



BIG

DATA

innovation challenge

*Pioneering approaches
to data-driven development*



WORLD BANK GROUP



© 2016

International Bank for Reconstruction and Development / The World Bank

1818 H Street NW, Washington DC 20433

Telephone: 202-473-1000; Internet: www.worldbank.org

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries. Nothing herein shall constitute or be considered to be a limitation upon or waiver of the privileges and immunities of The World Bank, all of which are specifically reserved.

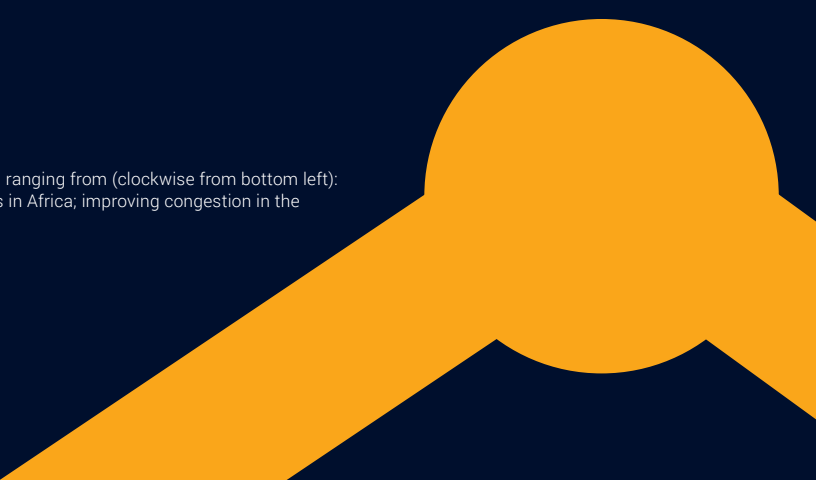
Rights and Permissions

The material in this work is subject to copyright. Because The World Bank encourages dissemination of its knowledge, this work may be reproduced, in whole or in part, for non-commercial purposes as long as full attribution to this work is given.

Any queries on rights and licenses, including subsidiary rights, should be addressed to the Office of the Publisher, The World Bank, 1818 H Street NW, Washington, DC 20433, USA; fax: 202-522-2422; e-mail: pubrights@worldbank.org

Front Cover:

Big data analytics offers potential for groundbreaking initiatives in development, ranging from (clockwise from bottom left): climate-smart agriculture in Latin America; boosting access to financial services in Africa; improving congestion in the Philippines, and linking crime to urban infrastructure in Colombia.



Foreward



Dear Readers,

We are truly delighted to share with you this compilation of an almost two-year journey, on which we embarked with 14 courageous World Bank innovators. Belonging to such a data-rich organization as ours, we realized in the summer of 2014 that a growing trend in the private sector – big data – could really enhance the way we achieve our results.

Big data capitalizes on the vast sources of data available to us today, from cellphone records to taxi GPS data, satellite images or new media. These reams of data, when anonymized and analyzed, can provide us with a wealth of information, such as the travel patterns of individuals in a crowded city or the path of a flu virus.

So we started to look for big data ideas within the Bank and found a few colleagues who were trying out new initiatives. We learned that many faced similar challenges, such as lack of access to certain types of data, a desire for world-class data science expertise, lack of storage and computational capacity, a desire for good practices on handling privacy, opportunities for peer-to-peer learning, and platforms and norms for sharing data and software. There was also need for seed and growth funding to kick-start the high-potential, but often high-risk, big data initiatives.

We realized the importance of creating a big data program with strong projects that could inspire others. Following up on this instinct, we launched the Big Data Innovation Challenge – not knowing what to expect, as this was a concept so new to our staff. We weren't even sure if we would get more than 10 proposals, and were pleasantly surprised to receive 131 innovative ideas, from which we chose 14 winners.

And what a diverse group of winners they were, from tracking rural electrification in India using satellite imagery, to analyzing taxi hailing services in the Philippines to reduce congestion, to using cellphones to measure road conditions in Belarus, and analyzing social media content in Brazil to understand citizens' political sentiment.

In November 2014, the winners all received seed funding to test their pilot ideas, as well as technical help from data scientists from the newly established Big Data team in the Bank's Innovation Labs.

Our journey with these teams has continued during the course of several months, during which we have seen ideas tried, failed, refined and field-tested. Each story became a personal quest for us and we are proud that many of these are now ready for scale-up and adoption in operations.

This publication profiles stories from the Challenge winning teams and finalists, showing how data on an unprecedented scale has the potential to be transformational in its effects.

We hope you will find these stories as inspiring as we have.

Adarsh Desai

Program Manager, Innovation Labs
The World Bank Group

Contents

| | |
|-------------------|-----|
| Introduction..... | iii |
|-------------------|-----|

At-a-Glance:

| | |
|---|----|
| Innovation Challenge winners and top finalists..... | vi |
|---|----|

Winners and Top Finalists

| | |
|---|----------|
| Mining Big Data for Climate-Smart Agriculture..... | 1 |
|---|----------|

*Erick C.M. Fernandes, Daniel Jiménez, Andy Jarvis
and Sylvain J. Delerce*

| | |
|--|----------|
| Big Data for Financial Inclusion..... | 7 |
|--|----------|

Sven Harten

| | |
|--|-----------|
| Improving Road Investments through Mobile Data..... | 13 |
|--|-----------|

Kai Kaiser

| | |
|---|-----------|
| Securing Property Rights through Geo-spatial Data..... | 19 |
|---|-----------|

Kathrine Kelm

| | |
|---|-----------|
| Open Traffic: Easing Urban Congestion..... | 25 |
|---|-----------|

Holly Krambeck

| | |
|--|-----------|
| Observing People's Feelings about State Institutions... | 31 |
|--|-----------|

Victoria L. Lemieux

| | |
|---|-----------|
| Monitoring Rural Electrification from Space..... | 37 |
|---|-----------|

Kwawu Mensan Gaba

| | |
|--|-----------|
| Mapping Poverty by Satellite..... | 43 |
|--|-----------|

David Newhouse

| | |
|--|-----------|
| Understanding How Infrastructure Affects Crime..... | 49 |
|--|-----------|

Camila Rodríguez, Andrés Villaveces

| | |
|--|-----------|
| Revamping Road Condition Monitoring with Smartphones..... | 55 |
|--|-----------|

Wei Winnie Wang

Projects to Watch

| | |
|---|-----------|
| Targeting Poverty by Predicting Poverty... | 61 |
|---|-----------|

Melissa Adelman

| | |
|--|-----------|
| Assessing Whether Markets are Working for the Poor..... | 62 |
|--|-----------|

Alvaro S. Gonzalez

| | |
|---|-----------|
| From Cellphone Data to Poverty Maps..... | 63 |
|---|-----------|

Marco Hernandez Ore

| | |
|---|-----------|
| Testing Cellphone-Derived Measures of Income and Inequality..... | 64 |
|---|-----------|

Tariq Khokhar

| | |
|---|-----------|
| Satellite-Based Yield Measurement..... | 65 |
|---|-----------|

Talip Kilic

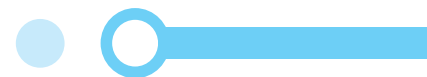
| | |
|--|-----------|
| Understanding Individual Travel Patterns in African Cities..... | 66 |
|--|-----------|

Nancy Lozano Gracia, Talip Kilic

| | |
|-------------------------|-----------|
| Key Lessons..... | 69 |
|-------------------------|-----------|

| | |
|----------------------|-----------|
| Glossary..... | 72 |
|----------------------|-----------|

| | |
|------------------------------|-----------|
| Acknowledgements..... | 74 |
|------------------------------|-----------|



Big data can sound remote and lacking a human dimension, with few obvious links to development and impacting the lives of the poor. Concepts such as anti-poverty targeting, market access or rural electrification seem far more relevant – and easier to grasp. And yet some of today's most groundbreaking initiatives in these areas rely on big data.

This publication profiles these and more, showing how data on an unprecedented scale has the potential to improve lives in unprecedented ways. The featured case stories illustrate the diverse range of big data applications in development. For the World Bank, with twin goals of ending extreme poverty and boosting shared prosperity, big data is big news – and this is just the beginning.

What is big data and why does it matter?

Big data is an umbrella term used to describe the constantly increasing flows of data emitted from connected individuals and things, as well as a new generation of approaches being used to deliver insight and value from these data flows. It is said that more data has been generated in the past two years alone than in all previous years combined. While most of the attention given to big data has focused on the developed world, the rapid diffusion of technologies such as the internet, cellphones, ground sensors and satellites – to name a few – is driving big data innovation in the developing world. And while data flows in the developing world are typically smaller and less diverse than in the developed world, they still present incredible opportunities for data scientists, economists and statisticians to use big data to enhance or supplement traditional analytical approaches.

Unlike traditional sources of development data, such as household surveys, which address specific research questions, big data is usually produced in the course of some other activity (such as making a cellphone call). This, along with the size and complexity of some datasets, requires different research methods. Big data analytics is the emerging set of tools and methods to manage and analyze this explosive growth of digital information. It includes data science methods like machine learning, predictive analytics, and visualization. These methods open significant potential for drawing on real-time information to address development challenges – potential that can't be ignored. To this end, the World Bank's Innovation Labs, housed in the Leadership, Learning and Innovation Vice Presidency, launched its Innovations in Big Data Analytics program in November 2014.



Analysis of climate and yield data can tell farmers what and when to plant for anticipated weather conditions

The World Bank's Big Data program

The Big Data program brings together data scientists, social scientists and sector specialists in a work program with two main objectives:

- To **accelerate organizational capabilities** in big data analytics for use in research and operations – to help the World Bank Group (WBG) better work towards ending extreme poverty and boosting shared prosperity.
- To **position WBG as a leader** in the use of big data solutions in development.

The program aims to scale early pilots into projects that solve significant development challenges, and to establish best practices for using big data analytics to steer evidence-driven development. To mainstream and embed big data analytics across the organization, the program:

- Provides data science **technical assistance** to projects with high potential to demonstrate early value through big data
- **connects** big data practitioners through workshops, training, knowledge events and online communities to foster collaboration, knowledge flows and learning-by-doing between data scientists, sector specialists and external entities
- works with internal technology and service providers to develop big data **technologies and tools** that meet evolving needs

- develops **training**, learning events and knowledge products exploring big data use in sectors from agriculture and health to energy and disaster risk
- builds **partnerships** to develop collective capacities and strengthen the WBG's role as a leader in big data
- **incentivizes** the use of big data through innovation challenges to solve development problems.

Taking up the Challenge

Launched in September 2014, the Big Data Innovation Challenge has been key in encouraging big data approaches. Exceeding all expectations, it attracted 131 innovative proposals and awarded 14 with funding and expertise to enable big data analytics in their projects. The winning initiatives cover an exciting range, from using satellite imagery to improve poverty mapping, to mining social media data to understand political sentiment, or cellphone data to increase the use of banking services. Others promote traffic flows or accountable road building, anticipate crop yields, predict violent crime and promote registration of land rights.

This publication profiles 16 extraordinary initiatives from the Challenge winning teams and finalists. The case stories examine the application of big data analytics and how it can help achieve project goals. By demonstrating the impact, use and value of big data in development, they show that you don't need to be a data scientist

to understand how these approaches can improve people's lives. Work now continues beyond the Challenge to take several of the projects to the next stage of growth. The Big Data program is also currently engaged with several WBG Global Practices to accelerate their progress in big data. These engagements involve a range of activities to deliver technical assistance, knowledge and learning, and essential resources to operationalize big data capabilities in each practice.

A rich learning process

Innovative paths inevitably involve hurdles. Each case story revealed useful lessons, and several testify to the value of perseverance. However, the potential rewards of big data make the effort worthwhile. Key themes include the need to capture, prepare and store data meticulously – and to plan enough time to do so. Successful big data solutions involve approaching existing situations from new angles or combining previously unrelated data sources – but these novel data approaches must be tested, validated and adapted for mainstream use.

Despite the central role of computational power, the human element remains vital to the success of these projects. Big data analysis can be enhanced by traditional research techniques, such as socioeconomic surveys. Several stories stress the need to invest in partnerships, or to combine human and computational power for optimum results. Big data analytics is a team sport: Effective collaboration between data experts, technologists and business sector specialists is also crucial.

Putting big data into action for development

The Big Data program will share these lessons widely, supporting early adopters and helping transition big data to mainstream use by providing advice, infrastructure, expertise and resources. We hope the exciting approaches in this publication will demystify big data in development and motivate its applicability to development challenges. Above all, we hope it will inspire you to put big data into action in your own work. These case stories demonstrate that big data can improve development effectiveness and help World Bank operations achieve results through solutions with better evidence, efficiency, awareness, understanding and forecasting. At the same time, we hope others outside the World Bank will find inspiration from these experiences in applying big data approaches for development. Ultimately, big data analytics can be an accelerator for ending poverty and boosting shared prosperity.



In Bogotá, big data analysis has identified urban features and times of day associated with violent crime

*For more information on big data and updates on any of these case stories, email innovation@worldbank.org
To download the World Bank's 2014 report 'Big Data in Action for Development,' visit <http://bit.ly/bigdatainaction>*

Big Data Innovation Challenge

At-a-Glance: Winners and Top Finalists

Mining Big Data for Climate-Smart Agriculture

*Erick C.M. Fernandes, Daniel Jiménez,
Andy Jarvis and Sylvain J. Delerce*

Climate change is making seasonal planting decisions unreliable, so researchers are comparing long-term data on harvests and climate, to identify climate sequences favorable or unfavorable for cropping in specific areas. By matching these with forecasts, farmers can tell what and when to plant for the anticipated conditions.

Big Data for Financial Inclusion

Sven Harten

By combining big data analytics and socio-economic research in Africa, the team established a statistical customer profile for users of digital financial services. They then identified matching profiles among phone customers not using such services, and mapped their locations. These insights are helping financial companies reach previously unbanked people.

Improving Road Investments through Mobile Data

Kai Kaiser

To track nationally-financed projects to improve local road networks in the Philippines, this project developed OpenRoads, an interactive multi-media portal and set of digital tools. The platform links official road data with crowd-sourced geo-tagged video and image data from mobile devices, to increase accountability in road network investment projects.

Securing Property Rights through Geo-spatial Data

Kathrine Kelm

Using unmanned aerial vehicles (drones), this project recorded imagery which is processed into maps and 3D computer-generated landscape models. Communities can use these to identify property boundaries and work with authorities to register their land. The process helps secure land rights in developing countries in a fast, cost-effective way.



Open Traffic: Easing Urban Congestion

Holly Krambeck

To help traffic management agencies in developing countries monitor real-time traffic conditions and mitigate congestion, this project developed an open-source platform, Open Traffic, for collecting, visualizing and analyzing traffic speed data derived from taxi drivers' smartphones.

Observing People's Feelings about State Institutions

Victoria L. Lemieux

By conducting sentiment analysis of tweets made during civil unrest around the 2014 Soccer World Cup in Brazil, researchers found that people protest when feeling deprived relative to some external standard – in this case, spending on the tournament. Their analytical method increases understanding of the relationship between civil unrest and citizens' sentiment.

Monitoring Rural Electrification from Space

Kwawu Mensan Gaba

Through analyzing two 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

David Newhouse

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, and have potential to improve traditional poverty maps.



Understanding How Infrastructure Affects Crime

Camila Rodríguez, Andrés Villaveces

Latin America has above-average crime rates and unplanned cities with high inequality. To define the association between crime and infrastructure, this project drew on rich existing data about Bogotá in Colombia. Through Risk Terrain Modeling, the team identified specific urban features and times of day associated with assault and homicide.

Revamping Road Condition Monitoring with Smartphones

Wei Winnie Wang

This project developed an app called RoadLab, which uses crowdsourced data from accelerometers in smartphones carried in moving vehicles to evaluate the roughness of road surfaces. RoadLab gives national road agencies comprehensive and frequent information on road surface condition, to help them manage assets cost-effectively.

