Big Data
Volume 3 Number 1, 2015
Mary Ann Liebert, Inc.
DOI: 10.1089/big.2014.0061

ORIGINAL ARTICLE

Targeting Villages for Rural Development Using Satellite Image Analysis

Kush R. Varshney,[1,2,*] George H. Chen,[3] Brian Abelson,[1,4] Kendall Nowocin,[3] Vivek Sakhrani,[5] Ling Xu,[6] and Brian L. Spatocco[7]

Abstract

Satellite imagery is a form of big data that can be harnessed for many social good applications, especially those focusing on rural areas. In this article, we describe the common problem of selecting sites for and planning rural development activities as informed by remote sensing and satellite image analysis. Effective planning in poor rural areas benefits from information that is not available and is difficult to obtain at any appreciable scale by any means other than algorithms for estimation and inference from remotely sensed images. We discuss two cases in depth: the targeting of unconditional cash transfers to extremely poor villages in sub-Saharan Africa and the siting and planning of solar-powered microgrids in remote villages in India. From these cases, we draw out some common lessons broadly applicable to informed rural development.

Key words: data mining; machine learning; predictive analytics

Introduction

Although urbanization has been one of humanity's defining storylines in the past century, the World Bank estimates that 47.0% of the world's population still lives in rural areas; this estimate is 70.6% for the least developed countries according to the United Nations classification.[1] Poverty is a serious issue in many rural areas. The International Fund for Agricultural Development (IFAD) recently reported that at least 70% of the world's very poor people live in rural areas, with South Asia having the largest number of rural poor people and sub-Saharan Africa having the highest incidence of rural poverty.[2]

IFAD reports that[2] "rural poverty results from lack of assets, limited economic opportunities and poor education and capabilities, as well as disadvantages rooted in social and political inequalities. … Households fall into poverty primarily as a result of shocks such as ill health, poor harvests, social expenses, or conflict and disasters. … The need to minimize [the possibilities of shocks] undermines people's ability to seize opportunities, which generally come with a measure of risk." Many rural people are stuck in *poverty traps* along several dimensions.[3]

In this article focused on social good, we discuss development activities aimed at uplifting rural populations out of these poverty traps, specifically caused by lack of assets and limited opportunities. In particular, we show how data-driven approaches can enable such programs to be more efficient and effective. We detail two such examples: one addressing lack of assets and the other addressing lack of opportunities due to insufficient infrastructure.

One way to cushion the poor against shocks is via unconditional cash transfers, in which households are given money with no strings attached; recipients are free to pursue their own goals with those funds along whichever dimension or dimensions make sense to

[1]*DataKind DataCorps, New York, New York.*
[2]*Mathematical Sciences and Analytics Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.*
[3]*Department of Electrical Engineering and Computer Science,* [5]*Engineering Systems Division,* [6]*Health Sciences and Technology Division, and* [7]*Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts.*
[4]*Enigma, New York, New York.*

*Address correspondence to: *Kush Varshney, PhD, Mathematical Sciences and Analytics Department, IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598,* E-mail: krvarshn@us.ibm.com

them. Unconditional cash transfer programs have been shown to alleviate poverty in randomized controlled trials because the poor are able to seize opportunities that would have been too risky without the cash transfer.[4]

Another way to cushion rural communities against shocks is by improving diversification opportunities and creating less risky environments through better infrastructure and utilities, including roads, electricity, and water, as well as services such as education, basic healthcare, information and communication technology services, and financial services.[2] Within the infrastructure and utilities category, decentralized energy production is especially effective to alleviate poverty in rural communities.[5]

Investment in rural development requires a good deal of planning. Especially under budget constraints, one is required to select which rural locations, villages, or other jurisdictional divisions to target for development activities in order to optimally utilize resources according to criteria of interest. A major obstacle for the planning of rural development activities is the lack of existing infrastructure. Many of the rural poor live in remote villages without access to roads and communication services. This presents logistical and financial challenges to gathering data across large numbers of villages in a region. It is precisely this information that is essential to the targeting and planning of rural development activities.

One option to overcome the logistical challenges of acquiring data from people on the ground is *remote sensing*. Remotely sensed satellite imagery can be analyzed from the rural development perspective to yield actionable insights for planning. In this work, we describe two sample cases of satellite image analysis for prioritizing villages for different types of rural development; our overall intention is to put forth a general paradigm for using remote sensing for social good in the rural setting.

The first case we detail is the selection of villages to receive unconditional cash transfers in East Africa by the nongovernmental organization (NGO) GiveDirectly.[6] The criterion is to select the poorest villages as indicated by the types of roofs of households, which are distinguishable in satellite images. Poorer households tend to live under thatched roofs and less poor households under metal roofs; thus, GiveDirectly would like to operate in villages with the largest proportions of households living under thatched roofs. The second case we detail is the selection of villages to receive investment in microgrid and renewable energy generation development in India by the NGO SELCO Foundation. The objective in this case is to determine the optimal grid layout and design in villages based on the locations of houses, other buildings, and roads extracted from satellite imagery. Then villages can be ranked by the cost-effectiveness of the design.

These two cases of rural development are not unique. One of the classical problems in remote sensing, the determination of land use and land cover,[7] can be used to inform agricultural decision making, quantify deforestation, and perform other tasks for the benefit of the rural poor. Night-time illumination imaging provides valuable data for estimating access to electricity (or lack thereof).[8]

Examples more similar to the two rural selection problems mentioned above include selecting locations for the extraction of potable water,[9] selecting sites for the placement of microscale hydropower stations,[10] and selecting warehouse locations and villages to be served by logistics companies for the delivery of goods ordered via e-commerce. Another related problem is the selection of paths over rural areas to be served by Internet-service-providing high-altitude balloons such as Google's Project Loon.

The remainder of the article is organized as follows. In the section Satellite Imagery, we discuss the availability and characteristics of satellite images of rural areas. We describe the GiveDirectly unconditional cash transfer case in the section Targeting Unconditional Cash Transfers, followed by the SELCO Foundation electrification case in the section Planning Rural Electrification. In Discussion and Conclusions, we discuss some of the common lessons we can draw from deploying these types of solutions and conclude.

