Putting Big Data Innovation into Action for Development

Trevor Monroe, Stephanie Debere, Kwawu Mensa Gaba, David Newhouse, and Talip Killic

Abstract As part of the global data revolution, an increasing number of World Bank projects are based on insights from big data sources, including satellite-based measurements. Many use innovative machine- and deep-learning techniques to understand factors key to development. Leading satellite-based initiatives include:

- *Monitoring Electrification from Space:* Through analyzing 2 decades of satellite images for nightly light output from India's 600,000 villages, this project developed a novel data-intensive strategy to improve the monitoring of rural electricity provision. The data is accessible via an online visualization platform to help optimize electrification planning.
- *Mapping Poverty by Satellite:* To generate inexpensive, timely poverty estimates, this project examined how well satellite indicators contribute to poverty prediction, and how this depends on the type of prediction model. When compared with Sri Lankan census data, high-resolution satellite indicators track poverty very well, with potential to improve traditional poverty maps.
- *Satellite-Based Yield Measurement:* Through trials in Uganda, this project is testing a pioneering approach which relates satellite-based data to plot-level ground measures of yields. This enables future yield predictions, which can inform better policymaking to help farmers improve productivity.

Putting big data innovation into everyday practice requires collaboration between data scientists, technologists and sector specialists. The World Bank's experience has shown the value of learning by doing, collaboration and persistence.

As recently noted by the UN,[1] the data revolution is fueling waves of innovation and experimentation for data-driven development. The World Bank is seeing a rise in the number of projects using remote sensing data—as well as other sources such as ground sensors, social media and mobile phones—for insights and action in development.

---

[1] http://www.undatarevolution.org/report/

T. Monroe (✉) • S. Debere • K.M. Gaba • D. Newhouse • T. Killic
World Bank Group, Washington, DC, USA
e-mail: tmonroe@worldbankgroup.org

The "Innovations in Big Data Analytics" program provides technical assistance to World Bank teams to help operationalize big data innovation. The program has supported several projects that use satellite-based measurements for understanding factors key to development, such as poverty, urbanization, agriculture, road infrastructure and electrification. Particularly notable is the use of machine- and deep-learning techniques to extract insights from the growth of satellite data and tools. Here we profile three of our leading initiatives based on the use of big data from satellites.

Tracking Light from the Sky: Monitoring Rural Electrification from Space

Electricity is essential to human well-being worldwide, yet 1.2 billion people still live without it. Tracking the availability and supply of electricity at the local level is critical to improving service provision.

Data processing technologies are now enabling new ways to monitor access to electricity. Night lights data measured by satellite has been a useful resource for the development community for several years. However, the complexity of accessing, processing and manipulating this data has been a barrier to widespread use. While analysts have previously examined summaries or subsets of historical nighttime lights data, there had been no systematic effort to study the entire raw nightly data stream.

In 2011, a team from the University of Michigan, the US National Oceanic and Atmospheric Administration (NOAA) and the World Bank Group's Energy and Extractives Global Practice began to explore how to use night lights data in a scalable, systematic way. Their early work focused on validating the relationship between satellite-detected light output and the availability of electricity in several hundred villages across Senegal, Mali and Vietnam. The next step was to develop a strategy to exploit the detailed information from the full archive of nighttime satellite imagery to improve the monitoring of electricity supply around the world. The team refined its approach and scaled it up to look at all of India, which launched a major rural electrification program in 2005 to bring power to over 100,000 villages.

To evaluate the electrification program, the team acquired the complete historical archive of nighttime satellite imagery from the NOAA's Defense Meteorological Satellite Program. This has taken pictures of the Earth every night for over 20 years, creating an archive of multiple terabytes of high-resolution image data. Using geographic information systems (GIS) and data processing tools, the team analyzed the nightly light signatures of India's 600,000 villages (identified by geographical coordinates). The resulting dataset of almost five billion observations represents the most comprehensive database known describing electricity access and variability. Drawing on official electrification program records, the project

linked newly electrified villages to their nighttime light signatures, covering around 8000 nights during a 21-year period (1993–2013). This enabled verification of improvements to electrical supply, and identification of potential implementation problems.