Projects to Watch – results still pending

Targeting Poverty by Predicting Poverty

Melissa Adelman

Targeting errors are common in development programs. By applying machine learning techniques to data-sets used for targeting poverty, this project seeks to improve methodologies for identifying the poor.

From Cellphone Data to Poverty Maps

Marco Hernandez Ore

Outdated poverty maps in many developing countries limit governments' ability to make effective anti-poverty decisions. Using anonymized cellphone call records in Guatemala, this project aims to create a tool to produce inexpensive, near-real-time poverty maps.



Assessing Whether Markets are Working for the Poor

Alvaro S. Gonzalez

Many markets in which the poor transact are volatile and fragmented. This project analyses existing micro-level price data to provide near-real-time information on how well markets are working, to help policymakers improve them for the poor.

Testing Cellphone-Derived Measures of Income and Inequality

Tariq Khokhar

Many countries' official measures of poverty are outdated and inconsistent. This project evaluates techniques that use cellphone call records to offer timely and complete poverty estimates, and examines how these techniques can enhance government workflows.

Satellite-Based Yield Measurement

Talip Kilic

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.

Understanding Individual Travel Patterns in African Cities

Nancy Lozano Gracia, Talip Kilic

This project combined personal interviews and analysis of big data from smartphones' GPS sensors. It aims to capture accurately and affordably the route, purpose, mode and cost of individuals' travels, to help improve urban transport and land-use planning.



Mining Big Data for Climate-Smart Agriculture

*Erick C.M. Fernandes (pictured),
Daniel Jiménez, Andy Jarvis
and Sylvain J. Delerce*



Data analysis can reveal the combination of climate factors resulting in high or low crop yields in a specific region



Photo: Neil Palmer

Data-driven agronomy will help increase the sector's capacity to adapt to climate change

SUMMARY

Climate change means traditional calendar-based decisions about what and when to plant are no longer reliable for farmers. New approaches are urgently required to provide relevant and timely information for decision making. In Colombia, the International Center for Tropical Research (CIAT) has combined long-term data on rice harvests and climate patterns, and analyzed it to provide essential information for crop planting decisions. This project draws on the Colombian experience. The team partnered with rice-growing experts from Argentina, Brazil, Chile and Uruguay, helped them prepare cropping and climate data, and held a workshop to analyze their data. The technique combines the two databases on harvests and weather, and relates each harvest to the corresponding climate sequence for approximately 120 days between sowing and reaping. The data is analyzed to unravel underlying correlations between climate factors and yield variability. This enables identification of climate sequences that are favorable or unfavorable for cropping in specific areas. By matching these with seasonal forecasts, the system can advise farmers on what variety of rice to plant for the anticipated conditions, and when to plant it. To support field data capture by farmers, the project also created a web platform and an Android app. The tool can be applied to any crop or location, providing farmers with data-driven information to help them decide what, when and where to plant.

CHALLENGE

Greater weather variability due to climate change means calendar references are increasingly outdated in helping farmers make the right planting decisions. New approaches are urgently required to provide them with relevant information and enhance their resilience to climate variability. Recent climate change analysis suggests Latin America and the Caribbean could have a future climate more suitable for growing rice – the world's most important food crop in terms of energy consumed by humanity. In Colombia rainfall and temperature extremes have changed differently in each region, and the climate is increasingly unpredictable, resulting in national average rice yields falling from six to five tonnes per hectare in less than five years (without noticeable changes in soil or crop management). Faced with an increasingly volatile climate, farmers need new methods for deciding when and what to plant.

This would foster data-driven agronomy and increase the sector's capacity to adapt to climate change



INNOVATION

CIAT's technique combines and analyzes two country-wide databases, covering commercial harvests and weather. Data on harvests has been collected by the Colombian National Rice-Growers' Federation for almost 20 years, covering variables such as yield, grain humidity, sowing and harvest dates, cultivar, municipality and cropping system (irrigated or upland rice). Data from the Colombian National Meteorology Institute provided daily records of five variables: Maximum and minimum temperature, precipitation, relative humidity and solar radiation.

By combining these two databases, each individual harvest event can be related to a corresponding climate sequence for around 120 days between sowing and harvest. Data is analyzed using machine learning techniques such as Artificial Neural Networks (ANN), random forest and clustering, to reveal the combination of climatic factors that result in high or low yields in a specific region.

Based on the harvest events database for two municipalities, analysis of the 120 days from sowing to harvest for the five climate variables revealed 17 different climatic sequences. By matching these against the yields obtained under each sequence and evaluating the relevance of specific climate indicators against the growing phases of the crop, CIAT identified favorable and non-favorable climatic situations. Projecting these onto seasonal forecasts, the system can advise farmers on optimum sowing dates and rice varieties, given the weather ahead. By providing farmers with data-driven information for planting decisions, the tool can

compensate for the increasing irrelevance of traditional knowledge due to climate variability.

Expanding the project reach

In pilot tests, analysis of crop data together with weather data generated advice against planting rice that season, due to projected adverse climate conditions. Farmers received the planting advice via the growers' association. Those who ignored it harvested nothing and lost considerable inputs, while those who followed it saved seed, labor, fertilizer and water. The historical data allowed insight at a site-specific level. For example, in Saldaña – where rice is grown year-round, due to irrigation policies – the analysis showed that rice yields were limited mainly by solar radiation during the grain-ripening stage. This implies that farmers can boost yields by aligning sowing dates with sunnier seasons, or choosing crop varieties resilient to low solar radiation.

With this success suggesting other countries and crops could benefit from the same approach, the project worked to scale it up in Latin America. The team visited Argentina, Uruguay and Brazil to seek partnerships with agricultural associations and assess data availability. CIAT then helped partners with data collection and preparation for five months before holding a workshop in Uruguay to teach them to prepare and analyze their own data. The training involved participants using the system for themselves, to ensure they could replicate the process in their own contexts.

Building a real-time system

Equally important was the move from using only historic data on weather and yields, to

collecting real-time data and further automating the analysis to build a real-time system. To enable field data capture by farmers for factors such as soil type and crop management techniques, the team developed an Android app, using open-source software and an external provider to build the app. Supporting a wide range of Android versions, it can capture GPS coordinates using one button. Stored in a cloud, the data is instantly available and can be exported to any format for reuse. The app is linked to a web platform which generates personalized reports, real-time information, interactive graphs and mapping, allowing for optimum site-specific crop management. The system will be further refined by increased use of machines such as drones or inexpensive wireless sensors for data capture. This allows high frequency measuring, without requiring effort from farmers and eliminating the margin of human error.

RESULTS

The workshop teams spent most of the time preparing their data: standardizing and cleaning it, and relating crop data with weather series. This reflects the reality of work with data. Each team completed at least one analysis using random forest models that allowed them to identify the most relevant factors behind yield variability, and evaluate the relationships between input and output variables.

The group quickly grasped the approach. Some teams achieved satisfying results which coincided with previous work, while others obtained unexpected results due to incomplete and insufficiently processed datasets.

All experienced the problems that occur while handling real data (such as outliers or variables out of the range when cleaning data). This generated a rich debate around how best to capture data. The limited capacity of participants' personal computers restricted the exercise to small datasets using only ANN and random forest techniques, prompting CIAT to run the analysis on its more powerful server facilities, with whole datasets. Exploratory results using information from one rice mill showed that previous land use also influences rice production.

The app is being finalized after initial user feedback reported that it was too slow loading and saving information, and did not work on all platforms (IOS/Apple, Windows, etc.). This caused some frustration among users (who wanted quick responses to problems, as if from a service provider rather than a research project).

The system could potentially be adapted to almost any crop, and the team aims to pilot it in Africa in 2016. It is hoped the workshop will form the basis for the establishment of a community of practitioners in big data in agriculture. This would foster a culture of data-driven agronomy and increase the sector's capacity to adapt to climate change.

LESSONS LEARNED

The scaling-up and move towards real-time data capture revealed that data quality is of paramount importance to the reliability of the system, but human relationships through local partnerships and a user community will also be crucial to its success.

- ***Work specifically to create a positive user experience***

User experience is central to success. Technical issues undermine user trust, making the adoption process harder. A tool must offer sufficient services to engage users, and be easy to operate, otherwise it will not be used.

- ***Invest in regional and local partnerships***

Partnerships on the ground are key to scaling up – particularly with agricultural organizations, which can source data from numerous locations and deliver recommendations to many farmers efficiently. It's important to gain credibility with partners, so they will share information: Open data-sharing is still in its infancy in many places and primary data holders often have legitimate concerns about how information they share will be used.

- ***Capture, prepare and store data meticulously***

High-quality input data is essential for accurate results. Agricultural data is readily available, but there is big potential to modernize the methods used to capture, store and share it – for example, through machine-based data sourcing. The data preparation step (addressing missing data,

outliers, correlated variables, etc.) must be completed, and organizations should evolve towards cloud-based technologies, so data is centralized and always available.

- **Create a community of practitioners**

Networking will be powerful in promoting uptake and refinement of these methods. CIAT hopes to support a user community for data mining techniques in agriculture, similar to the 'R community' – a global network of more than 2 million users

and developers who voluntarily contribute technical expertise to maintain and extend the 'R software' language. This approach demonstrates how a tool can be continuously improved.

www.open-aeps.org:8080/



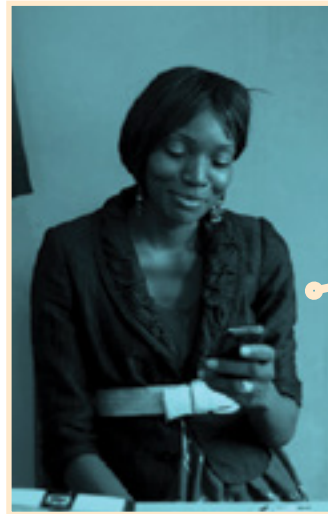
Teams at the project workshop learned data analysis techniques to identify the factors behind yield variability

Big Data for Financial Inclusion

Sven Harten



Analysis of mobile phone data can help increase subscribers' use of banking services, boosting their economic resilience



SUMMARY

Access to financial services is essential to development efforts. Across Africa, financial inclusion remains below potential – partly due to the challenge for financial institutions in developing products for the low-income mass market. This project uses available data in an innovative way to help providers offer affordable financial services to previously unbanked people. By combining big data analytics and socio-economic research in Ghana, Uganda and Zambia, the team created a powerful tool to increase adoption and use of digital financial services (DFS). This enabled them to establish a statistical customer profile for an active DFS user. They then searched the big-data set to identify matching profiles among phone-service customers who are not yet DFS users, and mapped their locations. These are the customers most likely to become active users of DFS. This intelligence can be used for product development and targeted marketing campaigns to increase the supply of financial services to the previously unbanked. It also offers valuable insights for the World Bank's drive for full global financial inclusion by 2020. In Ghana, the use of project findings has already led to the financial inclusion of more than 70,000 people.

CHALLENGE


Across Africa, the coverage of formal financial services is poor, with low-income and rural customers largely excluded. Yet their need for affordable financial services is well documented. Financial inclusion increases resilience to economic shocks and helps grow small and medium-sized enterprises. The International Finance Corporation (IFC) is committed to helping create 600 million bank accounts in the developing world.

Through the Partnership for Financial Inclusion – a joint initiative with the MasterCard Foundation – the IFC is working to expand microfinance and advance digital financial services (DFS) in Sub-Saharan Africa. The growth of DFS offers huge potential to increase financial inclusion in developing countries, especially given the proliferation of mobile phone use. With high mobile phone penetration across the continent, many microfinance institutions, banks and mobile network operators (MNOs) are developing DFS. However, although many customers have registered for these services, only a minority use them regularly. Financial institutions also lack information about potential customers, which products address their needs and how to provide access to those products.

Big data offers the opportunity to mine existing information about mobile phone and DFS users to help MNOs and DFS providers to deliver products and services to previously excluded customers. Using a company's DFS transactions database and call detail records, the project team sought to characterize active mobile DFS users and understand what drives inactivity.

They hoped to identify behavior patterns among customers and use that information to stimulate better use of DFS and identify potential new users.

The study has already led to the financial inclusion of over 70,000 people



INNOVATION

The project used big data from MNOs and financial institutions in Ghana, Uganda and Zambia to calculate profiles for users of mobile financial services. Using statistical predictors derived from the usage patterns of the current active network user base, the team firstly identified which MNO customers are highly likely to become active DFS users. As mobile phone data does not contain socio-economic or demographic information, they also designed classic surveys to achieve more complete profiling of users and non-users of financial services. Potential users can then be targeted by marketing campaigns.

Big data analysis

The study started with a big data analysis of call detail records (CDRs) covering one MNO per market, each with an average of 4 million mobile subscribers. Six months' of CDRs and DFS

transaction records, nearly two terabytes in size, were extracted from the MNOs' servers. The team segmented users into three categories:

- Voice only
- Registered but inactive DFS users
- Active DFS users.

These segments showed very distinct patterns of voice calls, social network structures and geographical mobility. Active DFS users make on average almost twice as many phone calls as non-users. These calls also last significantly longer. They also send and receive the most text messages and have a much larger social network. They are therefore the high-value customers and early adopters each MNO seeks to attract.

The research found that many telecoms-only customers had a demographic profile similar to these highly active DFS users, indicating a strong correlation between high users of telecoms services and the potential to be an active DFS user. The team therefore scored all telecoms subscribers according to the extent to which users are similar to the profile of highly

active DFS users. Using machine learning techniques, they modeled the 15 most powerful variables (such as the number and length of calls or number of call contacts) which predict whether a subscriber is likely to become a DFS user. Based on the findings, the team compiled maps showing the actual distribution of DFS users, the distribution of predicted adopters, and districts with highest concentrations of likely adopters.

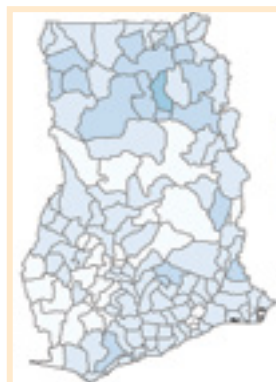
Classic socio-economic profiling interviews

To strengthen the profiles of different types of DFS customer, the team also carried out a socio-economic study in Ghana:

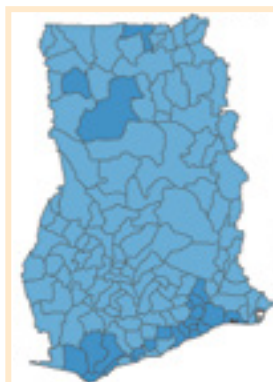
- ***Demographic profile***

Based on the average three-monthly volume of voice, text and data usage, network subscribers were organized into high, mid and low users. A random selection of 500 from each segment was interviewed by phone. The results indicated that mobile phone users are 61 percent male and relatively young (45 percent under 35), with good literacy and access to financial services (66 percent have a bank account).

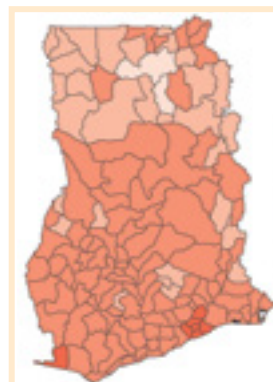
District-level adoption rate



Predicted adoption rate



Top target districts



Through big data analysis, the team mapped the distribution of current, predicted and likely users of digital financial services in Ghana

There is an apparent gender effect at all three levels of mobile activity: Women switch less between providers and have lower mobile activity levels on the network they use. This suggests opportunity for MNOs that develop marketing strategies targeting women.

- ***Use of Digital Financial Services***

Ninety-five percent of MNO subscribers, including users over 55, are aware of DFS. However, there is a gap between awareness and usage, especially among low-activity subscribers. There is also great disparity between male and female subscribers, with low-active male network users having the same level of DFS usage as high-active female network users. The gap between awareness and usage decreases the younger the users are. The key reason why so many MNO subscribers are not using DFS is around understanding and products. Twenty-eight percent of non-users declared they have no need for DFS, which suggests they need explanations of how DFS can help their financial management. MNOs should also consider whether they have the right products for these customers. Twenty-three percent of customers reported having no money to use with DFS, reinforcing the need for customer education, as even with irregular incomes, many could still benefit from DFS.