Satellite Imagery

Several satellites in orbit above the Earth are used for acquiring images of the ground. Most are very expensive endeavors, but a recent trend has seen startup companies such as Skybox Imaging, Planet Labs, and UrtheCast launch low-cost satellite imaging solutions.[11] The usefulness of satellite images can be examined along several dimensions of satellite design characteristics and acquisition conditions, such as spatial resolution, temporal resolution, spectral range, spectral resolution, cloud coverage, time of capture, and off-nadir angle. Finer resolutions and a larger spectral range tend to incur higher costs of acquisition. It is possible to more easily distinguish foliage if the spectral range includes infrared frequencies. A large off-nadir angle impacts the visibility of the sides of buildings rather than just

their tops. Clouds are opaque except at radio frequencies, and so typical satellite images taken of cloud-covered areas do not show anything of use. Very low temporal resolution may result in images of different areas being acquired in different seasons of the year, in which visual characteristics of the ground are different.

Keeping the costs of imagery down is particularly important in social good endeavors. Fortunately, satellite imagery is generally available for free or at low cost with enough quality to be useful for tasks such as the ones described in the sequel, including areas such as India and sub-Saharan Africa. For example, the Google Maps API allows the free downloading of visible-light satellite images with enough spatial resolution to distinguish buildings; with the startup companies joining the fray, the field of satellite imaging is poised to be democratized. Free or low-cost imagery sometimes lacks quality in some dimensions, such as images available in adjacent regions from different seasons or images taken with large off-nadir angles. Solutions need to be robust against these issues, for example, by building different machine learning models for images taken in different seasons or by sufficiently rectifying images before further processing. Free imagery may also suffer from image compression artifacts, but such artifacts have been shown to not significantly degrade machine learning tasks encountered in remote sensing.[12]

Targeting Unconditional Cash Transfers

The first case we describe is a pro bono satellite image analysis project developed through the DataKind DataCorps in conjunction with GiveDirectly, an NGO that aims to help people living in extreme poverty by making unconditional cash transfers to them via mobile telephony.[6] Unlike other charitable giving, unconditional cash transfers do not presuppose that livestock, training programs, or any other items are best for the recipients. Evidence from several randomized control trials shows that this method of charity has large positive effects on multiple measures of recipients' well-being.[4] Additionally, by forgoing large intermediary infrastructures, donations can be used extremely efficiently; to date, more than 90% of each donated dollar has reached a recipient. In this section, we first provide a summary of the project and then provide more technical details.

Summary

Currently operating in Kenya and Uganda, GiveDirectly takes an end-to-end operations model comprising several steps and does not outsource or subcontract work to other organizations. First, funds are solicited from individual and institutional donors. Once a sufficient amount has been collected, the NGO initiates a campaign to disburse the donations to the extremely poor. The first part of a campaign is enrolling extremely poor households to be recipients; the second is transferring funds to recipients via mobile money systems; and the third is conducting telephone and in-person follow-up with each recipient.

Targeting extremely poor households starts by identifying broad regions of the country with high poverty rates through data from the national census (which does not have the granularity to make village-level determinations of poverty). Within those regions, villages and households are selected through a transparent criterion that is associated with extreme poverty. In Kenya and Uganda, where the NGO currently operates, the extremely poor tend to live in homes with thatched roofs, whereas the less poor tend to live in homes with metal roofs. Villages with a large percentage of low-quality housing (as well as access to a mobile money agent) are targeted. After villages and households within villages have been selected, money is transferred directly to recipients' mobile accounts.

The existing approach to village selection involved multiple rounds of site visits because governmental or other poverty data do not exist at village-level resolution. In this project, we developed a remote sensing system that estimates the quality and density of housing in large areas of Central East Africa to facilitate village selection, thereby eliminating the need for the most expansive and expensive first round of site visits. Note that although other data and factors can help refine the indication of poverty, GiveDirectly uses roof type as the only indicator and thus it is the only indicator we consider in our approach. Specifically, we applied a combination of image processing and machine learning methods on free satellite imagery from Google Maps, trained on crowdsourced data, to build a regression model for the number of houses of different roof types in an image patch. The machine learning model allowed us to construct heat maps of housing density and thatched-roof proportion from individual image patch-level estimates and aggregate these estimates to produce village-level proportions of thatched roofs, resulting in a ranked list of villages to target.

The data-driven prioritization of villages was initially deployed in three districts in Kenya in February 2014. In total, the villages selected through the proposed approach received over 4 million U.S. dollars in direct

cash transfers. This satellite image analysis approach saved approximately 100 person-hours of effort that would have been incurred as a result of manual village selection methods.[6]

Technical Details

As discussed above, our objective was to estimate the number of thatched roofs and the number of metal roofs in small image patches in East Africa, or equivalently the total number of roofs and the proportion of roofs that are metal. Example satellite image patches from the Google Maps API are shown in Figure 1. The human visual system is able to distinguish metal roofs from thatched roofs because the metal roofs are bright white or gray rectangles, whereas thatched roofs are brownish with less crisp and slightly less straight edges.

We wanted to learn machine learning models for the two variables of interest because it would have been too cumbersome for human annotators to manually label the approximately 50,000 image patches needed to cover the three Kenyan districts in the pilot campaign. However, it was not unreasonable to use human annotators to produce a training set. Therefore, over one weekend, we enlisted 10 volunteers to label approximately 1,500 image patches through an interactive crowdsourcing platform we developed that allows the user to click on roofs and label them as metal or thatched. A screenshot of the application may be found in Ref.[6]

All roofs in the region are approximately the same size and shape, and have consistent appearance conditioned on the type of roof. Therefore, we used template matching[13] with two different small templates to locate all roofs in an image patch: one completely white and the other a typical thatched roof center from one of the training images. Template matching algorithms have a threshold value as a parameter; since our ultimate objective was to derive features for regressions, we did not fix a single threshold but took counts of roofs in the image patch at several different threshold values as features.

Moreover, we trained a random forests classifier based on the distribution of colors in a small area surrounding the roof locations in the training set to distinguish metal and thatched roofs, and applied it to obtain soft classifications for each roof location obtained via the template matching. Each pixel is represented by three numbers: a red, a green, and a blue intensity value; the color distribution is simply the set of three histograms constructed from the intensity values in the small area. The average roof type classification score at each template and threshold provided us with another set of features for the ultimate regressions. We also used the color distribution values of the entire image patch as features.