The approach is a departure from prior research on nighttime lights, most of which uses annual composite images, which describe the average brightness of a locality over a calendar year. Yet in India and elsewhere, day-to-day variability in access to electricity is a far larger concern. By applying statistical and machine learning techniques, the team developed new methods to visualize patterns of supply disruptions. One objective is to use variability in light output data to identify electrical supply problems as they occur.

To make the data accessible to governments, power companies, regulatory agencies and other users, the team developed an online visualization platform, Nightlights.io. The site allows users to see, compare and contrast how light output has evolved over 2 decades, from state level to individual villages. Freely explored from any part of the world, it has the potential be a powerful tool in driving rapid electrification.

The project demonstrated that nighttime satellite imagery can reliably indicate the use of electricity in the developing world, even in rural contexts characterized by low power loads, few and dispersed users, limited infrastructure and erratic service provision. The team now wants to refine the online platform, build new capabilities
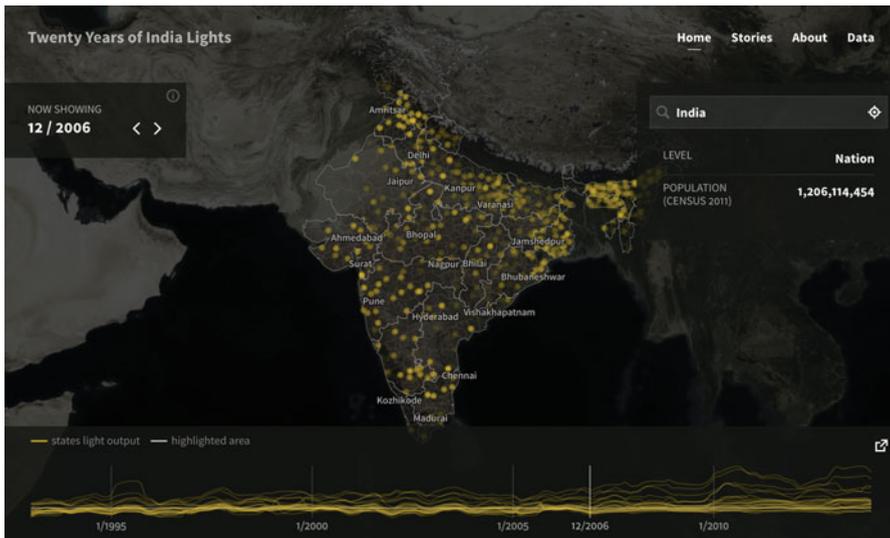
Fig. 1 Night-light map of India. The India Nightlights platform visualizes data from satellite images, so each point on the map represents the light output of a specific village at a specific time. Photo: World Bank

and generate nuanced reports to meet multiple stakeholder needs. It also wants to see how this approach could be replicated across the developing world (Fig. 1).

Tracking Poverty from Space

Poverty must be located accurately if development interventions are to be effectively targeted and monitored. However, long lags in processing mean that national estimates of poverty in developing countries are often several years old, while local-level estimates require census data that is expensive and collected infrequently. Big data, in the form of satellite imagery, has so far been largely untapped by policymakers wanting to understand where exactly the poorest people live. Little is known about which satellite-based indicators help predict poverty, and there is uncertainty around the best prediction model. Although numerous models have been developed, there has been little rigorous comparison of different approaches.

In response, this project explored the use of indicators derived from satellite data to predict geographic variations in poverty. Satellite data can generate a comprehensive picture of a particular area, and can be collected frequently at fine geographic levels, even in conflict areas inaccessible to surveys. Satellite-enhanced maps would be a key step towards the goal of real-time poverty estimates.

The project first examined how well publicly available low-resolution satellite indicators such as nighttime lights and land type contribute to poverty prediction, and how this depends on the method used to build the prediction model. To examine different models, the team applied out-of-sample validation techniques to household data from Pakistan and Sri Lanka. A randomly selected portion of the sample was repeatedly withheld when generating the prediction model, and accuracy was assessed by comparing extrapolated poverty rates from the prediction model to actual poverty rates in withheld areas. The team compared models derived using manual selection, stepwise regression and Lasso-based procedures. They also augmented the set of prediction variables with publicly available low-resolution satellite data to see whether this improved traditional poverty mapping techniques.

