting customer consumption? Does site-level data offer more insight than customer-level data?
2. How much can developers improve IRR by right-sizing new grids, and how does that change as the size of the grid changes?
3. Does combining all customer-level predictions at a site increase the accuracy of site-level consumption estimates, even if each individual prediction is inaccurate?

VI. What we're going to do about it: the Lab is investigating what datasets *are* effective in predicting customer consumption

The Innovation Lab improves the mini-grid business model by 1) proving innovations that improve the unit economics for mini-grids and then 2) scaling those innovations with developers and other implementation partners across the continent.

Before scaling, the Lab must identify datasets and build a model that can be used to predict customer demand

	1 Prove	2 Scale
	…the optimal predictive model to accurately forecast consumption and right-size new grids.	**…the predictive modelling tool so developers can better match supply with demand on all new sites.**
Innovation Lab 	***In the next six months…*** • Iterate on our analysis using geospatial data • Test the existing model using improved survey data • Test the accuracy of aggregating individual consumption predictions to estimate site-wide demand • Engage with survey providers on improving surveying methods ***In the next year…*** • Train mini-grid developers to use the model to predict consumption at potential new sites	***In the next year…*** • Measure the improvements in IRR resulting from using the model to right-size new grids
Mini-Grid Developers 	***In the next six months…*** • Collaborate with the Lab to identify the most promising publicly available datasets • Trial early versions of the model to allow rapid iteration and calibration	***In the next year…*** • Use the Lab's model to right-size new grids

Funders

In the next three months…

- Fund ongoing research to calibrate the model using additional datasets and the Lab's customer-level data

In the next year…

- Use the Lab's model to more accurately forecast the IRR of potential mini-grid investments
- Use the Lab's model to more accurately forecast consumption to optimize tender and subsidy program design

Government

In the next year…

- Use the Lab's model to estimate rural customers' energy demand to inform countries' integrated electrification planning
- Use the Lab's model to more accurately forecast consumption to optimize tender and subsidy program design