With these three types of features and the response variables from the crowdsourced data, we trained random forests regressions for the total number of roofs in the image patch and for the proportion of roofs metal

FIG. 1. Sample satellite image patches of the region with **(a)** thatched roofs in the center and **(b)** a metal roof in the center.

in the image patch. We found that separate classifiers in the feature extraction and separate regression functions for different satellite acquisition conditions (i.e., hazy, wet, and dry) yielded better accuracy than the same models for all conditions due to significant variation in visual appearance of imagery under the different conditions.[6]

We obtained a 10-fold cross-validation mean absolute error of 1.95 on the number of roofs and 0.162 on the proportion metal. While our final step was to aggregate individual image patches to village-level proportion thatched, we did not have ground truth labels for images of entire villages, and so we characterized the error as follows. Villages cover an area approximately 150 image patches in size, which is also the number of image patches in the testing partition of each fold of the tenfold cross-validation. Therefore, we averaged the estimated aggregate proportion thatch in the test partition and compared it to the true aggregate proportion thatch, and averaged across the 10-folds. The resulting absolute error of the "village-level" proportion thatched was 0.020, which is 2.87% when expressed as an absolute percentage error. The village-level error was much smaller because the individual image patch estimates were fairly unbiased.

Using the learned random forest regressions, we scored all three Kenyan districts of interest to obtain the heat maps in Figure 2. We aggregated the estimates within village boundaries to obtain the village-level results shown graphically in Figure 3, which can also be presented as a ranked list of villages for GiveDirectly's targeting. Obtaining geographic coordinates for village boundaries in order to carry out the aggregation, as shown in Figure 3, was nontrivial. Processing the only source of these boundaries, a collection of pdf files, required significant manual effort.[6]

In the course of the February 2014 campaign, the GiveDirectly staff members who visited the targeted villages to enroll individual households also manually collected statistics on household roof types. In comparing the village-level proportion thatched values obtained from the satellite image analysis to the ones obtained on the ground, we found that the estimated values were much larger than the values obtained by staff members. We also found that the correlation between the two sets of values was not as high as we would have liked. In analyzing these observations further, we found the root cause to be the fact that households may construct thatched-roof structures separate from the main house to be used as kitchens, sleeping quarters for sons who have reached puberty but have not yet married, and so on. The algorithm counted each of these structures as a separate roof, whereas the humans on the ground only counted the roofs of the main houses. This issue presents an area for future research.

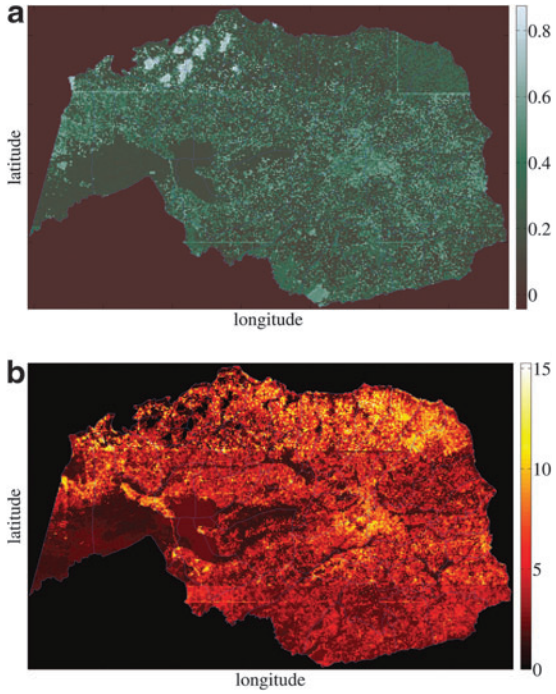

FIG. 2. Heat maps of three districts in Kenya showing **(a)** estimated proportion of roofs that are metal, and **(b)** estimated number of roofs per 400×400 pixel image.

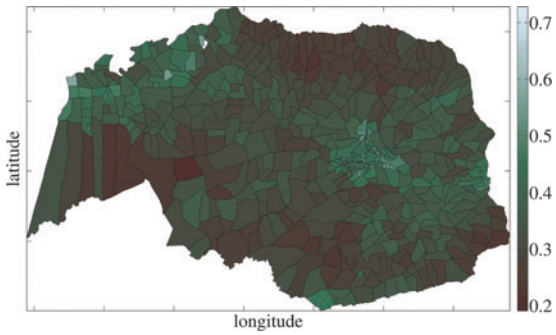

FIG. 3. Estimated proportion of roofs that are metal per village.

Planning Rural Electrification

The second case we describe is a project initiated through the 2014 MIT IDEAS Global Challenge and in partnership with the SELCO Foundation, an India-based NGO that focuses on energy access for underserved communities. Of India's 1.2 billion citizens, approximately 300 million (or 25%) lack access to electricity.[14] Although most urban poor have access to grid power, although unpredictable and erratically priced, rural Indians have no such hope of being connected to the larger network due to their distance from primary infrastructure, lower population densities, and exorbitant cost of power on a per kWh basis due to transmission losses from physical distance and stealing. For these reasons, rural Indians (70% of the total population[15]) often have the greatest need but fewest alternatives beyond biomass and kerosene. Unfortunately, neither of these offers the developmental promise that electricity does in the form of education, communication, or manufacturing. Of the few electrification options available, one of the most exciting opportunities is the creation of stand-alone village-sized microgrids powered by local generation and distributed among the participants. Such distribution grids are exciting not only because they allow for more efficient use of renewably generated energy but also because they are designed to operate independently from any larger grid and therefore are unaffected by the remoteness of a site.

Summary

In spite of the excitement surrounding microgrids, there have been relatively few successful long-term players in the market as a result of the significant challenges faced during implementation. Compared to first-world grid development, microgrid companies face distinct challenges that greatly hinder their ability to succeed—challenges that we believe can be surmounted by data analytics. First, most microgrid companies face large knowledge gaps pertaining to village geographies and topologies. Little to no information about the village's size and building layout exists prior to project selection, and as a result village identification and modeling is mostly done via arbitrary personal connections and ad hoc paper-and-pencil planning methods. Compounding this problem is the fact that most rural power companies have historically focused on solar home lighting (SHL) systems as their main tool for electrification. These systems are installed house by house and operate in a standalone manner from the rest of the buildings in the village. Although several companies have found solvent business models deploying such systems, they are not likely candidates for impacting the much larger multimillion person electrification challenge due to the amount of time and human capital it would require to go house by house. A better method would be to leverage existing repositories of satellite imagery from which we can automatically map villages and extract valuable house-level data.