When satellite indicators were applied to the models in Pakistan—which generates district poverty estimates from a detailed household survey—they did not improve the accuracy of predictions. The rich survey information meant satellite-based indicators contributed nothing new. However, in Sri Lanka—more typical in generating poverty estimates from a census (meaning fewer indicators)—even freely available satellite indicators improved the accuracy of predictions. In both cases, the team found models selected using Lasso techniques work best for predicting poverty at local level, with sizeable benefits when there are many variables.

The team then purchased high-resolution (0.5 m per pixel) satellite imagery covering approximately 5% of Sri Lanka, including both rural and urban areas. They used multispectral imagery to capture variations in roof texture and surface material, enabling far more accurate identification of possible correlates of income. Novel methods are also emerging to detect smaller indicators, such as cars, which

evolve rapidly with economic growth. The team produced pan-sharpened mosaics (merging several smaller scenes) of the raw high-resolution imagery, and worked with experts to develop detection algorithms to identify possible poverty predictors. These include built-up area, building and car density, type of roofing, amounts of shadow, road type and agricultural land-use. Using open-source image processing algorithms, the team also constructed indicators such as the shape of buildings and amounts of paved road or built-up area. They assessed the density of each feature by local district and correlated the satellite-based measures with poverty estimates from the 2011 census data.

When these high-resolution satellite data indicators were combined with Sri Lankan census data, preliminary results showed that satellite indicators track regional differences in poverty extremely well. The satellite data proved a valuable complement to household survey data, with potential to generate accurate and updated local poverty maps to help refine targeting in development initiatives.

This approach is the first step of an exciting research agenda. Imagery can deliver new insights into a variety of development challenges, such as the scale of urbanization, infrastructure and the state of natural resources. More work is needed to explore which indicators best track local variations in poverty in a variety of contexts. These might include building density, roads, agricultural land or forest cover. More analysis is also needed to better understand the trade-off between
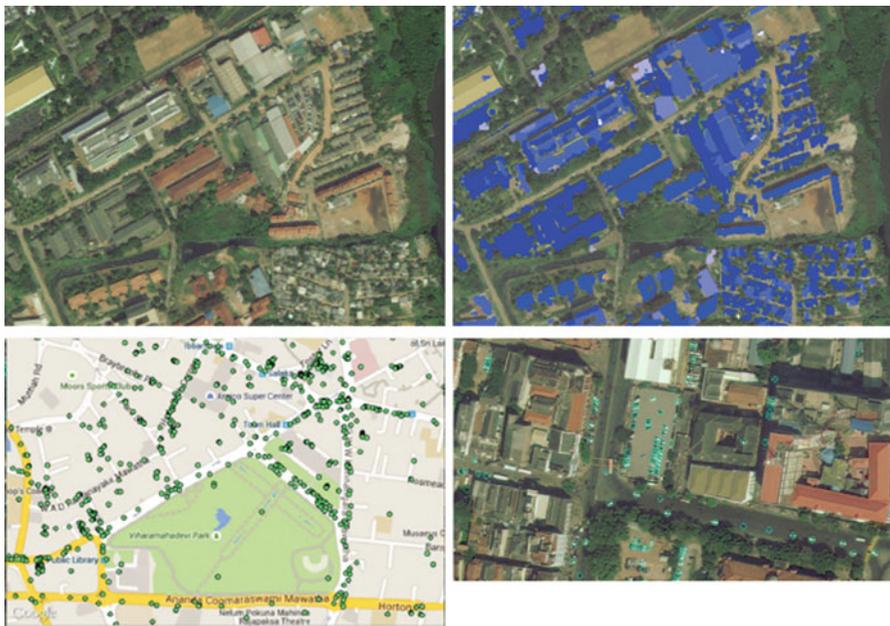

Fig. 2 Satellite image of rooftops. High-resolution satellite indicators such as roof type or density of cars track poverty, based on census estimates, extremely well. Photo: World Bank

the quality and cost of the imagery on one hand, and its benefits in terms of predicting local variations in poverty. Eventually, satellite-based imagery could also be a valuable tool for measuring inequality, monitoring development projects and "nowcasting" poverty rates (Fig. 2).