RESULTS

In this study, big data was used to discover the profile of those MNO voice customers most likely to become regular DFS users. Research then identified socio-economic groups that fit these profiles but were not using DFS. It is reasonable to expect that a combination of targeted marketing and the provision of DFS relevant to these profiles should result in significantly increased active usage of DFS.

Comparing the findings from the socio-economic survey with big data analysis, the team found that infrequent users of voice calls are also more likely never to have used DFS. While younger people are the most active users of voice, they also have the largest share of registered yet inactive DFS accounts. This suggests improvements needed in DFS products, as mobile-savvy younger customers are ignoring them despite using their phones regularly. An important finding was that DFS users display increased network usage and loyalty to an MNO than non-DFS users. The socio-economic research showed high potential for growth, given that nearly half of voice subscribers have never used DFS. In particular, the youth segment and infrequent female voice users are high potential target groups who could be approached with tailored products and communication strategies. Appealing to these consumers could lead to increased use of DFS and ultimately to greater financial inclusion. Findings from big data analysis are already helping MNOs promote both the use of DFS and their telecoms business. In Ghana, use of information from this study has so far led to the financial inclusion of more than 70,000 additional people. The MNO

in Ghana called the list of potential customers directly to promote DFS. These calls were far more effective than previous indiscriminate efforts to attract customers.

The team now hopes to develop more sophisticated metrics of phone use, such as social network structure, behavioral traits, geographic segmentation and mobility analysis. Statistical analysis could also identify key traits that differentiate groups, such as voice versus DSF users or active versus inactive DSF users. Ultimately the researchers aim to construct a model to compute the likelihood that each voice subscriber will sign up for DFS or become an active user.

LESSONS LEARNED

The combination of big data analysis and classic research methods offers valuable lessons in getting the most from big data:

- ***Enhance big data analysis with traditional research techniques***

Despite the current hype, big data is ultimately 'only' a new (albeit very rich)

data source, and it does not make other data obsolete. Talking to people remains a powerful source of information. For example, big data analysis offers huge potential to support financial inclusion, but only by enhancing it with consumer profile research it is possible to target customers with precision.

- ***Plan for imperfect data***

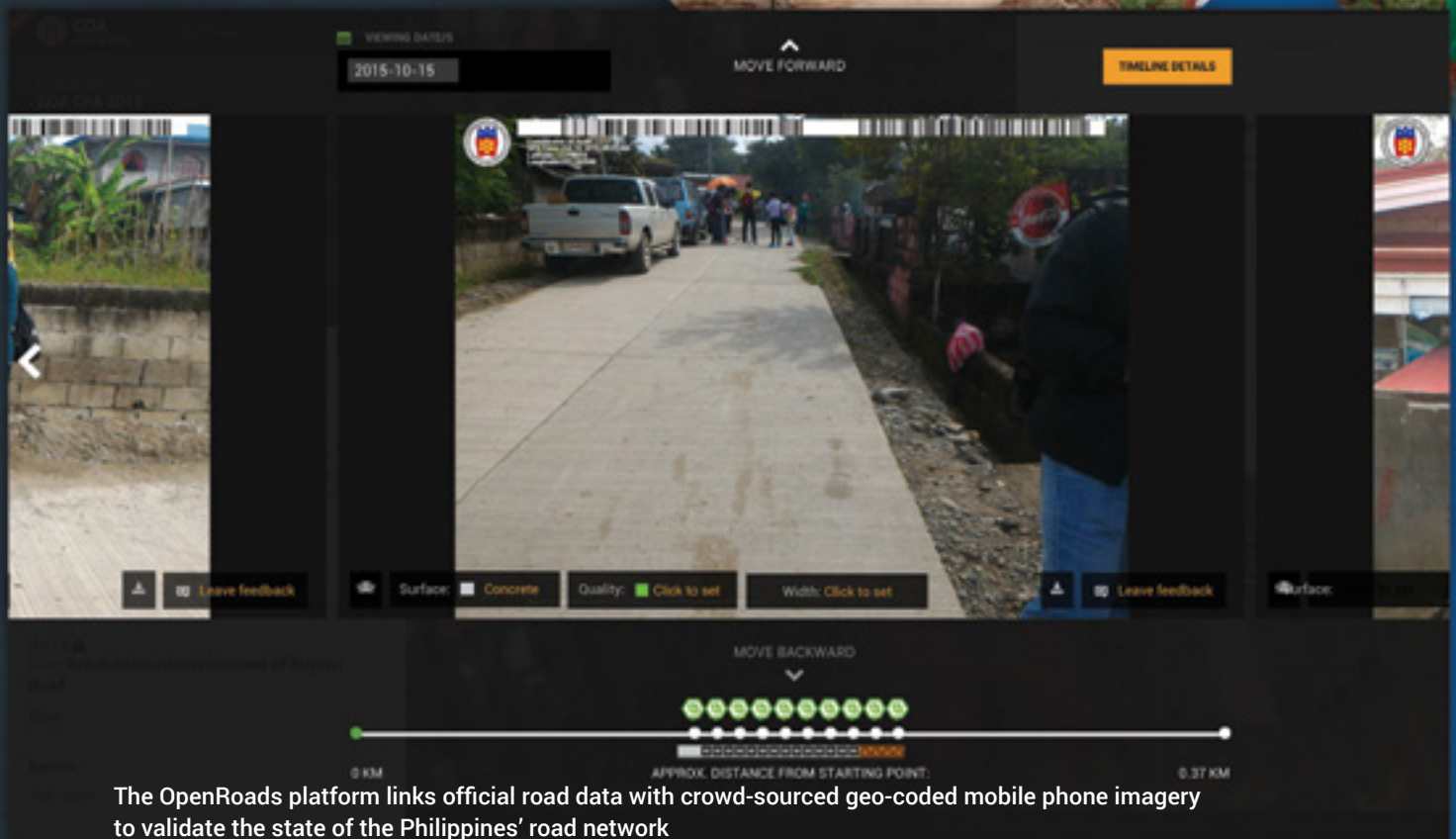
It's rare to have perfectly extractable data. There may be gaps or reliability issues, especially if data sources span different machines or archives. Different data (such as call details and DSF transactions) usually sit on different servers, which can complicate and prolong data extraction. Before starting the full extraction, request samples of data from all sources, to ensure they can be unified without problems. Even with the best preparation, extraction often takes longer than anticipated, so include potential delays in project planning.



Big data can help the World Bank reach its target of full global financial inclusion by 2020

Improving Road Investments through Mobile Data

Kai Kaiser





The Philippines' 180,000km of local roads are vital for linking people to social and economic opportunities

SUMMARY

Local roads that give 'last-mile' access to a destination are essential for promoting inclusive growth. But improvements in local road networks typically depend on national financing and involve numerous decentralized projects, making it hard for policymakers and citizens to ensure the development of efficient road networks. In response, this project developed OpenRoads, an interactive multi-media portal and set of digital tools to track nationally-financed projects to improve local road networks in the Philippines. The platform links official road data with crowd-sourced geo-coded video and image data from mobile devices, to validate the state of the country's extensive road network. It can be used by all stakeholders working to promote better road investments, including ordinary citizens. Leveraging low-cost open-source mapping technology and mobile pictures and videos, it organizes a wealth of rich geo-tagged imagery and feedback contributions, turning fragmented data into timely information. By geo-tagging, processing and analyzing the imagery, it captures the state of local road networks and road investment projects. This allows greater digital transparency and feedback, improving results and value for money for public investments in last-mile rural road access.

CHALLENGE

Beyond national highway networks, local roads that give last-mile access to a destination are vital for linking people to economic and social opportunities. But improvements in local road networks typically depend on national financing and involve large and diverse portfolios of decentralized projects. This makes it hard for policymakers and citizens to ensure that such programs are effective in prioritizing, delivering and maintaining efficient road networks.


In the Philippines only 31,000km (15 percent) of an estimated 210,000km of road are designated national roads. Improving last-mile access means upgrading the remaining 180,000km of local roads – a distance halfway to the moon. Several government programs worth almost US\$1 billion in 2016 target local roads, ranging from tourism to farm-to-market schemes. However, with patronage politics playing a key role in the country, it is hard to ensure that the right roads get built at the right cost. Traditional monitoring and evaluation systems for road projects cannot quickly and cost-effectively improve transparency and feedback.

Incomplete local road maps compound the problem, but geo-tagged photo and video data offer a reliable approach to mapping and assessing the Filipino local road network. In recent years, government agencies' collection of such data on public infrastructure programs (including through crowd-sourcing) has exploded. There is a need to effectively leverage this data overflow to provide real-time information on the road network and improvement projects. However, the data

is so unstructured that it cannot be systematically analyzed.

In response, this project sought to implement new data mining and evaluation protocols, develop a clear reference model for the road network, and ensure that image and video data can be referenced to this model. With the 2013-16 national budget for improving local last-mile roads at over US\$3 billion, the value for money in advancing transparency and feedback could be significant.

Users can see all nationally-financed local road projects and provide feedback by project or location



INNOVATION

The project developed OpenRoads, an interactive multi-media portal and set of digital tools to track nationally-financed projects to improve local road networks. The platform turns fragmented data into timely information, organizing a wealth of rich geo-tagged image, video and feedback contributions over time. Its premise is that visible, mapped information on existing road networks and the physical and financial progress of investment projects will

enhance transparency and accountability for public investments in last-mile roads.

OpenRoads leverages innovative, low-cost open-source mapping technology and mobile picture and video imagery. By geo-tagging, processing and analyzing the imagery, it captures the state of local road networks and road investment projects. The platform maps both projects and the road network assets these investments are seeking to improve. OpenRoads Mapping and Network Analytics builds on Open Street Map protocols, an open-source global mapping platform. This provides for systematic data analytics and validation, as well as public disclosure of information.

Joining up disparate data sources

Various Filipino government agencies are financing or implementing local roads projects. OpenRoads is not designed to duplicate these agencies' tracking systems, but rather to provide a platform for bringing this data together and augmenting it, particularly with rich geo-tagged information. For example, the Department of Public Works and Highways electronic project lifecycle system provides monthly summaries on the financial and physical completion of an estimated 20,000 projects per year. While many focus on the national road network, others benefit local roads. Open Roads augments this information with locational information, as well as videos and images. The system can also scour image and video data from various dates to show the lifecycle of road segments and projects.

The platform allows all stakeholders wanting better roads to create the missing digital road map using Routeshoot, a mobile video application that works on basic smartphones. After simple training, people can upload these movies and maps onto OpenRoads. The platform provides a comprehensive overview of government investments in local roads. It can be used by citizens concerned with the use of tax money, development and private sector partners promoting investments, or government bodies showcasing public infrastructure delivery. OpenRoads offers project updates and virtual tours at different points in time. Users can view national and local road networks, see all mapped nationally-financed local road projects and provide feedback by project or points on the map. This enables stakeholder collaboration in improving road networks.

Measuring program performance

The platform's Geostore stores mapped images and videos over time, allowing dynamic project-based review and updates by responsible agencies. Government bodies can map and geotag all projects, which oversight agencies can track systematically. Users can measure program performance by project or local government. The mobile app assigns point-or track-mapping to photos or videos, which geo-processing tools convert to diagrams summarizing project information (such as road length, surface and quality). Users can validate the quality of budgeted road projects against observed progress on the ground. The platform also offers road mapping tools to assess project connections and the extent of local road networks. Through

its Dashboard, OpenRoads allocates scores to project portfolios based on the availability of physical and financial information and the extent of basic project mapping.

By bringing projects to life through its dynamic digital maps, OpenRoads deepens information exchanges, allowing stakeholders to assess the road network and prioritize the right improvement projects.

RESULTS

OpenRoads is already making a tangible difference to the Philippines' road network. Local roads in Palawan Province have been surveyed using Routeshot, and the imagery geo-processed to rapidly summarize their condition by segment. The provincial government is collaborating with 24 cities and municipalities to complete comprehensive road mapping across the province. The Commission on Audit has engaged local civil society to conduct a citizens' audit of more than 200 Farm-to-Market roads across the country. OpenRoads' Geostore was used to manage the process, resulting in a People's Audit Report based on data analytics and visualization. The Tourism Road Infrastructure Program is upgrading over 4,000km of roads across more than 450 projects and 1,000 contracts. The Department of Tourism is using Routeshot and geo-processing to conduct a rapid appraisal and validation of all these roads.

Under *Kalsada*, a new public investment financing program, provinces are eligible for national government financing for up to two

provincial roads under the 2016 budget. A prerequisite is that agencies submit their proposals to OpenRoads using Routeshot videos. Over 70 provinces have been trained and are submitting projects. Worth nearly US\$150 million in 2016, the Kalsada program is set to increase in value six-fold in 2017. This 'No map, no money' approach from sponsoring agencies underscores the importance of accountability for taxpayers' money. The platform's approach could in future be applied to other decentralized infrastructure investments or special programs, such as re-greening.

OpenRoads is no substitute for major institutional reforms to deliver sustained effectiveness in last-mile roads investments, but better information and stakeholder engagement and feedback enable the country to check whether the right roads are being built at the right time and cost.

LESSONS LEARNED

OpenRoads shows how big data can bring stakeholders together to promote transparency, but technology cannot substitute for institutional reform to guarantee accountability.

- ***Complement big data with non-digital approaches to deliver change***
Online roads transparency is no silver bullet. Digital technology needs 'analog' complements, such as institutions that link transparency to accountability, to improve service delivery.

- ***Use big data to promote dialog***

OpenRoads promotes a multi-stakeholder conversation about how to improve planning, budgeting and implementation of decentralized road investment programs. Citizens can provide constructive feedback on progress in last-mile access for their communities. Mapping reframes conversations about road investments so they resonate with politicians, policymakers and citizens alike. Goals such as providing every community with road access can be costed, and choices evaluated.

- ***Be prepared for evolution in system usage***

The Kalsada program institutionalized the OpenRoads platform, linking the mandatory use of geo-tags to receiving performance-based grants. However, this also meant that OpenRoads increasingly needed to serve

as a project management system, rather than just a disclosure platform. The major lesson from this evolution was to ensure that the technology was both scalable and able to accommodate specific program requirements (e.g., tracking budget release requirements).

- ***Make road project geo-tagging comprehensive and mandatory***

The only way to avoid ghost roads is to ensure all local road projects are mapped. It was critical to underscore the need for OpenRoads coverage to be comprehensive. The Commission on Audit is enforcing this principle in the Philippines. Engaging Supreme Audit Institutions in geo-tagging must be an important part of advancing the notion that 'no road project should be a state secret'.



OpenRoads enables the country to check whether its roads are being built in the right place, at the right time and the right cost

Using Geo-spatial Data to Secure Property Rights



Kathrine Kelm



Ultra-light commercial drones record aerial imagery that can be quickly processed into accurate, cost-effective maps



The orthophoto produced from the drone and the software on the tablet are used to gather property information from local residents in Kosovo. The property boundary information is then updated on the orthophoto

SUMMARY

Property rights are critical to economic growth and social stability, yet almost 75 percent of the world's population lacks access to formal systems to register their land rights. In a new approach to recording property rights quickly and cheaply, this project used Unmanned Aerial Vehicles (UAVs), commonly known as drones. These record imagery which is processed into high-resolution orthophotographs (aerial photographs corrected to have the same lack of distortion as a map). The process generates accurate, cost-effective and up-to-date maps and 3D computer-generated landscape models in a fraction of the time of conventional aerial surveys. In Kosovo, the team used a UAV to map villages where the men were killed in the Balkan conflicts and the women lack formal property title. They are now using the new maps to help the women define their property boundaries and officially register their rights. They also deployed the UAV successfully in the fast-growing city of Ferizaj, to support a government program for unregistered land owners to legalize their property rights. The initiative can be scaled up globally, especially to secure land rights in developing countries.