Another failure mode of many microgrid companies is a distinct inability to scale up power supply for existing consumers and to expand to future sites. Microgrids are often designed to provide a particular level of service at minimal cost, and as a result the hardware (converters, wire gauges, etc.) lock consumers in at maximum power levels in spite of the fact that once electrified, per-capita demand tends to grow quickly. With digital village modeling, this evolution in demand can be statistically modeled to assist decision making. At a larger business model level, companies currently rely on site-by-site tailored approaches for electrification and as a result are unable to translate success from one location to another and unlock the enormous market potential. Modeling what characteristics make locations similar and suitable could improve a company's ability to transfer success across locations.

We propose to empower microgrid and renewable generation development in rural India through a software tool that rapidly identifies, digitizes, and models rural development sites. Unlike SHL systems that install electrification systems at the individual house level, microgrid installations demand a greater degree of planning as an increasing percentage of the cost–performance nexus is functional on the extent of wiring and sizing of shared generation and storage systems. In this respect, microgrids have significant spatial challenges that do not exist in more simplistic rural electrification solutions. To address this, our tool will enable microgrid developers to effectively and efficiently analyze rural sites and plan infrastructures most optimal for long-term growth through three main ways:

1. Site identification and selection: India has vast numbers of unique rural areas without electricity. These sites vary substantially in their topologies and housing profiles, both of which strongly influence microgrid planning for a specific site. In order to provide these key parameters, the first part of our tool rapidly analyzes satellite images and automatically extracts villages and building

features. These hundreds of digitized villages can then be rapidly compared on the basis of population, size, density, and shape.

2. Site modeling and optimization: Once a site has been located and its features have been analyzed, grid developers need to design an energy infrastructure overlay. Our software models the village buildings and simulates various grid topologies with user-input cost functions. This allows developers to select the most optimal design specific to the site.
3. Site expansion planning: Once electrified, demand tends to grow in time as a result of expanding individual usage profiles as well as village development. Using past and present satellite images of electrified villages, our tool projects the growth of a village following initial electrification to allow designers to oversize bottleneck hardware and allow for grid growth in time.

Technical Details

Unlike the satellite imagery analysis for targeting unconditional cash transfers where the goal was to estimate the number of thatched and metal roofs, in planning rural electrification, we aim to estimate the cost of installing a microgrid in a village with the help of information on locations and shapes of buildings in the village. In contrast to the Google Maps satellite imagery used in the GiveDirectly example, we used here commercial satellite imagery that included near-infrared measurements as well. At a high level, our tool has two steps. In the first step, it identifies where buildings are in satellite images of villages. In the second step, it simulates thousands of different microgrid wiring topologies per village and estimates the cost for each wiring topology based on the actual costs incurred by our partner organizations. This second phase of work benefits from an iterative process of comparing projected costs with the actual costs for former projects. We describe these two steps next.

Building detection. It would be an immensely time-consuming task to manually identify the large number of houses necessary to our desired analysis. However, we can manually label a few villages and learn a model of buildings from these labeled villages.

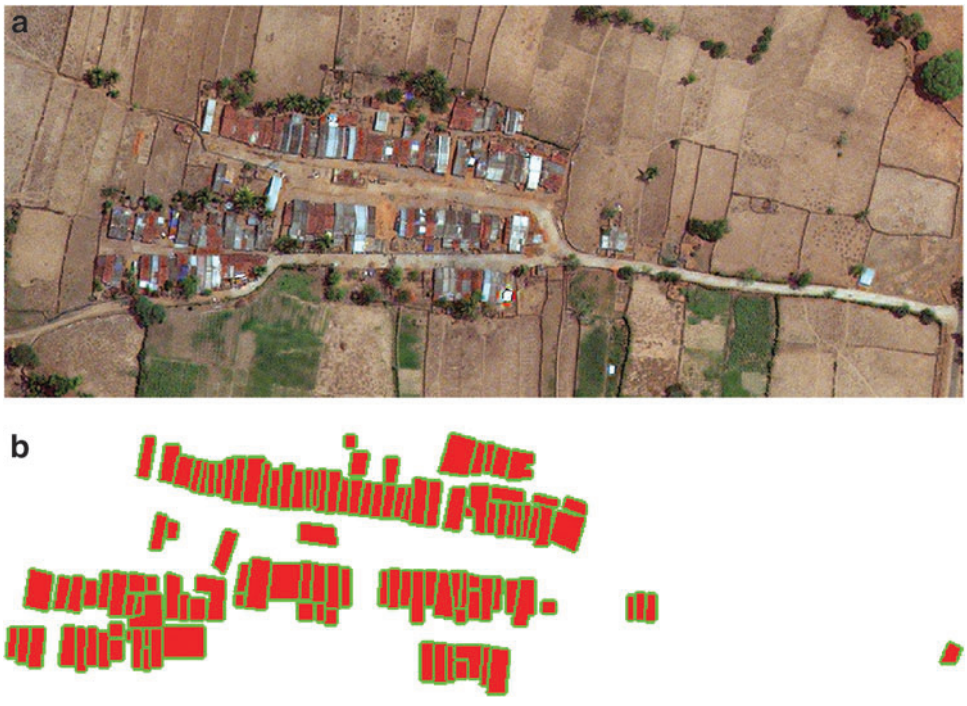

FIG. 4. **(a)** Sample satellite image of a rural village in India. **(b)** Ground truth label of the village where red indicates "building," white indicates "not building," and green indicates "not sure," which includes pixels very close to the building boundaries.

We manually labeled 596 buildings across 10 villages in Rayagada, Orissa, India, to serve as ground truth. In particular, for each satellite image of a village, the ground truth data consist of another image that indicates for each pixel in the satellite image whether it is part of a building, not part of a building, or "not sure." An example satellite image of a village and its corresponding ground truth label are shown in Figure 4.