Satellite-Based Yield Measurement

Reliable data on crop productivity is essential for policy decisions that will improve agricultural yields and reduce poverty. Traditional approaches to measuring yields and productivity (such as household surveys) are resource-intensive and difficult to implement, particularly for smallholder systems. However, pioneering techniques using data from satellite imagery now offer more accurate, timely and affordable agricultural statistics.

Through trials in Uganda, this project is testing a pioneering approach to deriving reliable data on crop productivity from satellite imagery. The technique relates satellite-based data to plot-level ground measures of yields. This enables future yield predictions, which can inform better policymaking to help farmers improve productivity.
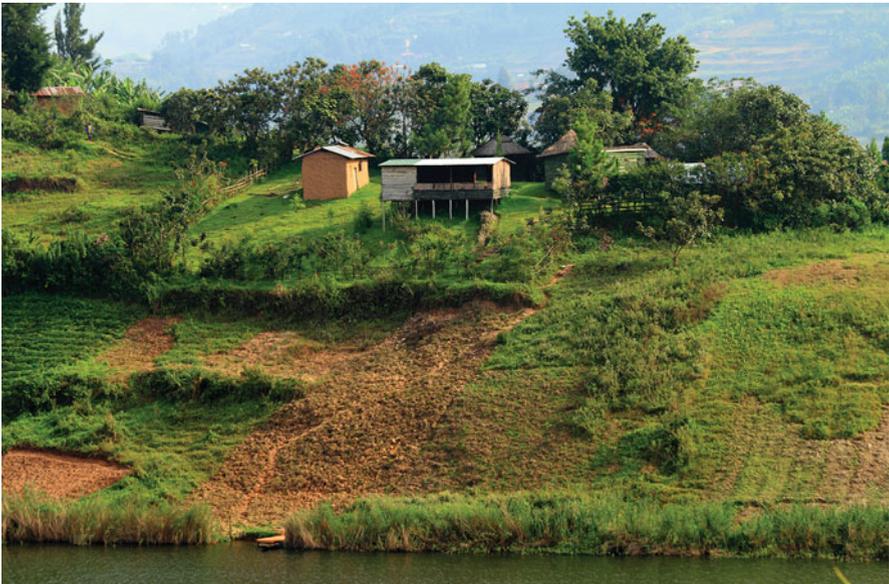
Fig. 3 Ugandan smallholding. Pioneering trials in Uganda are linking satellite-based data to plot-level ground measures of yields. This enables future yield predictions, to inform policymaking and improve productivity. Photo: bigstock.com

The team is testing Scalable Satellite-based Crop Yield Mapper, a statistical approach newly developed at Stanford University which relates satellite data to plot-level ground measures of yields in order to make future yield predictions. To validate the approach, this project tested satellite-based yield predictions against results on the ground for 900 maize plots in Uganda, a country highly dependent on smallholder agriculture. The outlines of each maize plot were captured via handheld GPS devices and used in conjunction with satellite imagery.

The project took objective and subjective measures of soil fertility (through conventional analysis of subsamples, and farmer reporting), maize variety (through DNA fingerprinting of leaf and grain samples, and farmer assessment assisted by photographic prompts). It is the first to test yield estimation in smallholder production via high-resolution satellite imagery against farmer self-reported harvest and objective ground research into actual yields. The results are still undergoing analysis, but they suggest that the approach could be scaled up across different crops and regions. Uganda is the first of several countries in Sub-Saharan Africa in which the team plans to validate this satellite-based remote sensing approach (Fig. 3).

Conclusion

Big data innovation is a stimulus for organizations such as the World Bank to adapt methods and practices to help our country clients thrive in an increasingly connected and data-filled world. Putting big data innovation into everyday practice in World Bank operations is a team sport that requires collaboration between data scientists, technologists and sector specialists. The World Bank's experience in driving big data innovation has shown the value of learning by doing, collaboration and persistence.