CHALLENGE

Almost three-quarters of the world's population lack affordable access to formal systems to register and secure their land rights. Poor people, including indigenous and vulnerable groups, are disproportionately affected. Even where affordable formal systems exist, data quality often remains low, yet property rights are critical to economic growth and social stability. Without secure rights, land remains underdeveloped and underutilized. Access to secure land rights eliminates threat of eviction, increases investment and improves agricultural productivity. Women's property rights have been shown to improve children's health and education, foster inclusive family decision-making and reduce domestic violence.

There is urgent need to develop ways to identify and record land rights information far more quickly and cheaply than conventional methods. New technology is enabling the local capture of high-resolution geo-spatial data and processing into accurate maps. Combined with open-source software programs, this provides huge potential for a more cost-effective and inclusive approach to securing property rights. The project therefore aimed to produce faster, cheaper spatial data using processed imagery from Unmanned Aerial Vehicles (UAVs) to help the majority of

the world's population who lack secure property rights.

This mapping approach could be used to complete or update cadastral maps. By facilitating the registration of land rights, it would allow poor people to acquire a tangible asset, which they could use directly or as collateral to invest in other assets. Vital for economic and social inclusion, this security of tenure would help people escape the vicious cycle of poverty.

INNOVATION

UAVs offer a new approach to producing accurate, cost-effective and up-to-date maps and 3D computer-generated landscape models. More commonly known as drones, commercial UAVs are small and ultra-light, facilitating an affordable mapping service through a process that takes days or weeks from planning to product, rather than months or years. These maps can easily be disseminated and used by citizens, local government, utility companies and businesses, among others.

To refine the identification of land plots by UAV, the team built on a 2014 World Bank project which successfully used UAVs for mapping

The approach significantly reduces the cost and timescale of high-quality cadastral mapping, and empowers local communities to participate

and citizen engagement in defining property rights in Albania. They used drones to collect aerial imagery and produce high-resolution orthophotographs – aerial photographs geometrically corrected for topographic relief, lens distortion and camera tilt ('orthorectified'), so that their scale is uniform and there are no distortions. Unlike an uncorrected aerial photograph, an orthophotograph can be used to accurately measure distances. The team is now combining the orthophotos with open-source, customizable registration software to record property rights information from citizens. Together, UAVs and the new software offer an innovative and cost-effective property mapping and registration toolkit.

The process results in geo-referenced maps with boundaries and information on ownership and use for each property. These can be distributed to each owner and the local community, allowing land owners to easily identify and verify the boundaries of their properties. Although the maps do not constitute formal registration, the team also works with government agencies on how the field information can be integrated into the official cadaster or registration database.

Securing women's property rights

To scale the Albania pilot up to operational levels, the team turned to Kosovo, where the World Bank is helping the government produce a national cadaster system. Working with the Kosovo Cadastral Agency, the team began to integrate UAVs into the national mapping program. This would significantly reduce the cost and duration of cadaster development, and facilitate informed planning. The process would also empower local communities to participate.

For their first UAV deployment, the team worked to support the property rights of women in Krushe e Madhe, where most of the men and boys were killed in 1999 in the Balkan conflicts. The women have slowly rebuilt their lives and have organized several agricultural cooperatives. However, official data from the Kosovo Cadastral Agency shows less than 10 percent of properties in Krushe e Madhe are registered in a female name. This prevents women from using their land as an effective economic asset – in particular, as collateral for credit. The time, cost and complexity of conventional cadastral registration, along with poor knowledge about their rights, often exclude the women from the benefits of registration.

To map Krushe e Madhe's property boundaries, the team used a Sensefly eBee fixed-wing UAV, owned by the World Bank's Innovation Labs. The drone carries an 18 megapixel camera and flies within remote control contact at an altitude of around 100 meters. It covers a predefined area, prepared by setting survey 'ground control points' (marked with spray paint) that ensure the highest accuracy of the maps. In a week, the team completed 25 flights covering 12 square kilometers and processed the images locally, with the help of the Kosovo Cadastral Agency, into high-resolution maps of around three centimeters per pixel.

Mapping cityscapes

The UAV was also deployed successfully in a fast-changing urban context. In the past two decades many cities in Kosovo have experienced rapid, unplanned expansion resulting in informal settlements, illegal constructions and chaotic development.

In response, the government recently introduced a program for land owners to legalize their property rights. To facilitate property registration in the city of Ferizaj, the team spent a day carrying out six flights, covering three square kilometers in a total of three hours' flying time. The data were processed in 24 hours using two local high-end desktop computers, resulting in orthophoto maps with 1.9cm resolution from which land owners can easily identify their property.

RESULTS

The project showed that UAV technology supported by customized open-source software can produce accurate, cost-effective and up-to-date maps and ownership information for the registration of property rights. This approach significantly reduces the cost and timescale of high-quality cadastral mapping activities, and empowers local communities to participate by identifying and verifying boundaries on the maps. In Krushe e Madhe, the team is using the new maps to help the women define their property boundaries, and working with officials to develop a system for completing official registration using the maps and community information on property ownership. The maps and digital elevation model produced in Ferizaj will be made available free of charge to citizens who are participating in the legalization program, as well as to the municipality and other authorities via the national Geoportal (which offers web access to maps and other geospatial information).

The project also proved the versatility of the UAV approach. While flying near the

construction site for a new national highway, the team responded to a spontaneous request for assistance from a local official. The road crew had recently found an archaeological site, but existing aerial imagery and maps provided no evidence of it. Using the UAV, the team was able to plan, fly and process a high-resolution 3D map of the area in less than 24 hours. This provided accurate information for rerouting the road and preserving the cultural heritage site.

Other examples of UAVs' potential include utility inventories, supervision of major infrastructure contracts, post-conflict or disaster response assessment, recording the rights of indigenous and vulnerable communities, and road engineering. The initiative can be scaled up globally, especially to secure land rights in developing countries. The results will also inform the global discussion on how to help people currently unable to register their land rights and how to build sustainable systems to identify the way land is used.

LESSONS LEARNED

The project confirmed the potential of UAVs for collecting geospatial information. Paradoxically, the use of big data and powerful technology can empower people at community levels.

- ***Use drones for a wide range of aerial imaging requirements***

UAVs are becoming a more commonly accepted tool for producing high-resolution mapping products for targeted areas. They support 'fit-for-purpose' mapping principles, which hold that land administration should be designed to meet the needs of people and

their relationship to land, support security of tenure for all, and sustainably manage land use and natural resources.

- ***Empower local communities through new technology and big data tools***

Although a certain level of resources and capabilities is needed to process and manage big data such as the high-resolution geospatial information produced by UAVs, the project demonstrates how this new technology allows a more decentralized approach to traditional mapping. The

potential to empower communities and local government to visualize their environment and make informed decisions can be significantly increased by using these new tools and technology.



Drones enable production of accurate orthophotos in less than 24 hours for applications such as supervising infrastructure contracts, disaster response assessment and mapping archaeological sites (above, in Kosovo)

Open Traffic: Easing Urban Congestion

Stephanie Debere

*With thanks to Task Team Leader
Holly Krambeck (pictured)*



Urban traffic congestion affects poorer people disproportionately, as they have longer commutes and suffer more from the health effects of pollution



The Open Traffic platform can generate travel time survey data without the cost of manual fieldwork and analysis



SUMMARY

Congestion has known negative impacts on economic growth and can exacerbate urban air pollution and greenhouse gas emissions. Effectively addressing congestion requires accurate traffic speed and flow data, but resource-constrained transport agencies are challenged to collect these data, as modern tools tend to be financially and technically out of reach. In response, the Open Traffic program leverages open-source software and innovative partnerships to substantially reduce the cost of traditional traffic data collection and analysis, while simultaneously improving the quality. The first scalable, open-source program of its kind, the project built on work with the Cebu City Government in the Philippines to develop an open-source platform for collecting, visualizing and analyzing traffic speed data derived from taxi drivers' smartphones. Using GPS data from an on-demand taxi service, Open Traffic successfully analyzed peak-hour congestion, travel time reliability and corridor vulnerability across 10 Southeast Asian cities, and has prepared travel time analyses for select origin-destination pairs. This analysis would not previously have been possible without substantial time and resources. It shows that the next generation of congestion management solutions will leapfrog the capital-intensive approaches of the past, enabling traffic management agencies to make affordable, evidence-based planning decisions. Open Traffic has now been deployed in Cebu City for live testing.

CHALLENGE

Urban traffic congestion affects poorer people disproportionately. They generally have longer commutes than the affluent and suffer more from the health effects of higher pollution, as many work outside. Congestion also generates excess greenhouse gas emissions, and it is often the poorest people who live in areas most vulnerable to climate change. Time lost in traffic jams also has a significant negative impact on urban GDP growth.

In many developing countries, decisions about traffic signal timing plans, public transit provision, roadway infrastructure, emergency traffic management and travel demand management are made without observed, quantified congestion or travel-time data. Such data is costly to collect and can also require substantial technical expertise to analyze. This causes avoidable congestion, as well as unnecessary fuel consumption.

In higher-income countries, transport agencies rely on a combination of manual survey methods and installed physical sensors – underground detector loops, pneumatic tubes, laser-based sensors, cameras and Bluetooth device detectors. However, these require initial capital outlays, ongoing maintenance and

technical expertise beyond the capacity of poorer cities. They can also record data only in places where they are deployed – select corridors during select time periods. There is urgent need for a viable, inexpensive alternative to traditional travel-time and congestion data collection and analysis. This would allow resource-constrained agencies to make evidence-based decisions to promote traffic flow.

INNOVATION

The project leveraged three trends to develop a traffic management system reliant on GPS data instead of fixed-location equipment: Growth in global smartphone usage, the emergence of taxi-hailing app companies and increased use of open-source software.

Over a third of the world's population is expected to have a smartphone by 2017. This has inadvertently created a new source of traffic data, derived from handset GPS signals and Wi-Fi pings. Viewed as traffic probes, smartphones can create a sensor network that is unrestricted to specific corridors, is continuously updated in real time, requires no maintenance and provides a level of sampling unachievable through manual methods or equipment-based sensors.

The project leveraged three trends: Growth in smartphone usage, the emergence of taxi-hailing app companies and increased use of open-source software



Recently, international smartphone-based taxi-hailing app services have also emerged. These companies maintain databases of millions of urban GPS points, often spanning hundreds of cities across many countries. The project combined smartphone 'sensors' and taxi GPS databases to develop a single cloud-based traffic management application, which could support services in numerous cities simultaneously. By using open-source software, it offers unprecedented economies of scale in capturing and analyzing traffic data.

Linking disparate data sources

The initiative builds on a successful pilot in Cebu City, where the team created an open-source platform that uses GPS data generated by taxi drivers' smartphones to derive meaningful statistics for traffic planning. The platform, called Open Traffic, is a graphical user interface allowing government agencies to easily query and visualize stored traffic statistics derived from GPS data collected from drivers' phones.

The team partnered with Malaysia-based Grab, the largest taxi-hailing app company in Southeast Asia, to further develop and pilot the smartphone 'traffic sensor' approach. Through the partnership, traffic management agencies in Malaysia, Singapore, Indonesia, Vietnam, the Philippines and Thailand will have access to anonymized traffic data generated by 250,000 vehicles in Grab's fleet, free of charge for at least the two-year pilot and scaling-up phases of the project.

The platform uses Open Street Map (OSM), a global geographic dataset populated

by volunteers without cost or licensing requirements. The map may be freely updated and improved by transport agencies and others, using open-source editing tools. Its 'Highway' feature includes all OSM mapped roads, from unpaved rural tracks to expressways, covering much of the planet. Drawing on the OSM Highway, Open Traffic links average traffic speed calculations to OSM road segments via several steps:

- Open Traffic downloads the relevant portion of the global OSM map.
- It prepares the map sections by assigning virtual 'detectors' to every approach where road map segments intersect.
- It calculates the travel time for a single vehicle traversing a road segment across two detectors, as the distance between the two detectors divided by the time the vehicle spent traveling between them.

From raw data to travel times

The estimated travel time for each road segment on a given trip is stored on a server. Neither raw GPS data nor information associated with a particular vehicle is retained. Data are stored as the number of travel times for every hour of the day (how many travel times of, for example, five, six or seven kilometers per hour, etc.). These travel times can be queried to calculate average traffic speed for different time specifications (a specific day, a specific hour each day, etc.) for single or multiple road segments.

Open Traffic can query the database of stored travel times by road segment to generate a map of average travel speeds for selected time periods. It also facilitates travel-time queries

between select origin and destination pairs, either for automatically generated routes (based on the shortest path) or manually defined ones. A ‘confidence indicator’ is provided, based on the number of observations used to derive the travel time and average speed estimates.

RESULTS

Using GrabTaxi’s data from 10 major Southeast Asian cities, the team tested whether Open Traffic’s analytical results made intuitive sense. They used the platform to observe weekday peak and non-peak travel patterns in each city. These peak-hour graphs mostly reflected expectations about urban traffic, with travel speeds highest at night and slowest during commuting times. The results meant Open Traffic could be used to monitor the efficacy of congestion mitigation measures.

The team also tested the platform’s suitability for examining peak-traffic duration and variation, conducting travel-time surveys and understanding how externalities and traffic interventions affect traffic speed. Tests successfully examined the predictability of congestion, checking the consistency of expected travel times between origin and destination pairs along key corridors.

The Open Traffic platform could also be used to generate inputs such as travel-time survey data for traditional transportation planning, without the cost of fieldwork, encoding and analysis. The team compared manual survey data in Cebu City to Open Traffic data, and found that the Open Traffic results provide less variation between road segments. This is unsurprising,

as the traditional survey represents only a single sample, whereas the Open Traffic dataset represents thousands of samples over the same time period.

These results show that the next generation of congestion management solutions will leapfrog the capital-intensive approaches of the past. They will enable traffic management agencies to make better, evidence-based decisions about traffic signal timings, public transport, road infrastructure, emergency traffic management and travel demand management.

The Open Traffic platform has now been deployed in Cebu City for live testing. Next steps include development of a methodology for optimizing traffic signal timing plans using GPS data instead of traditional sensors, as well as a standardized methodology for estimating the cost of congestion (in terms of fuel usage, greenhouse gas emissions and economic impact). Discussions are underway with Grab on launching the platform in other cities.

LESSONS LEARNED

Open Traffic illustrates that much of big data’s potential lies in combining existing disparate sources of information in unprecedented and innovative ways.

- ***Keep seeking the potential in new combinations of data***

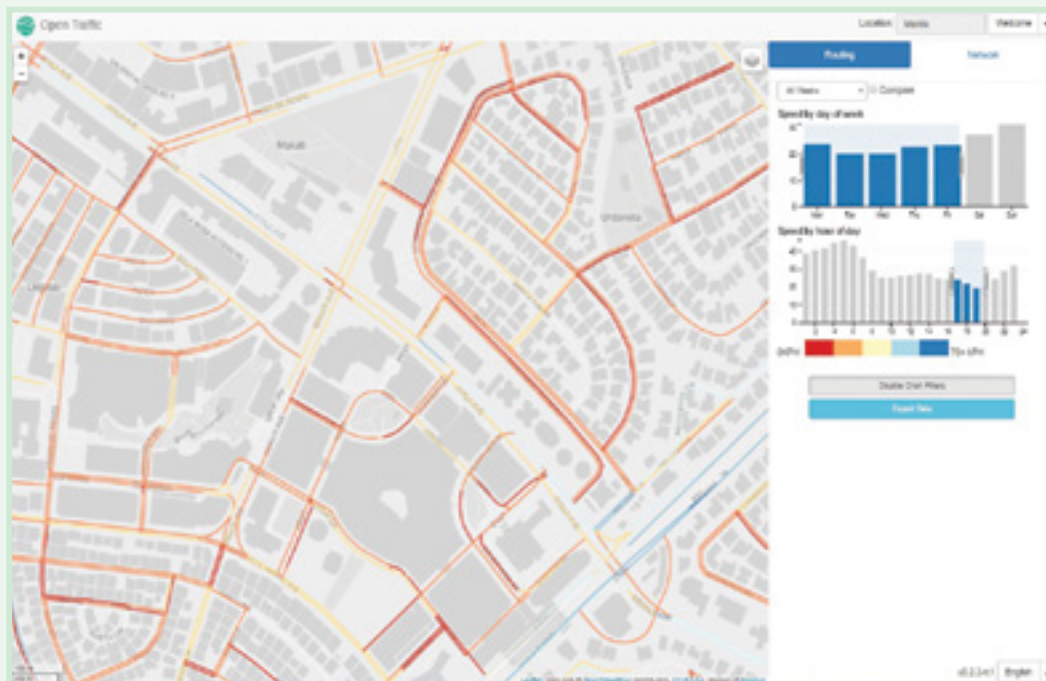
By bringing together existing independent data sources, the team was able to mine a rich new source of information. To exploit big data fully, it will be important to remain alert to the potential of combining disparate and

even seemingly unrelated sources of data for analysis.