To learn a model of buildings, we begin by partitioning a satellite image of a village into small regions using a seed-based region-growing image segmentation algorithm that groups together adjacent pixels of similar color. The algorithm also uses a recent boundary detector[16] that can parse various instances of weak color gradients between a building and the ground surrounding it. An example image segmentation result is shown in Figure 5a. Next, for each image region, we compute

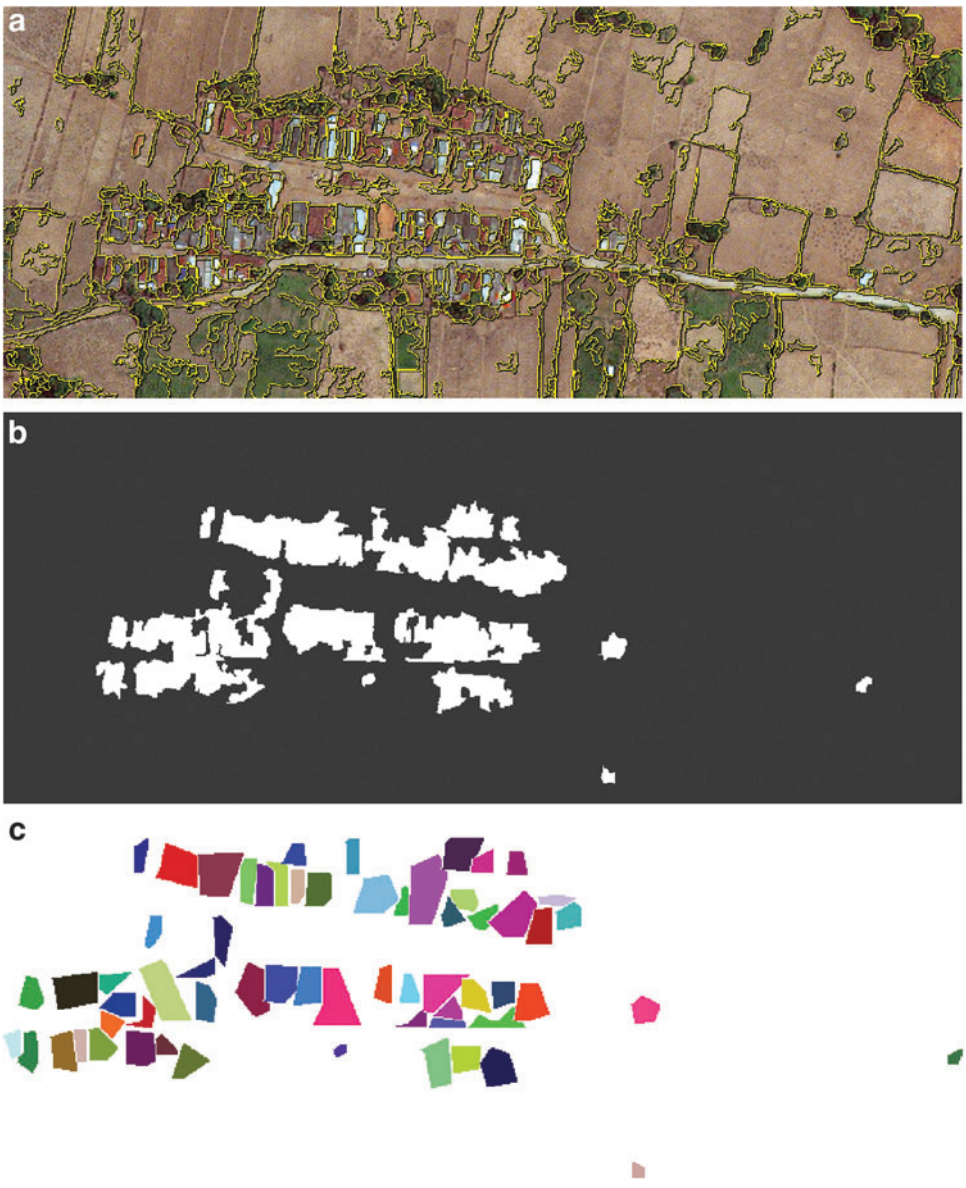

FIG. 5. For the village in Figure 4, we show results of our data processing: **(a)** segmentation of the satellite image, **(b)** estimate of which image segments are part of a building, and **(c)** polygonization of estimated building segments.

a feature vector, which describes that region's basic color and geometry information as well as some ratios that may be indicative of either building or vegetation. Specifically, the basic color and geometry information we include in the feature vector are the region's color distribution (of pixels within the region), area, and diameter. We also include each region's ratio of perimeter to area (gives a coarse measure of polygon complexity), standard deviation of grayscale values (tree regions tend to have noisier grayscale values than building regions), the normalized ratio G/(R+G+B) (vegetation regions tend to have a larger fraction of green compared to all colors), and the R/NIR ratio (this ratio has previously been found to be indicative of vegetation[17]). For each region, we can obtain a label for whether it is part of a building or not by looking at what the most popular ground truth label is among pixels in the region. Regions with a most popular label of "not sure" are ignored. Finally, we train a random forest classifier on image regions, each of which is now represented as a feature vector.

To estimate the location of buildings for an unlabeled satellite image of a village, we again begin by segmenting the image into regions and computing feature vectors per image region. Then we use the trained random forest classifier to determine whether each region is part of a building. We can assemble the classification results per region into an image that shows where the buildings are estimated to be, as shown in Figure 5b. To evaluate the quality of such an image of estimated building locations, we compute the Dice coefficient[18]—a measure of overlap where 0 is no overlap and 1 is perfect overlap—between the detected building pixels and the ground truth building pixels. Specifically, for two sets A (of pixels declared to be buildings by our algorithm) and B (of pixels that are part of buildings according to ground truth), the Dice coefficient is given by $2|A \cap B|/(|A|+|B|)$, where "$|\cdot|$" denotes set cardinality and "$\cap$" denotes set intersection. Using this evaluation criterion, we obtained a leave-one-out cross-validation mean Dice coefficient of 0.66. Importantly, this error metric is computed based on pixels rather than regions. Measuring error based on regions we computed does not account for the arbitrariness of the choice of segmentation algorithm used to compute regions. Another algorithm could, for instance, either use different regions or not use regions at all in order to estimate whether each pixel in an unlabeled image is part of a building or not. Regardless of what algorithm is used, the Dice coefficient above can be used to evaluate the quality of detecting buildings in an image.