- **Think about data storage from the beginning**

Big data analytics can generate unwieldy amounts of data. Open Traffic data was initially stored as city-specific files in Amazon 'buckets', a cloud-storage service which facilitates automated uploading and downloading of data directly from and into

applications and databases. However, this required Grab to set up specialized buckets for the project, and the process and file sizes soon became unwieldy. On the team's recommendation, Grab began aggregating its global data as a single stream, rather than individual files, using Amazon's Kinesis service, which can upload real-time data streams from multiple sources.



Open Traffic can query the database of stored travel times by road segment to generate a map of average travel speeds for selected time periods

Observing People's Feelings About State Institutions

Victoria L. Lemieux



To explore links between citizen's feelings and civil unrest, the project analyzed tweets made during protests around the 2014 Soccer World Cup in Brazil



SUMMARY

BIG DATA *innovation challenge* 31

CHALLENGE

Civil protest has been described as moving from discontent of the populace, to politicization of that discontent, to actualization of frustration as aggression against the state. Such unrest can lead to fragile states, corruption, terrorism and economic impediments. It can be causally linked to poverty, but dissatisfaction with institutions can also be relevant. In countries where citizens lack means of voicing discontent with institutions and governments, they can embrace extreme forms of protest that quickly escalate – as demonstrated by the so-called ‘Arab Spring’.

Typical methods for studying internal conflict include qualitative case studies and econometric analyses. However, case studies can be dangerous and costly, and econometric approaches can overlook localized conditions and the evolving dynamics of protest. There is a need for granular information about local conditions and the causal chain of events, without the dangers and costs of case study approaches. Analysis of microblog data has the potential to deliver this information. This project sought to discover whether such analysis could uncover how citizens feel about their institutions and government, and how these sentiments translate into collective behavior.

Evidence from online media data offers an innovative approach to investigating social and political issues



INNOVATION

The team chose to analyze Twitter postings in Brazil during the 2014 Soccer World Cup. The country has relatively high levels of inequality and high social media use. It is the world’s second-biggest Twitter user (with roughly 41.2 million tweeters). Costing up to an estimated US\$14 billion, the World Cup was the country’s then-largest and most expensive sporting event. During and afterwards, it sparked public protests in several Brazilian cities.

To explore whether the protests related to citizens’ trust in state institutions, the project produced an innovative visual analytics tool and a novel analytic methodology. Visual analytics combines human reasoning with machine reasoning through an interactive visual interface. This ‘mixed initiative’ approach overcomes the limitations of each type of reasoning. Human analysts can detect subtleties of humor or satire that a computer might miss, and computers can rapidly process and manipulate volumes of data that humans cannot.

How do you feel?

Through sentiment analysis and text analysis (machine reasoning) of a large-scale collection of tweets, the project tracked general public distress and trust in institutions, hypothesizing that increased negative sentiment signaled declining public trust in government. The team collected publicly available Twitter data, in two phases. The first was an initial 'big picture' harvest of approximately 11 million tweets. The sample was then analyzed using Natural Language Processing and a visual analysis tool that clusters documents together by determining key themes in each. Collections of texts are displayed like a galaxy of stars, in which each star is a single document, and clusters represent document similarity. Through this, the team identified tweets on political opinion and analyzed them to define 'naturalistic' search terms representing key concepts (such as government and institutions) underlying the study. These search terms were then used to harvest historical tweets for the 2014 World Cup period.

The team performed sentiment classification of the harvested tweets using SentiStrength, a tool which classifies text in terms of positive and negative sentiment. It can also compute a single evaluation based on both positive and negative classifications. The project used this single-sentiment approach, resulting in nine sentiment categories: Positive or negative with a magnitude of 1 to 4, and neutral. The results of the sentiment classification were visually presented as horizon charts, allowing analysts to see how sentiment varies in polarity and intensity over time. From a basic text analysis of historical tweets, they depicted the top 500

word-frequency terms in a text-cloud, as well as the single use of terms over time.

Translating sentiment into behavior

Results of the sentiment classification and text analyses were then represented visually, so analysts could detect patterns in the data (human reasoning). They used a structured analytic methodology, aided by the tool's interactive features, such as the ability to search, sort and see data from different views. The sentiment analysis enabled the team to observe patterns in the data such as high negative or positive sentiment toward a particular institution, increasing or decreasing positive or negative sentiment over a period, or correlation of negative sentiment about an institution with rising negative sentiment about government as a whole. They explored correlations between patterns of sentiment, government policies and observed citizen behavior (such as protests).

Using the visual analytics tool, the team carried out a pair analysis to make basic observations from the data. This analysis pairs a subject-matter expert and a visual analytics expert, combining contextual knowledge with technical expertise. The tool was also used for an analysis of competing hypotheses, a methodology which analyzes the degree to which evidence supports the relative likelihood of alternative hypotheses. From background literature, the team identified 68 hypotheses on the relationship between citizen trust and social protest. The analysis drew tentative conclusions about the relative likelihood of each by trying to disprove rather than prove each hypothesis.

RESULTS

By connecting the visual analysis to hypotheses on the relationship between citizen trust and social protest, the study found support for the relative deprivation theory of social protest. This suggests that an individual or group lacks something that another group has and to which they feel entitled. Deprivation is felt in relation to some external standard, not in absolute terms.

The methodology showed that around the 2014 World Cup, Brazilians expressed negative sentiment about low investment in services such as education, health and water, relative to spending on the tournament. At state level, water was a key issue, with tweeters criticizing investment relative to politicians' spending on priorities such as campaign financing. The negative tweets themselves constituted social protest, as well as forming part of larger politicized groups using protest hashtags and calling for other forms of protest (such as demonstrations).

While media reports and subsequent studies of the 2014 World Cup protests generally focused on single immediate causes, the analysis showed that the protests sprang from a range of long-standing grievances, coupled with relative deprivation triggered by spending on the World Cup and campaign financing. This sense of deprivation fueled sentiments that activated protest.

The project offers an innovative approach to investigating social and political issues. The methodology has already been used in a World Bank evaluation of higher education in Brazil, enabling analysts to link how citizens feel about

higher education to existing hypotheses about education in developing countries. Evidence from online media data suggests further avenues for research using complementary methodologies, such as surveys or comparative analysis with other data sources. The approach could also support the development of an observatory of citizen sentiment to inform investment decisions around strengthening institutions. It could ultimately lead to predictive models for negative sentiment towards particular institutions.

LESSONS LEARNED

The project demonstrated the potential of big data analytics to contribute to knowledge about development, as well as for evaluating development outcomes.

- ***Combine human and computational power for optimum results***

Visual analytics are useful for unpicking complex socio-political issues, such as those surrounding citizen trust. Alongside computational methods, analysis of online social data requires human input, supported by interactive visual interfaces.

- ***Build strong partnerships***

Development of this approach required close collaboration with experts in fields such as Brazilian society, history and government. Using 'design thinking' techniques, the project team partnered with experts to test assumptions and obtain feedback on approaches and results. This improved the project outcome.

- ***Use big data analytics alongside other research approaches***

Big data analytics makes observations possible from a distance, both in terms of space and time, and the project provided insight into Brazilians' thoughts, as expressed naturalistically. In contrast, surveys require researchers to spend time on the ground, and may prime citizens with questions that do not reveal their own thoughts. However, this type of big data analytics does not offer a representative sample, and is best used to complement other approaches to understanding development issues.

- ***Protect individuals' privacy***

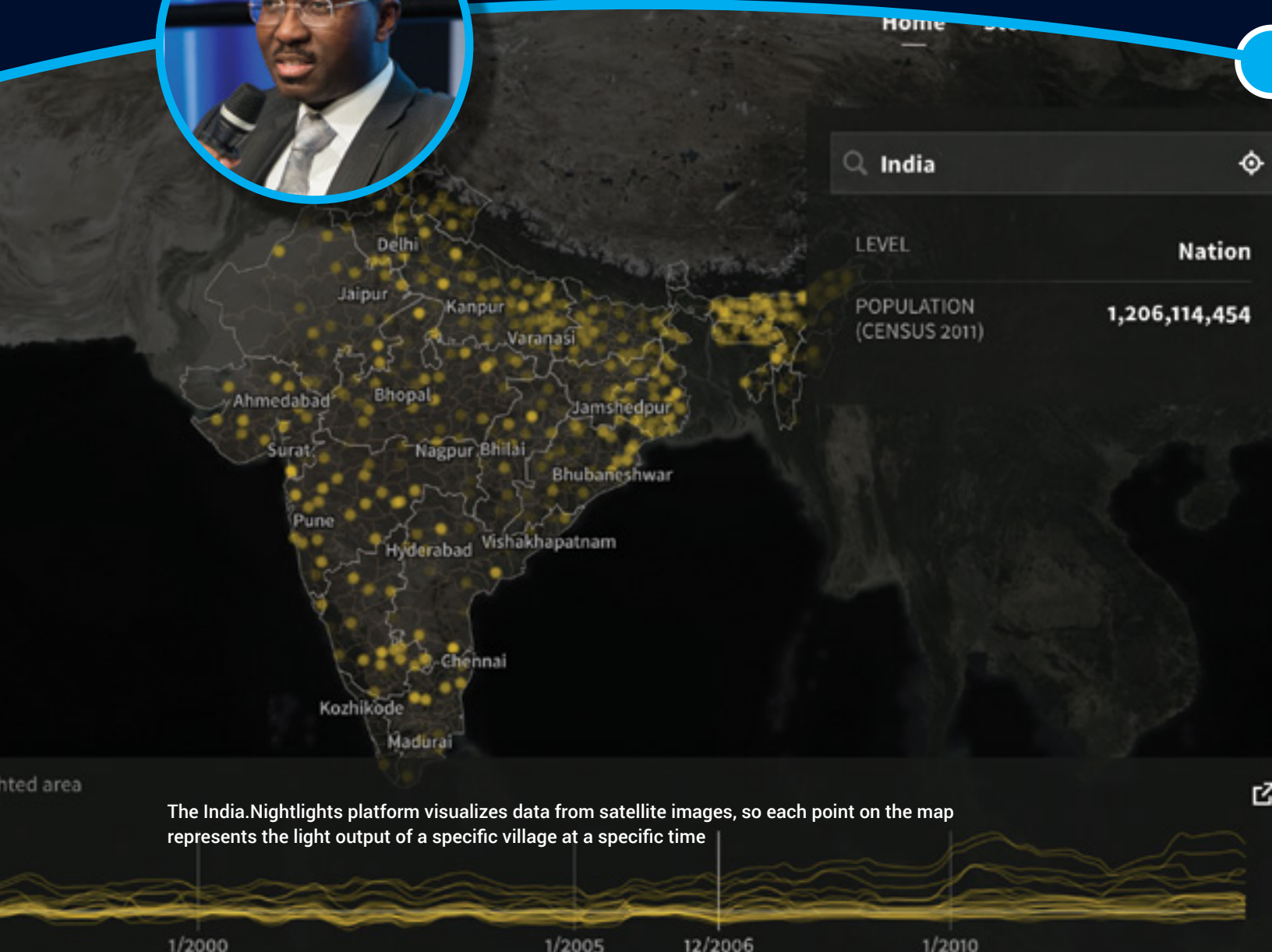
Twitter users do not consider use of their data for research when they tweet. Many tweets sent impulsively – especially during social protest – contain candid or critical remarks. Take care to protect citizens from potential harm, such as retaliation for expression of opinion or lawsuits for defamation.



Where citizens lack means of voicing discontent, they can embrace forms of protest that can quickly escalate

Monitoring Rural Electrification from Space

Kwawu Mensan Gaba



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



The approach shows the great potential of satellite-based monitoring to radically transform electrification planning

SUMMARY

Electricity is essential to human wellbeing worldwide, yet 1.2 billion people still live without it. Key to improving service provision is accurate tracking of the availability and supply of electricity at local level. By collecting and analyzing a unique historical archive of nighttime satellite imagery, this project developed a novel data-intensive strategy to improve the monitoring of electricity provision to rural areas across the developing world. Drawing on a multi-terabyte image archive spanning over 8,000 nights since 1993, the team used computationally intensive methods to extract and analyze patterns of light output observed nightly over all 600,000 villages in India. This pioneering dataset paints a dynamic portrait of rural energy access over two decades, enabling observation of how access to electricity has expanded, and identification of villages that remain dark. It also enables the detection of power supply irregularities. These insights are particularly useful in rural and remote regions where traditional monitoring is difficult. To make the data accessible to governments, power companies, regulatory agencies and other users, the team developed an online visualization platform, India.Nightlights. The site allows users to see how light output has evolved over two decades, from state level right down to individual villages. Ultimately, the approach could help optimize electrification planning according to an accurate understanding of electrical supplies on the ground.

CHALLENGE

In much of the world, access to electricity is uneven and irregular, undermining development and welfare. Rural electrification and lighting improvement projects are high on the development agenda. These are regularly monitored and evaluated, but there have been no mechanisms to track the sustainability of electrification schemes after projects end, or to identify easily and precisely who has electricity and who does not.

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 has been no systematic effort to study the entire raw nightly data stream. This stream reveals the distribution of electricity at high resolution over the last two decades.

In 2011, a team from the University of Michigan, the US National Oceanic and Atmospheric Administration and the World Bank Group's Energy and Extractives Global Practice began

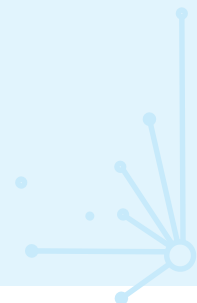
to explore how to use night lights data in a scalable, systematic way. Their early work focused on validating ('ground-truthing') the relationship between satellite-detected light output and the use and the availability of electricity in several hundred villages across Senegal, Mali and Vietnam. The 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.

INNOVATION

The team's pilot studies explored the use of night light data to monitor rural electrification in countries with low electrification rates, such as Mali and Senegal. This was expanded to Vietnam, which has near-universal electrification. Following the Big Data Innovation Challenge, the team refined its approach and scaled it up to look at all of India, a country with a high density of villages and a major rural electrification program. This project took two parallel tracks, in close collaboration:

- **Mapping India's power supply**
With its high density of villages and a flagship national electrification strategy, India was an ideal country to assess the

The resulting dataset represents the most comprehensive database known describing electricity access and variability



validity and reliability of a satellite-based approach for monitoring rural electrification over time. The first step was to evaluate the large-scale electrification program, launched in 2005 to bring power to over 100,000 villages.

The team acquired the complete historical archive of nighttime satellite imagery from the Defense Meteorological Satellite Program, run by the National Oceanic and Atmospheric Administration. 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 5 billion observations represents the most comprehensive database known describing electricity access and variability. It enables new analysis, exploration of signal-processing techniques and generation of data visualizations that better capture space and time patterns in electricity distribution.

Drawing on official electrification program records, the project linked newly electrified villages to their nighttime light signatures, covering around 8,000 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 analysis

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 regional instability in power supply, increased incidences of power cuts, and indications of electrical supply problems as they occur.

- ***Creating visual tools***

Building on the insights gained from the satellite data study of India, the team developed an online toolkit to provide power companies, regulatory agencies and relevant partners with geo-referenced maps and satellite imagery depicting current patterns and recent trends in electricity supply. It visualized electrification trends on a web platform, India.Nightlights.io.

The open-source platform comprises a pipeline to process massive amounts of data, an application programming interface that enables technical partners to query light output at village, district, region or state levels across India, and a dashboard map to allow users to explore light output trends. The platform offers high-level overviews or enables users to compare villages, plot trends and share data. Freely explored from any part of the world, it has the potential to be a powerful tool in driving rapid electrification.

Each point on the map of India represents the light output of a specific village at a specific time. At district level, users can filter to view villages that have participated in India's electrification program and see changes in light output, which can be used to complement research about electrification in the country.

The platform was tested by various users involved in expanding electricity supply, including private firms, universities, regional governments, non-governmental organizations and development partners. It was then refined ahead of a public launch at the World Economic Forum in Davos, Switzerland, in January 2016.

RESULTS

The project demonstrated that nighttime satellite imagery can be reliably used to detect the use of electricity in the developing world, even in rural contexts where electricity use is characterized by low power loads, small numbers of dispersed users, limited infrastructure and erratic service provision.

India.Nightlights presents an online platform that enables visualizations and interactive exploration of the night lights data over India. The team now wants to refine the platform, build new capabilities and generate nuanced reports to meet the myriad needs of potential country-level beneficiaries. It also wants to see how this approach could be replicated across the developing world.

The platform shows the great potential of satellite-based monitoring to radically transform rural electrification planning and assessment. Drawing on satellite-based data will sharpen program targeting in the village selection process, improve implementation assessment and allow ongoing monitoring by interested parties after projects have officially closed.

Next steps include promoting adoption of the tool and exploring where to focus future electrification efforts. The team is looking at where electrification has been successful, what other variables are related to faster electrification, and whether other development indicators can be added to the platform's dashboard. The tool could be used in poverty analysis, as the geo-referenced data can be combined with other databases, drawing more correlations between electricity access and development outcomes.

LESSONS LEARNED

The project's success rests on the importance of validation when pioneering big data approaches, and of persevering when faced with hurdles.

- **Validate novel data approaches thoroughly**
Rigorous validation or 'ground-truthing' is imperative to establish confidence in novel data sources and new methods. Because the team was able to demonstrate in its earlier work in Senegal, Mali and Vietnam the strong correlation between light outputs captured in satellite imagery and electricity supply on the ground, the project generated confidence in its approach.

- **Take the long view**

Perseverance is essential. The overall process took five years to reach the stage of publicly launching the web platform. The team had to overcome several hurdles in the process, in particular, data availability and processing requirements to ensure quality data.

- **Partnerships are crucial for success**

Working with other organizations brings valuable insight from new perspectives, and can generate solutions from unexpected quarters. In Vietnam, for example, the agency

that ran the survey also provided access to electricity consumption data not publicly available. For the visualization platform, the team ran a competition to select the website developer, having benefited from free technical assistance from the GIS mapping software company and contributions from other World Bank global practices to formulate the terms of reference.

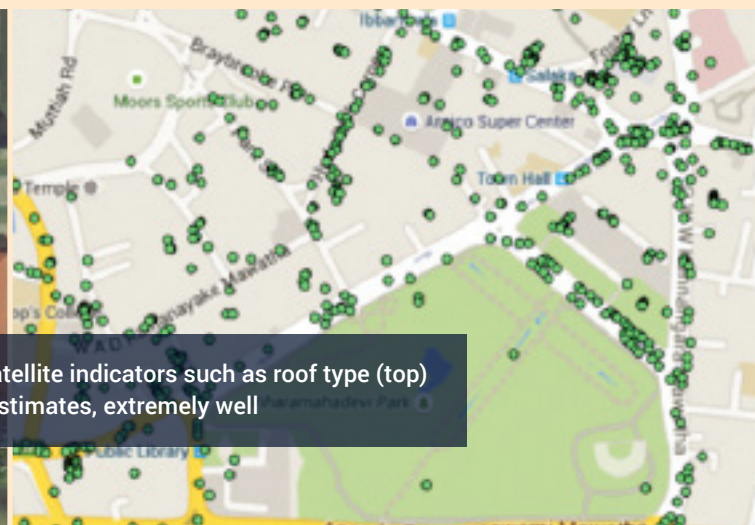
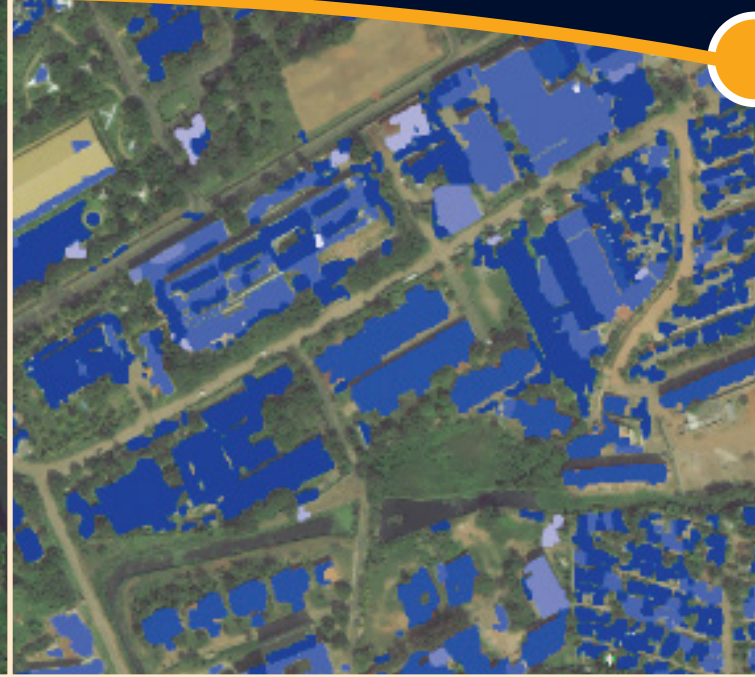
www.india.nightlights.io

Around 1.2 billion people worldwide still live without the electricity essential to human wellbeing



Mapping Poverty by Satellite

David Newhouse



Preliminary results show that in Sri Lanka, high-resolution satellite indicators such as roof type (top) and density of cars (below) track poverty, based on census estimates, extremely well



Satellite-based indicators improved the accuracy of poverty prediction in Sri Lanka (which has few existing poverty indicators) more than in Pakistan (which has many)

SUMMARY

Poverty must be located accurately if development interventions are to be effectively targeted and monitored. However, expensive data collection and processing mean national poverty estimates in developing countries are often outdated. In response, this project explored the use of indicators derived from satellite data to predict geographic variations in poverty. The first component 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. When satellite indicators were applied to the models in Pakistan, which is unusual in that it can generate 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, which is 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 an innovative statistical technique called Lasso to work best for predicting poverty at a more local level, with sizeable benefits when there are many variables.

When high-resolution satellite data indicators – such as cars, built-up area, shadows, roof type and road type – were combined with Sri Lankan census data, preliminary results showed that satellite indicators track regional differences in poverty extremely well. They demonstrate that high-resolution satellite imagery is a valuable complement to household survey data, with potential to help generate more accurate and updated local poverty maps and refine targeting in development initiatives.

CHALLENGE

Development interventions can be more effectively targeted and monitored if poverty can be located more precisely. However, long lags in processing and report-writing mean that recent national estimates of poverty in developing countries are often several years old. In addition, 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 way to build a prediction model. Although numerous models have been developed, there has been little rigorous comparison of different approaches.

This project aimed to assess different approaches to poverty mapping, as well as the extent to which high-resolution satellite imagery can be used to generate more accurate poverty estimates. Incorporating satellite data into poverty mapping is a first step towards using the wealth of non-traditional data generated daily to predict poverty more effectively. Satellite-based data analysis is particularly attractive because it can see a complete picture of a particular area, unlike, for example, mobile phone-based analysis, which typically captures only a subset of phone users. Satellite data can also be collected frequently at fine geographic levels, even in conflict areas not conducive to surveys. Using this data to better understand poverty can help development practitioners target interventions and evaluate their effectiveness more accurately. Satellite-enhanced maps would be a key step towards

the goal of real-time estimates of how pockets of poverty are evolving.

Satellite imagery gives insights into factors such as the scale of urbanization, infrastructure and natural resources.



INNOVATION

The study involved two stages:

- The first compared different methods of generating poverty prediction models, in both Pakistan (which has existing data for many poverty indicators) and Sri Lanka (where fewer are available).
- The team then examined how high-resolution satellite data indicators are correlated with poverty predictions based on the 2011 Census, for local administrative areas in Sri Lanka.

Assessing poverty prediction models

To examine different approaches to developing poverty prediction 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 those withheld areas. This technique was used to compare the accuracy of models derived using manual selection, stepwise regression and Lasso-based procedures. The team also augmented the set of prediction variables with publicly available low-resolution satellite data to see whether this improved traditional poverty mapping techniques.

Applying higher-resolution imagery

In the second stage, the team purchased high-resolution (0.5m per pixel) satellite imagery covering approximately 5 percent of Sri Lanka, including both rural and urban areas, and containing roughly 1,400 local administrative divisions. They used multispectral imagery (multiple images taken at varying spectrum wavelengths) 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 objects such as cars from such imagery. These additional predictors have not yet been tried in poverty mapping models. Because some indicators, such as car traffic, can change rapidly as economic growth occurs, these additional predictors could pave the way towards more frequent poverty estimates in the future. Contractors produced pan-sharpened mosaics (merging several smaller scenes) of the raw high-resolution imagery. The team then worked with experts to develop detection algorithms to identify possible poverty predictors that can be extracted from high-resolution satellite data. 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 calculated whether buildings were more rectangular or had more chaotic angles (indicating higher poverty) and constructed indicators such as the share of paved roads or built-up area. They assessed the density of each feature in each local district division and correlated the satellite-based measures with poverty estimates from the 2011 census data. This provided a measure of which indicators correlate most strongly with predicted poverty and other measures of economic welfare from the census. The process demonstrates the potential of high-resolution satellite data to capture new indicators correlated to poverty.

RESULTS

The assessment of poverty prediction models showed that the number of poverty indicators affects the performance of different models and the usefulness of adding satellite data to the indicators. In Pakistan, with many potential indicators, the team found that Lasso models outperform both discretionary and stepwise models. However, Lasso and stepwise models give comparable results in Sri Lanka, where the set of indicators is smaller. The accuracy of the prediction model also depends considerably on the poverty threshold. In Sri Lanka, models were better able to predict the bottom 40 percent than the bottom 10 percent, but in Pakistan the reverse was true. In Sri Lanka, including publicly available satellite data made poverty predictions more accurate, but in Pakistan, the satellite data makes predictions slightly less accurate. When the satellite data is included in Sri Lanka, the Lasso models significantly outperform the manual and stepwise models.

Overall, the team found Lasso-based models are preferred for generating poverty predictions, and that the benefits can be sizeable when the pool of candidate variables is large, as in the case of Pakistan and of Sri Lanka when satellite indicators are included. There is strong interest in testing different poverty modeling approaches and the research highlighted the value of using publicly-available satellite data to generate small-area estimates of poverty in contexts where census data is limited.

Understanding pictures of poverty

For the Sri Lanka study using high-resolution imagery, preliminary results show that indicators track regional differences in poverty, based on estimates from the census, extremely well. When looking at the full sample, the key indicators relate to building density and urbanization, including the number of buildings, a vegetation index, and in rural areas, the share of roads that are paved, shadow, and type of roof. But when predicting variation in poverty in local areas within urban areas, the number of cars and an abstract measure of rectangular buildings become strong predictors as well.

These preliminary results demonstrate that satellite-based data is a valuable complement to household survey data, strengthening the case for investing in high-resolution imagery to monitor poverty more generally, as well as project impacts. This approach is the first step of an exciting research agenda. Imagery can deliver new insights related to a variety of development challenges, such as the scale of urbanization, infrastructure and the state of natural resources. Much 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 tradeoff between the quality and cost of the imagery on the one hand, and its benefits in terms of predicting local variation in poverty. Eventually, satellite-based imagery could also be a valuable tool in improving measuring inequality, monitoring development projects and 'nowcasting' poverty rates.

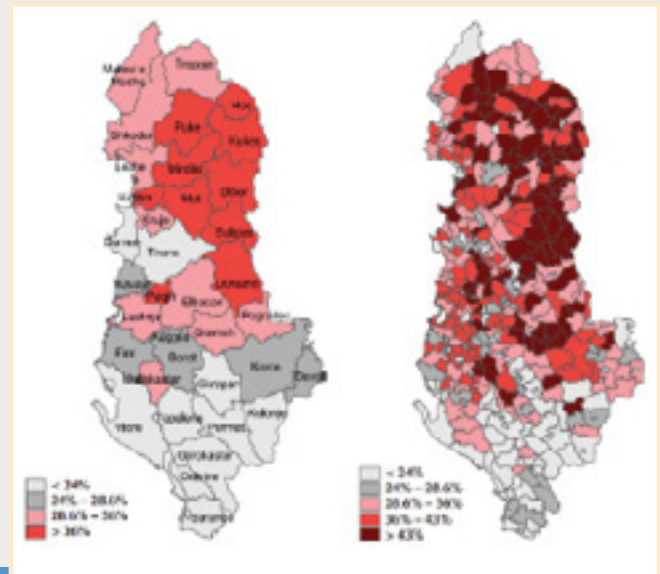
LESSONS LEARNED

As the price of high-resolution imagery continues to fall and coverage improves, satellite-based data will become an increasingly useful source of information about welfare in developing countries

- ***Drive research into mainstreaming the use of satellite imagery in poverty measurement***
Most poverty economists are unaware of satellite technology's potential to improve small area estimates. There is not yet sufficient evidence to mainstream the use of satellite imagery to improve poverty maps based on census data, but this project has increased awareness. Beyond poverty prediction, satellite imagery can help efforts to better understand poverty. For example, road network estimates can indicate whether new road construction benefits the poor.
- ***Allocate sufficient resources to sourcing raw satellite imagery***
Navigating the market for high-resolution satellite imagery and developing relationships with vendors and processing

experts can take longer than anticipated.
Ensure adequate project funding and planning to allow for the time this can take.

- **Watch for future potential in satellite imagery**
The accuracy and timeliness of satellite-derived poverty maps will continue to improve. More work is needed to see which satellite-based measures best predict poverty, both spatially and across time. There is broad scope for fruitful collaboration between poverty economists and geo-spatial image experts.