Once we have an estimate of where buildings are at the pixel level, we can fit polygons for where different buildings appear. An example polygonization is shown in Figure 5c. Unfortunately, many of the villages we encounter have buildings of similar rooftops and color lumped together, such as many of the buildings in the first row of the satellite image in Figure 4a. Our polygonization process is currently unable to tease apart such adjacent buildings. The clumping of like-roofed buildings can affect the accuracy of the load and cost estimate by misrepresenting the building's size. For example, if two small houses are clumped together and recognized as one building, it would be estimated as a single larger structure, such as school or barn, which would have different power requirements than the two small houses. A user could manually edit the polygons to obtain a more accurate result, in which case the automatic building detection could be viewed as providing a helpful initial guess, which mitigates the amount of manual labor needed. Even without manual editing, the simulations of wiring topologies using the automatic building detection yielded useful ballpark estimates for the cost of installing microgrids in a large number of villages.

Simulating wiring topologies to estimate cost of installing a microgrid. Once buildings have been polygonized (Fig. 5c), the village can be run through various simulations to identify optimal wiring configurations of a chosen topology. We run Monte Carlo simulations for one to "x" (user defined) number of generation

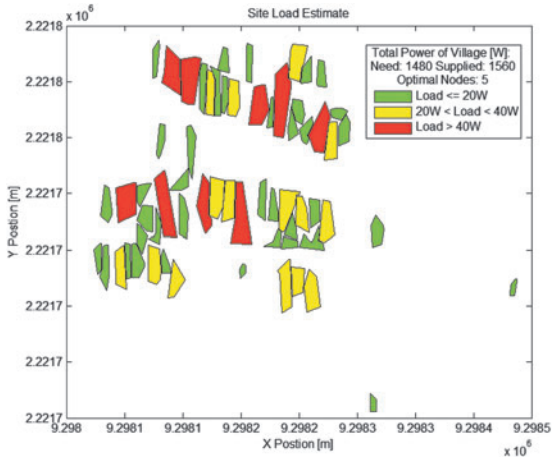

FIG. 6. For the village in Figure 4, we show the results of the load estimation model.

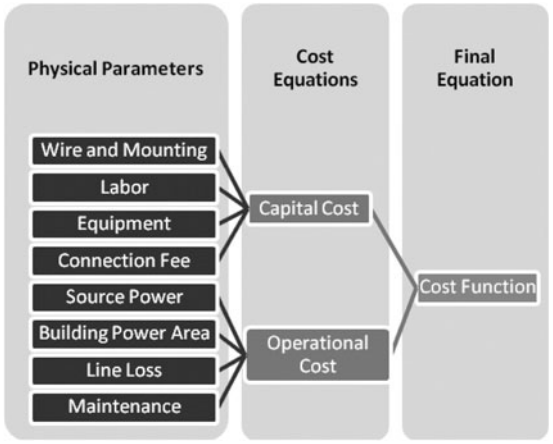

FIG. 7. User-defined and inputted cost factors that can be used to develop the total cost of the village.

nodes through a detailed cost function to simulate thousands of wiring schematics and their respective costs. The wiring topology can be defined by the user, and in this case a modular interconnected "hub and spoke" approach (each house is a spoke to a hub generation node that is interconnected to other hub generation nodes). The simulation can identify the configuration with the greatest savings in cost (wiring, components, labor, total cost, etc.) at a specified level of service per customer.

An input file of the microgrid parameters (voltage, labor rates, component costs, etc.) with information on the distribution and sizing of the various buildings in the village is used to run the simulation. First, the approximate power draw from each structure is estimated. We do this by using area as a proxy for total number of inhabitants and also modulate the demand on the basis of the proximity of the house to the centroid of the village and its number of nearest neighbors. These variables provide load estimates per house that are then validated against the experiences of the microgrid partner, in this case SELCO. Figure 6 provides a load estimate calculation for each house in the village.

With approximate load draws, we are then able to sample thousands of potential hardware placements for solar generation and battery storage locations. These locations act as important nodes within the wiring network and therefore influence variables like the total amount of wire, resistance loss, and labor implicit in a particular design. In particular, the cost equation for the village takes into account many variables, examples of which are outlined in Figure 7.

Each permutation of the number and placement of storage/generation nodes corresponds to a specific

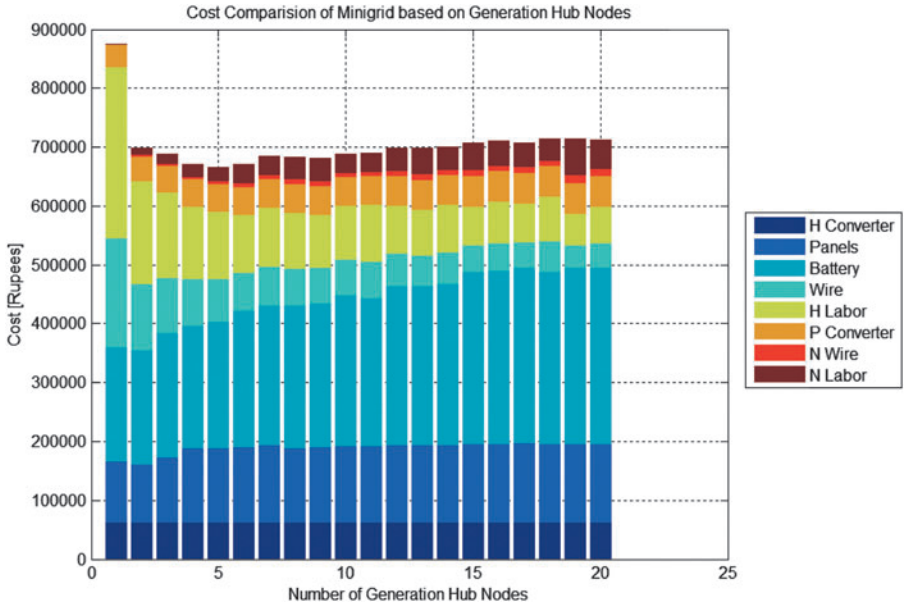

FIG. 8. For the village in Figure 4, we show the results of the minimized total cost for the Monte Carlo runs of each hub generation node from 1 to "x" ($x = 20$).

cost range. These variables are sampled in parametric space via a Monte Carlo method and the wiring costs are estimated via straight-line distances between points mated with a minimum spanning tree to connect a higher voltage-level architecture between nodes. Figure 8 shows the minimization curve of the total village cost for a given number of hub generation nodes.