Poverty must be located accurately if development interventions are to be effectively targeted and monitored

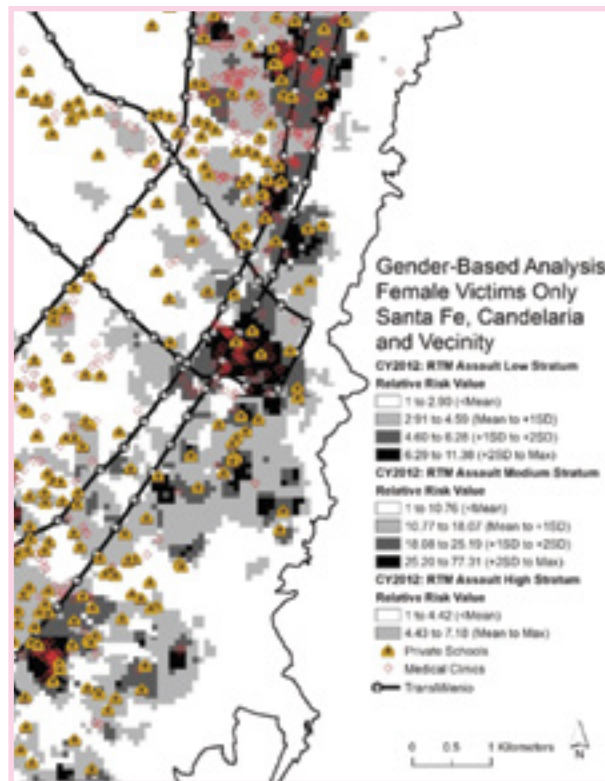


Understanding How Infrastructure Affects Crime

*Camila Rodríguez
and Andrés Villaveces*



Using geo-coded data to generate risk terrain models, the project identified specific urban features associated with violent crime in Bogotá



Risk terrain models help diagnose why crimes have clustered at certain places and forecast where they are likely to occur

SUMMARY

Latin America is highly urbanized, with above-average crime rates. Its cities are typically unplanned, with high socioeconomic inequality, yet the association between crime and infrastructure has not been clearly defined or quantified. Colombia's capital, Bogotá, collects considerable geo-coded data on urban infrastructure and has reliable geo-coded information on population and crime. The recent development of the world's largest Bus Rapid Transit system has led to the modification of infrastructure in several parts of Bogotá. These changes present an opportunity for studying the association between crime and infrastructure. Drawing on rich data, this project quantified the occurrence of crimes in relation to specific characteristics in the built environment. Through risk terrain modeling (RTM), the team identified locations near public hospitals, schools, drugstores and bus stations as being associated with assault and homicide. The modeling also revealed peak times of day for crime, and predicted areas of the city more likely to experience future crime. Combined with local stakeholder perspectives, RTM analyses can reliably suggest action to reduce crime associated with particular environmental factors. The methods are widely applicable in other locations and for other crimes.

CHALLENGE

Latin America is one of the world's most urbanized regions, with crime levels higher than the global average. Urban violence and crime disproportionately affect young, economically active populations, which are the region's largest segment. Yet the effects of urban characteristics on crime in Latin American cities are little studied. Understanding of this relationship could inform urban planning that helps deter crime.

Among initiatives to address Bogotá's transport problems has been the creation of the world's largest Bus Rapid Transit (BRT) system. The development of the system's trunk lines includes several modifications to the cityscape. This project sought to evaluate the association between crime and infrastructure in Bogotá using the BRT trunk routes as an intervention and comparing them with other areas. Carried out in collaboration with Rutgers University, the study aimed to generate reliable estimates of risk for different crimes around BRT stations and throughout Bogotá. If it could quantify crime variations and their relationship with population flows through the different bus stations, this would help city authorities understand the impact of urban infrastructure modifications in relation to crime. Such information could inform future urban planning to help create environments that are non-conducive or even inhibitive to crime. The study also sought to highlight and estimate risks of crime in 18 different urban features, including schools, libraries, tourist attractions, bridges and clinics. These findings would have implications for cities elsewhere in the world, especially those with similar BRT systems.

Clusters of crime can be partially explained by landscape features that attract criminal behavior at certain times



INNOVATION

To capture Bogotá's overall characteristics, the team drew on rich infrastructure data, including land-use information, service network data (gas, water, sewage) and city block and street audit information captured via previously validated tools. Geo-coded crime data from 2012 then allowed them to estimate the correlation of homicide and assaults with overall city infrastructure features, land-use patterns, socio-economic level, cadastral information, socio-economic survey data and BRT or non-BRT sections of the city. The BRT analyses included data on population flows through stations to understand crime variations related to population density at specific times of the day.

The crime data were analyzed in different ways. The team first conducted a Nearest Neighbor analysis to assess clustering within the distribution of crimes. The results suggested that the distribution of crimes in Bogotá is significantly clustered. Kernel density mapping was then used to identify where, at more localized places within the

study area, the highest concentrations of incidents of crime occur. Using hotspot analysis, the team found many micro-level locations around which crimes cluster. This patterning was statistically significant.

Drawing on the findings from these exercises and combining them with the infrastructural data for Bogotá, the project generated 16 different risk terrain models (RTMs) for assault incidents across the city. This modeling process uses a specific algorithm that identifies relationships between different layers of data and correlates them with crime using count regression models which are then linked to places on a digitized map. The approach represents spatial influences of crime risk factors as common geographic units, then combines separate layers of map (one per risk) to produce maps showing the intensity of all risk factors at every location throughout a landscape – the ‘risk terrain’. Risk terrain maps show where conditions are most conducive to crime. They help diagnose why crimes have clustered at certain places, and can help forecast where they are likely to occur in the future.

The 16 risk terrain models covered:

- ***The entire city for one calendar year***
City-wide risk terrain models suggest that different economic strata (Low, Medium and High) of Bogotá correlate with crime incident locations, but each stratum influences assault and homicide differently. Low stratum blocks have an increased likelihood of both homicide and assault. Proximity to drug stores and medical clinics

presented consistently high risk, with at least 50 percent greater likelihood of either assault or homicide. Places close to hospitals, public schools or BRT stations also presented significantly higher risks of assault or homicide compared to places lacking the defining landscape features of these public facilities (such as large pedestrian access routes).

- ***Different economic strata for one calendar year***
To understand temporal and spatial incentives to criminal behavior within each stratum, the team performed hourly-based RTM analyses. A heat map summarizing the incidence of crimes by hour and day of the week indicated that violent crime clusters at certain times, peaking from Saturday evening until Sunday dawn.
- ***Highest concentration of assault incidents***
Several RTM analyses explore these peak times in greater detail. The results suggest that proximity to drugstores and medical clinics increases the risk of assault and homicide significantly in low stratum sections of Bogotá. In medium stratum sections, proximity to public schools increases the likelihood of assault. Overall, crimes tend to occur mostly in the evenings.
- ***Peak and off-peak hours***
Hourly-based risk terrain models were carried out by stratum, for peak and off-peak hours on the BRT network. These suggest that the risk of assault more than doubles near medical clinics during peak hours in lower stratum sections. In medium stratum sections, the likelihood of assault increases

near private schools. The likelihood of homicide more than quadruples near BRT stations during peak hours in low stratum sections. The results suggest that crimes occur in different places at different times of the day.

- ***Gender-based maps for one calendar year***

Gender-based risk terrain models were generated for each economic stratum to analyze the locations of assaults on female and male victims. The results suggest that proximity to medical clinics and drugstores in low stratum sections almost doubles the likelihood of assault for both male and female victims. In medium stratum sections, being near tourist attractions increases the risk of assault on men, while proximity to private schools increases the likelihood of women being assaulted more than fivefold.

RESULTS

The results indicated that incidents of both homicide and assault cluster at certain places. This can be partially explained by certain features of the landscape that attract criminal behavior at certain times, linked to changing density of people. Such crime patterns occur beyond random chance and are statistically significant.

Findings were presented to key local and national stakeholders in Bogotá, who fed back useful information for contextualizing the results. They suggested the clustering of homicide incidents during peak hours near BRT stations in low stratum sections was related to known criminal activities organized

by certain ‘street vendors’ near a few BRT stations. They also explained that many people go to drugstores to buy prescription drugs unobtainable at hospitals. Offenders near these locations may see these people with cash as potential targets. Drug micro-trafficking may also be a factor, particularly in lower stratum sections.

RTM offers insights into the spatial dynamics of crime and how to mitigate key factors leading to criminal behavior, for example, by environmental modifications that improve passive surveillance and law enforcement activities. The approach can be applied to other crimes, such as residential or commercial robbery and vehicle thefts. The team has encouraged various entities to perform their own RTM analyses. Local stakeholders have already used the methodology to create maps of drug micro-trafficking in five Colombian cities for the Ministry of Justice. Their insider perspective increases the reliability and practical value of the RTM analyses.

LESSONS LEARNED

The project clearly showed how big data analytics reveals correlations which inform strategies to reduce crime.

- ***Draw on big data to inform decision-making***

RTM is useful as a reliable diagnostic tool that can evaluate patterns of crime and orient prevention and enforcement activities more efficiently – including towards specific areas and at specific times. Being cross-sectional, the methodology established certain correlations, rather than causal

associations, but in doing so it highlighted certain features of the environment that otherwise would not easily be detected by simpler statistical modeling.

- ***Use big data analytics to assess disparate data sources***

Big data approaches are useful for integrating disparate but complementary information to provide a very rich environment that can be analyzed to respond to key questions – in this case, about crime. Analyzing big data can also establish emergent patterns of correlation or association which can be related to specific outcomes of interest.

- ***Involve local stakeholders to help explain findings***

Insider perspectives and content analysis with local stakeholders help identify potentially plausible – though not causal – explanations for crime, given their knowledge of the context. Working with local stakeholders is important for generating explanatory hypotheses.



Stakeholders helped interpret the risk terrain maps, suggesting that the clustering of homicides near bus stations was linked to known criminal activity

Revamping Road Condition and Safety Monitoring with Smartphones

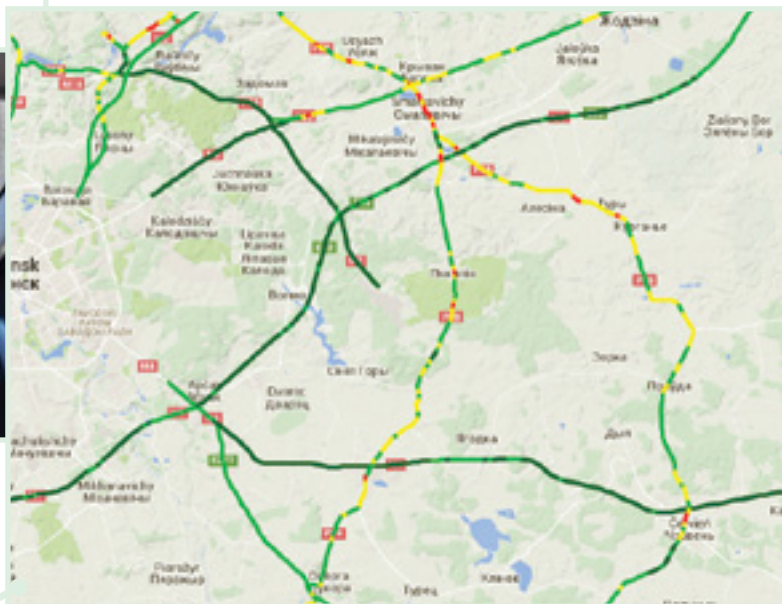
Wei Winnie Wang



Well-kept roads are needed to connect people to public amenities and reduce travel time, vehicle operation costs and crash risks



The RoadLab app uses accelerometers in smartphones to evaluate the roughness of road surfaces and identify major damage



RoadLab visualizes road condition in Google maps

SUMMARY

Road agencies constantly face the challenge of developing cost-effective asset management strategies despite limited resources and modest understanding of road infrastructure conditions from users' perspectives. This project piloted an innovative solution through an app called RoadLab, developed using accelerometers in smartphones carried in moving vehicles. This automatically evaluates the roughness of road surfaces and identifies major damage, such as potholes. It also allows road users to manually submit reports of road accidents or safety hazards, along with precise GPS information. The crowdsourced data is analyzed to extract information on the condition of road surfaces users travel over. Developed in collaboration with the Belarus authorities, RoadLab gives road agencies comprehensive and frequent information on road surface condition over wide areas. This enables them to prioritize investments effectively for maintaining infrastructure, and to assess road surfaces before and after maintenance work. By promoting citizen engagement and enabling road agencies to respond more effectively to users' concerns, this approach also enhances government accountability. Built with the aim of scaling up, RoadLab can easily be modified to other countries and has the potential to become a key input for road network management worldwide.

CHALLENGE

Well-kept roads connect people to public amenities and reduce travel time, vehicle operation costs and crash risks. In order to maintain road networks, government agencies must develop cost-effective asset management strategies, but many have only limited resources and poor understanding of road infrastructure conditions from road users' perspectives. Potholes, rutted surfaces and missing manhole covers are among the hazards identified by road authorities for assessing road surface condition and informing decisions on the maintenance of road assets. However, collecting information on such hazards through conventional methods is costly and time-consuming, involving engineers physically identifying locations that require maintenance or examining surface roughness using road surface profilers. This requires significant resources in both time and labor, especially for large road networks.

*Within 10 days,
the team collected
useful road surface
information for
3,000km of road*

As a result, road agencies often pay inadequate attention to road asset management. To maintain safe, efficient road networks and to improve infrastructure, the authorities need

more affordable, uniform and immediate data on road condition. One approach is to harness road users to report the surface conditions and safety issues they encounter, simply by using an app in their smartphones as they travel by car. Automatic data collection with accelerometers in smartphones can give road agencies large amounts of information from the road users' perspective. This can facilitate quicker and more effective decisions on road asset management. However, existing road-reporting apps faced challenges affecting the reliability of their readings, such as what parameters to include in the evaluation algorithm, and the accuracy of their vertical acceleration detection. The need was for a reliable smartphone app that addressed these common issues and provided valuable reference for future app development.

INNOVATION

The project developed a smartphone app called 'RoadLab,' which in effect harnesses moving vehicles as probes that detect real-time road conditions by using smartphone accelerometers to monitor and report the roughness of travel over stretches of road. The app was developed in close collaboration with the national road management agency in Belarus, alongside the World Bank-supported development of the country's Traffic and Road Safety Coordination Center.

Learning from previous approaches

Through a review of existing road-reporting apps, the team identified key factors often overlooked but affecting the accuracy of road roughness estimations. These included the position of smartphones – especially if changed

during driving – and the vehicle’s speed and suspension type. They realized that the filtering and smoothing of raw data would need to be conducted carefully through machine learning to accurately differentiate bumps from abrupt braking, a swing of the vehicle or the user moving their phone. In addition, repeated reporting of extraordinary vertical acceleration values at the same location by multiple users could be used as a supplemental tool to the data processing and filtering model to identify abnormal road conditions.

To develop the app, the team divided roads into 100-meter segments, with GPS coordinates at the start and end of each. The system gathers and analyzes data from smartphones, and calculates the vertical acceleration and average speed within each segment. Regression models then link the road surface condition with vertical acceleration and speed, and estimate the roughness in line with the global International Roughness Index, commonly used for measuring the roughness of road surfaces. This way, RoadLab estimates can be directly linked with existing road roughness measurements for comparison and updates. Reflecting the practice of Belarus road management agencies, RoadLab automatically categorizes road surface condition as excellent, good, fair or poor. It allows road agencies to set these threshold values themselves, given that what constitutes a poor surface or a major bump is highly subjective, and standards are also likely to vary across road agencies and countries. This allows maximum adaptation to local contexts.

Roadlab users must place the smartphone on a stable surface in a moving vehicle, such as the dashboard or mounted vertically in a cradle

on the windshield. The app will automatically detect phone positions and strength of GPS signals and remind the user to place the device correctly in order to obtain reliable data.

A global solution

Several adaptations were made as a result of field tests to verify road conditions reported by the app. It was found that when the vehicle speed is lower than 30km per hour, smartphone accelerometer readings are less sensitive to road surface, therefore such readings are discarded to avoid false reporting. Similarly, when the smartphone is put in a pocket or directly on the seat of a moving vehicle, the correlation between accelerometer readings and road conditions and bumps is not accurate, so these readings too are excluded. These adaptations also help prevent overloading road agencies with raw data.

Although developed in Belarus, the app was designed as a global solution for road asset management. It was built with parameters easily adapted to suit other countries.

RESULTS

Belarus road agencies have readily adopted the RoadLab app to screen road surface conditions. The approach is much more cost-effective than traditional road-monitoring methods, and the authorities are currently working on integrating it into their own road asset management database and the system of the Traffic and Road Safety Coordination Center.

The app was tested by engineers from the Belarus road management agency, before being launched to the general public in the capital

Minsk through a public campaign, including posters distributed to car-owners' clubs, whose members are keen to improve road conditions. Within 10 days, the team was able to collect useful road surface information for 3,000km of road. Analysis of this data compared with the International Roughness Index showed that the estimation from the smartphone app was reasonable. Despite the limited sample data size, the exercise clearly demonstrated the value of a big data approach to road surface analysis.

Following the pilot, minor refinements are being made to the app, such as inclusion of a chart showing measured road surface for the last 10km surveyed. A future option to make RoadLab more attractive to the general public might be to combine the standalone app with other apps that are more practical for travelers, such as navigation systems.