Once complete, the algorithm selects for the cheapest configurations and presents approximate per house costs for project deployment. Figure 9a shows an example of the projected cost of the village in its optimal configuration with wiring based on the extracted buildings. This is compared with the wiring results based on the ground truth buildings (Fig. 9b). The number of hub generation nodes and average cost per house obtained from the extracted buildings are comparable to those obtained from the ground truth buildings. This further validates the building detection algorithm.

These simulation results provide microgrid designers and implementers with critical information for

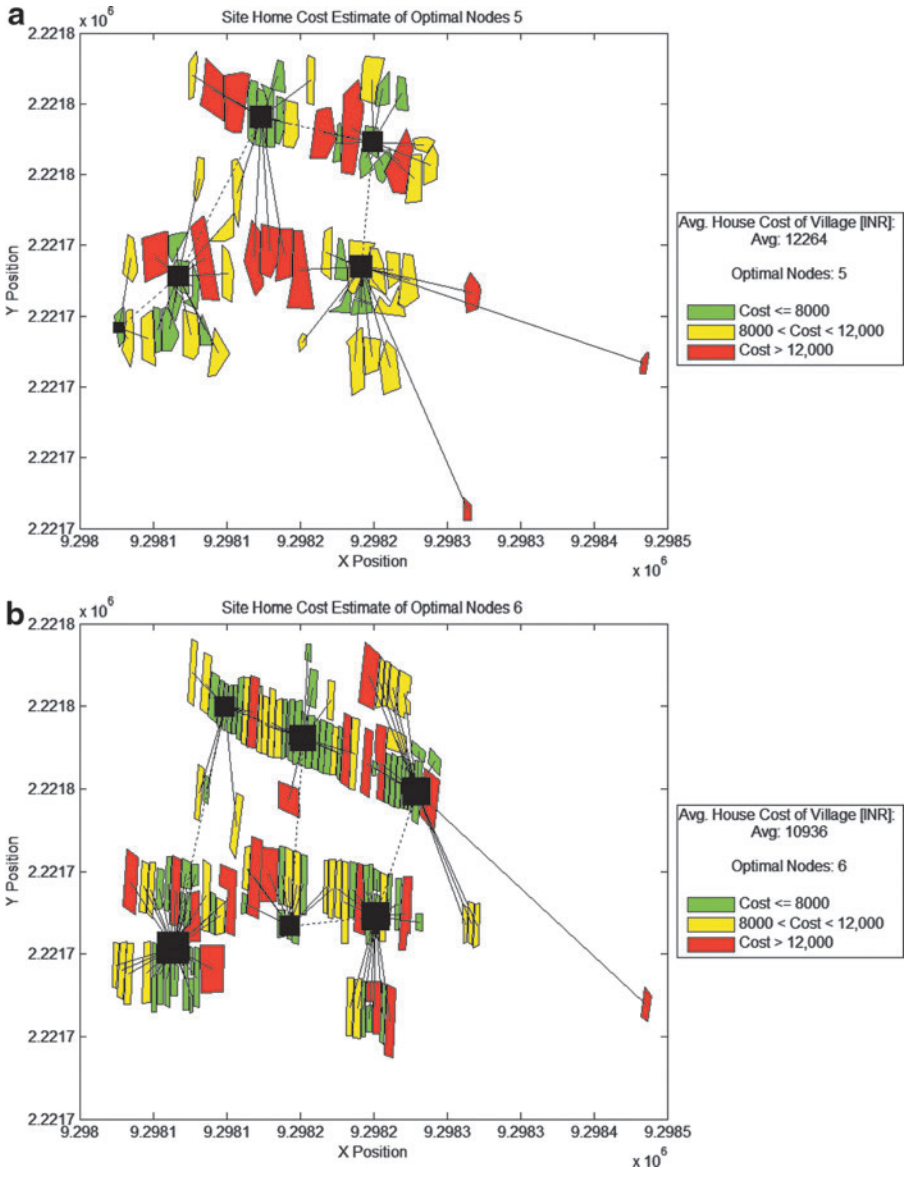

FIG. 9. The optimal configurations for the **(a)** building location estimate and **(b)** ground truth have a comparable number of hub generation nodes and average house cost.

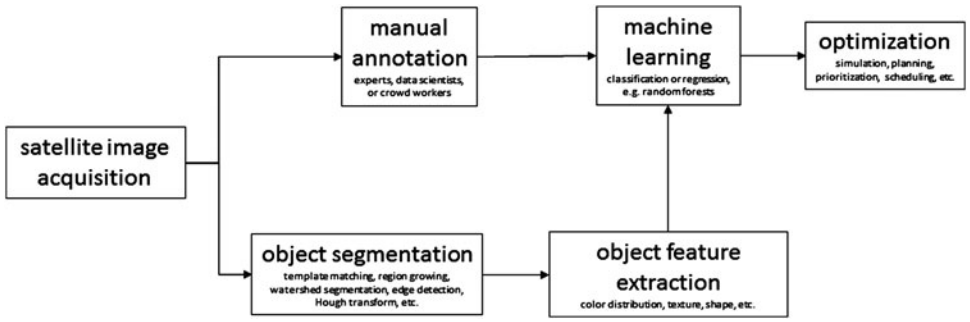

FIG. 10. A diagram illustrating the general workflow required for planning rural development activities based on satellite image analysis.

proposal writing and planning, such as power requirements, cost per house, total village cost, number and placement of generation nodes, and size and length of wire. Furthermore, by monitoring villages via satellite imagery postelectrification, we can begin to observe the impact of electrification on village growth. We can couple this spatial growth information with growth in per-capita demand to create a model of how villages grow postelectrification. From this growth model, we can simulate microgrid configurations based not on the village's current loads and topography, but on its projected future loads and topography. By intelligently designing microgrid systems to accommodate this future demand, we can select hardware to prevent premature bottlenecks in generation, storage, and current flow, which are common hindrances to the successful long-term implementation of microgrids.

Discussion and Conclusion

Social good projects in rural settings are hampered by lack of ground knowledge necessary for informed planning and decision making. This lack of knowledge stems from the lack of infrastructure to allow easy data gathering from remote rural locations, which is one of the reasons why social good projects are needed in the first place. Remote sensing provides an avenue to overcome this barrier to rural development programs when paired with satellite image analysis and other advanced modeling. With the emergence of low-cost satellite imaging and subsequent mass generation of data, we see remote sensing-supported planning playing an increasingly important role in rural development going forward. As we have discussed and illustrated through two real-world examples, remote sensing and satellite image analysis are poised to change rural development into a data-driven enterprise.

The two example problems and the several others mentioned in the introduction all involve an image processing component to derive features, a machine learning component to estimate suitability criteria, and an operations research component to determine optimal planning and resource allocation, that is, the full pipeline of descriptive analytics, predictive analytics, and prescriptive analytics. This is a common workflow that can be used in almost any rural development application that is constrained by cost or human resources. The workflow is summarized in Figure 10; we note that these stages should be continually evaluated and refined as informed by inputs from local experts.

When developing technical solutions or products for use in rural contexts, the onus of understanding the customer is significantly complicated by the fact that innovators, foreign or domestic, rarely live or interact with the final user in their normal spheres of activity. As a result, the ability to apply latent intuition during the design process is severely limited and the importance of engaging with local expertise is multiplied when creating a developing world product than one for consumers similar to the producer. For example, without consulting locals, it would not be possible to know that families in rural Kenya build separate thatched-roof sleeping structures for their unmarried, postpubescent sons. Recall that this issue of separate thatched-roof structures caused overestimation in the proportion of households with thatched roofs. Similarly, without iterative interaction with the microgrid companies in India, it would be quite difficult to determine the actual costs of available hardware and maximum capabilities of an ad-hoc labor force.

Another key learning has been that infrastructural growth dynamics are frequently quite different than those that innovators tend to be familiar with in their host environments. Images taken in different seasons can vary significantly and can have downstream consequences if unaccounted for. For instance, villages existing along the flood plains of the Ganges River frequently relocate depending on the degree of flooding or varying location of fertile soil. Assuming that homes and livelihood locations exist with the same degree of dynamism as in the western built environment would result in a picture that is artificially static in its mobility or economic activities. Lessons from western planning cannot easily be ported over to nonurbanized regions and, as a result, remotely sensed data like satellite imagery must be accompanied by deep knowledge of the social dynamic.

With all of these learnings in mind, we would still like to emphasize that existing image analysis techniques from the literature are usually sufficient to tackle selection problems in rural development when used properly, and thus should not be a bottleneck in such projects. The key is in determining and modeling the right criteria, whether it is something simple like type of roofing material or something complicated like the cost of the optimal grid topology. Though satellite imagery presents particularly a useful tool for tackling a new suite of developing world problems, the functionalities and dependencies are often outside of the innovator's set of expectations and must be investigated in parallel.

Acknowledgments

The authors thank Ramesh Sridharan for introducing them to each other, and Tilek Mamutov for discussions about Project Loon. Varshney and Abelson thank DataKind for teaming them, connecting them to GiveDirectly, and supporting the cash transfer village selection project. They also thank GiveDirectly for partnership, especially Joy Sun and Carolina Toth. Chen, Nowocin, Sakhrani, Xu, and Spatocco thank IDEAS Global Challenge for their financial support, Dr. Robert Stoner for his continued mentorship and support, the Tata Center for Innovation and Design for their financial support and guidance, and numerous conversations with SELCO in helping to ideate, design, and refine the tool toward optimal usefulness.

Author Disclosure Statement

No competing financial interests exist.

References

1. World Bank. World Development Indicators. Washington, DC, 2013.
2. International Fund for Agricultural Development. Rural Poverty Report. Rome, Italy, 2011.
3. Banerjee AV, Duflo E. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. PublicAffairs: New York, NY, 2011.
4. Haushofer J, Shapiro J. Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab, 2013.
5. World Bank. Decentralized Energy Services to Fight Poverty: Outcome Driven Engagement of Small and Medium Size Enterprises in the Provision of Energy Services in IDA Countries. Washington, DC, 2009.
6. Abelson B, Varshney KR, Sun J. Targeting direct cash transfers to the extremely poor. Proc ACM SIGKDD Conf Knowl Disc Data Min. 2014;1563–1572.
7. Verburg PH, Schot PP, Dijst MJ, Veldkamp A. Land use change modelling: current practice and research priorities. GeoJournal. 2004;61:309–324.
8. Doll CNH, Pachauri S. Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. Energy Pol. 2010;38:5661–5670.
9. Jaiswal RK, Mukherjee S, Krishnamurty J, Saxena R. Role of remote sensing and GIS techniques for generation of groundwater prospect zones towards rural development—an approach. Int J Remote Sens. 2003;24:993–1008.
10. Dudhani S, Sinha AK, Inamdar SS. Assessment of small hydropower potential using remote sensing data for sustainable development in India. Energy Pol. 2006;34:3195–3205.
11. Kumagai J. 5 Earth-imaging start-ups coming to a sky near you. IEEE Spectrum. 2014;51:20–21.
12. Hu Q, Wu W, Xia T, et al. Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping. Remote Sens. 2013;5:6026–6042.
13. Brunell R. Template Matching Techniques in Computer Vision: Theory and Practice. Chichester, UK: Wiley, 2009.
14. International Energy Agency. World Energy Outlook 2011 Electricity Access Database. Paris, France.
15. Census Data, 2011. Ministry of Home Affairs, Government of India, New Delhi.
16. Isola P, Zoran D, Krishnan D, Adelson EH. Crisp boundary detection using pointwise mutual information. European Conference on Computer Vision 2014.
17. Birth GS, McVey GR. Measuring color of growing turf with a reflectance spectrophotometer. Agronomy J. 1968;60:640–649.
18. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

Cite this article as: Varshney KR, Chen GH, Abelson B, Nowocin K, Sakhrani V, Xu L, Spatocco BL (2015) Targeting villages for rural development using satellite image analysis. *Big Data* 3:1, 41–53, DOI: 10.1089/big.2014.0061.

Abbreviations Used

IFAD = International Fund for Agricultural Development
NGO = nongovernmental organization
SHL = solar home lighting