RoadLab gives road agencies the tools to transform their approach to road asset management. It delivers comprehensive and frequent information on road surface conditions over wide areas, enabling them to prioritize investments effectively for maintaining road infrastructure and to assess road surfaces before and after maintenance work. By promoting citizen engagement and enabling road agencies to respond more effectively to road users' concerns, the app also enhances government accountability.

The team is currently disseminating this innovative approach to other countries and regions for replication. With increasing usage of smartphones, they expect the initiative

to grow into a key input for road network management worldwide.

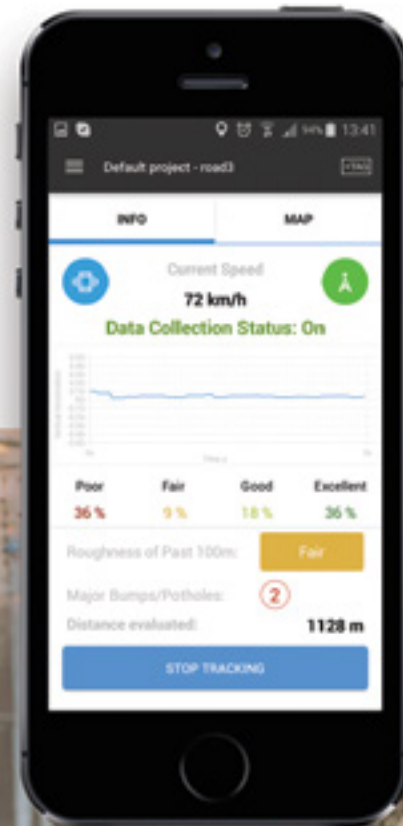
LESSONS LEARNED

The RoadLab project underlined the importance of close client consultation and of thinking in new ways to see links and potential in existing resources.

- ***Consult clients throughout the project cycle***
Successful client engagement was key to the project's success. By engaging and consulting with the Belarus road authorities, the team benefited from valuable end-user inputs, as well as developing the authorities' knowledge and ownership of the approach, making them willing and able to sustain it.
- ***Teach how to fish instead of giving fish***
The project showed the importance of building clients' technical capacity and skills, so innovative approaches can be sustained in the long run and have real impact on communities.
- ***Approach existing situations from new angles***
Researchers have used standalone accelerometers in moving vehicles to evaluate road roughness for decades. By linking wide smartphone ownership, people's use of phones for navigation, and the fact they all have accelerometers embedded anyway, a picture emerges of how phones could be used to extract information useful for road agencies. Through thinking more widely about GPS usage, the team is also now adding a tracing function to the app,

so road networks can automatically be mapped digitally.

RoadLab gives road agencies the tools to transform their approach to road asset management





BIG DATA

Projects to watch

Several of the Innovation Challenge winners and finalists are still awaiting the full results of their big data projects. But they all offer valuable examples of ways in which big data approaches can help improve development programs, and are worth watching for their results in coming months.

Targeting Poverty by Predicting Poverty

Melissa Adelman



Targeting errors are common in development programs, undermining their efficacy. By applying machine learning techniques to datasets commonly used for targeting poverty, this project seeks to improve methodologies for identifying the poor, as well as who will benefit most from particular interventions. With targeting integral to so many programs, improvements could have large impacts on development outcomes.



CHALLENGE

Targeted transfers are central to nearly every anti-poverty program, yet misclassifications are common, reducing program efficacy. Targeting poverty is also more than simply targeting the poor. Optimal targeting requires consideration both of who is poor and who is likely to benefit from an intervention. Someone for whom the expected benefit is large may be selected ahead of someone slightly poorer but for whom the expected benefit is small.

This project aims to improve methodologies for identifying the poor – in particular, by testing whether machine learning techniques can reduce targeting errors. These tools can also be applied to predicting the degree of benefit for different types of recipient, and to estimate the targeting method which maximizes poverty reduction. With targeting integral to so many programs, even small improvements could have large impacts on development outcomes.

INNOVATION

The project views targeting as the sort of simple prediction problem which machine learning tools are designed to address. Existing targeting rules typically predict low consumption via a Proxy Means Test, a model which combines household characteristics, such as roof material or television ownership, to predict

poverty. These models are deeply linear, yet the real world they predict is not. Household characteristics may interact in extremely complex ways: An earth floor may indicate poverty for a piece laborer but not for a farmer, yet a simple linear model would skip over the difference.

Machine learning provides structured methods to search over a very wide set of functions to maximize a model's predictive power. Some methods, such as Lasso techniques, are similar to linear regression but allow for much greater complexity. Others are very different and may better reflect the real world. Decision trees, for example, allow the impact of some characteristics to depend on others – for example, asking whether a person was a piece laborer or a farmer, and then only for the laborer basing the prediction on whether they had an earth floor.

The project applied machine learning tools to several existing programs and one new experiment, to compare the targeting outcomes to those of the actual target method used. This will show whether machine learning tools can improve targeting. From their results, the team aims to create an improved methodology that can be applied broadly across data sets already used for targeting.

Real-time Assessments: How Markets Are Working for the Poor

Alvaro S. Gonzalez



Research suggests that many markets in which the poor transact are volatile, fragmented and suffer weak competition, reducing people's ability to escape poverty. This project analyzes existing micro-level price data to provide policymakers with near-real-time information on how well markets are working, so they can implement policies to improve them for the poor.



CHALLENGE

Improving poor households' participation in markets – as buyers, sellers or producers – is central to combating poverty, but participation depends on well-functioning and efficient markets. The team's prior research in Nigeria found correlations between poverty and low market efficiency. Many markets in which the poor transact are volatile, badly integrated and characterized by weak competition. Based on analysis of price and poverty data, these inefficiencies are likely to result in lower incomes, lower investment, higher risk aversion and suppressed supply responses.

The root causes of these inefficiencies are unclear, as is precisely how they affect poverty outcomes. Micro-analysis of market trends and conditions would help governments understand these inefficiencies, and evaluate the effects on poor people of attempts to mitigate them, such as changes to price supports, trade regimes, exchange rate or interest rate policies and infrastructure. The dualism that exists in many developing economies – a rising (usually urban) core with a poor periphery – may also be due to how markets function. This too has yet to be understood. Governments can act to make markets work better for the poor – but they need the right information to do so.

INNOVATION

To increase understanding of markets in which the poor operate, this project analyzes largely unexploited, micro-level price data, to assess how well markets function. The team assessed monthly price data on hundreds of commodities, collected to estimate inflation indexes across the world. To geographically pinpoint markets serving poorer regions, the team combined spatially disaggregated price data with publicly available satellite lights data. This combined data can be used to track trends and conditions in markets and can alert policymakers to changes that may negatively affect the poor. The price data is updated monthly, allowing the team to monitor how markets function in near-real-time.

This analysis can also monitor market trends and conditions in the poorest regions, and assess the impact of reforms and changes in policies or the market. It will inform governments and development partners on how well the economic environment in which the poor transact is helping them raise their incomes and escape poverty.

From Cellphone Data to Poverty Maps

Marco Hernandez Ore



Accurate poverty maps require expensive, time-consuming surveys, meaning many developing countries only produce them every five or 10 years, often with a long time-lag. This limits governments' ability to make effective anti-poverty decisions. Using anonymized cellphone call records in Guatemala, this project aims to create a tool to produce inexpensive, near-real-time poverty maps and predictions based on phone users' behavioral patterns.



CHALLENGE

Accurate poverty maps are critical for effective anti-poverty initiatives. They help policymakers understand poverty and design better interventions, as national level indicators often hide important regional differences. However, they require expensive, time-consuming surveys, meaning many developing countries only produce poverty maps every five or 10 years, often with a long time-lag. This limits governments' ability to make effective decisions.

The lack of information on poverty is a key policy challenge in Guatemala. The country's latest poverty figures were produced in 2011, meaning policies are designed with limited and outdated data. With low government revenue and more than half of the population living in poverty, it is vital that resources are well-targeted. This project addresses Guatemala's urgent need for affordable, up-to-date poverty data to inform decisions and monitor progress.

INNOVATION

The wide use of cellphones in developing economies offers an opportunity to enhance poverty mapping. Cellphones generate large datasets which can be analyzed using machine

learning techniques to reveal users' behavioral patterns. These can be used to estimate poverty.

The project aimed to create a tool to produce inexpensive, rapidly updated poverty maps and to forecast trends in poverty across geographical regions. Using anonymized cellphone call records gathered by Movistar Guatemala, the country's largest cellphone provider, the team developed algorithms to extract and analyze users' behavioral patterns as a proxy to estimate poverty. The call records allowed them to compute three sets of variables to characterize human behavior: Phone service consumption, social networks and mobility. These behavioral features could predict regional poverty data, easily visualized as maps. For example, longer traveling patterns or smaller social networks could be correlated to higher or lower income levels respectively.

The validity of the approach is being assessed by analyzing the similarity between the predicted social indicators and Guatemala's 2011 poverty map. This approach aims to allow governments to use call records to produce low-cost, near-real-time information on poverty and to forecast trends.

Testing Cellphone-Derived Measures of Income and Inequality

Tariq Khokhar



In many developing countries, official measures of poverty and inequality – vital to responsive policy design and implementation – are produced with a multi-year time lag and inconsistent coverage. This project evaluates techniques that use Call Detail Records (CDRs) to offer more timely and complete estimates of poverty and inequality. It also examines how these techniques can be incorporated into government workflows.



CHALLENGE

Many developing country governments operate in resource-constrained environments with significant gaps and lags in data needed for effective policy design and service delivery. Official measures of poverty and inequality are currently produced with a multi-year time lag and have varying levels of coverage across and within countries. More timely, complete and disaggregated socio-economic measures are urgently needed.

Techniques are emerging for using Call Detail Records (CDRs) to offer more comprehensive and timely estimates of poverty and inequality. This project aims to evaluate these techniques and explore how they could be incorporated into the routine work of government agencies. Such knowledge could form the basis of fundamental improvements in key data availability for developing countries.

INNOVATION

The project takes a two-part approach, first working with the Colombian National Statistics Office to identify opportunities for integrating

CDR-derived techniques for measuring poverty into routine data work, then evaluating the technical performance and methodological issues around using CDRs as proxy measures for wealth. The team will examine how well CDR-derived models perform, including in comparison with official sources, and whether they can be valuable in environments with sparse calibration data. They will also assess the impact of biases such as infrastructure development and mobile phone penetration on CDR-derived measures of wealth and inequality.

The project will produce papers identifying opportunities for CDR-derived measures to be incorporated into the workflow of government statistical programs, and presenting the evaluation of CDR-based techniques for estimating key socioeconomic variables. All software developed will be published as open-source code intended for re-use by others. The project's long-term goal is to provide governments with new tools and methods for estimating socioeconomic variables in their countries.

Satellite-Based Yield Measurement

Talip Kilic



Through trials in Uganda, this project is testing a novel approach to derive 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.



CHALLENGE

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. 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.

INNOVATION

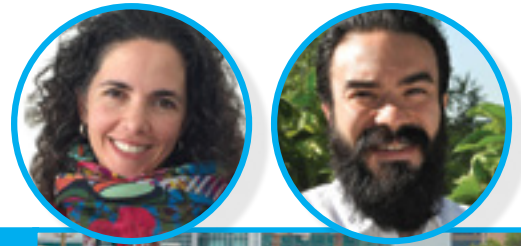
The team used the Scalable Satellite-based Crop Yield Mapper, a statistical approach newly developed at Stanford University which relates satellite data to plot-level ground measures of yields in order to make future yield predictions. The outlines of each maize plot were captured via handheld GPS devices 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). Questionnaires were also submitted to each household. The combined satellite and field datasets provide an unprecedented opportunity for testing the ability of satellites to improve and predict yield measurement in smallholder systems.

The project 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 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.

Understanding Individual Travel Patterns in African Cities

*Nancy Lozano Gracia
Talip Kilic*



This project combined face-to-face and phone interviews with analysis of big data from sensor-embedded smartphones in Dar es Salaam, Tanzania. It aims to capture accurately and affordably the route, purpose, travel mode and cost of individual's journeys within the city, to support well-informed urban transport and land-use planning.



CHALLENGE

To make informed, coordinated decisions on transport investments and land-use planning, policymakers need reliable information about how individuals move around cities and the constraints they face. These decisions affect the locations of households and businesses, influencing residents' quality of life and economic opportunities.

Traditional methods for understanding individuals' travel patterns (purpose, mode and cost) involve collecting travel diaries kept by respondents and supervised for completion by field staff over an extended period. These methods are resource-intensive, subject to recall error and demanding on respondents.

To support well-informed urban planning, this project aimed to develop a tool that could capture accurately and affordably the route, purpose, travel mode and cost for every trip respondents made within a city over a period of time.

INNOVATION

The team sought to create a dataset highly informative about individual's travel patterns as a function of their socioeconomic background, the purpose of their travels and the associated costs. They combined face-to-face interviews with analysis of big data from sensor-embedded smartphones and follow-up phone interviews, to assess individual travel patterns in Dar es Salaam, Tanzania (although the methodology could be applied to other cities).

After developing sensors and software that were installed in GPS-enabled smartphones, the team selected a random sub-sample of respondents to the World Bank's 2013-14 Measuring Living Standards Survey. Each had already taken part in a face-to-face interview covering their socioeconomic background and travel patterns. They were supplied with a smartphone able to collect and transmit the time and GPS location of individual movements at one-minute intervals for a one-month period. To encourage continued participation, respondents were told they could keep the phones after the study. A total of 533 people from 300 households took part. Journey records from their phones were then validated



via follow-up phone interviews every three days, covering the origin, destination, route, purpose and cost of each trip.

Initial data analysis is now underway, focusing on understanding the key determinants for how people choose modes of transport. The team aims to combine project data with cellphone call records to assess the feasibility of using lower-cost phone data for informing transport planning. Subsequent analysis will also examine how the travel patterns recorded compare to those based on traditional data sources, such as the Living Standards survey.





BIG DATA

Key lessons

Key lessons from the challenge winners



Despite the diversity of the winning solutions, several overarching lessons emerge. These can help big data practitioners harness the wealth of information generated by everyday activity, and use it to promote development.

Spot the opportunities

- Much of big data's potential lies in combining disparate and even seemingly unrelated sources of information in innovative ways for analysis. This creates rich new approaches – as in Bogotá, where the team combined existing data sources into a potent dataset from which to predict crime. The climate-smart agriculture project merged climate and yield data, while Open Traffic combined smartphone data and taxi company databases.
- Approach existing situations from new angles. Researchers have used standalone accelerometers in moving vehicles to evaluate road quality for decades, but the RoadLab team linked wide smartphone ownership, the use of phones for navigation and the fact phones all have accelerometers, to extract information useful for road improvement.
- Seek opportunities in emerging technology. Poverty mapping in Sri Lanka and electrification monitoring in India show the untapped potential of satellite imagery. Ultra-light drones were used for high-resolution cadastral mapping in Kosovo, and will help record agricultural data in Latin America.

Quality data means quality results

- High-quality input data is essential for accurate results. Even where data is readily available, there is often significant potential to modernize the methods used to capture, store and share it – for example, through machine-based data sourcing.
- Prepare and store data meticulously (addressing gaps, outliers, correlated variables, etc.). Use cloud-based storage technologies so data is centralized and always available.
- It's rare to have perfectly extractable data, so plan for imperfection. The financial inclusion project grappled to unify different data sources on different servers. Even with the best preparation, extraction often takes longer than anticipated, so include potential delays in project planning.
- Consider data storage from the beginning. Open Traffic's data was initially kept as city-specific files in cloud-storage, but the file sizes soon became unwieldy, so the data was aggregated as a single stream, using a cloud-based service which uploads real-time data streams from multiple sources.
- Respect privacy and protect individuals from potential harm resulting from use of their data (such as retaliation for expression of opinion on social media).

- Focus on creating a positive experience for users of big data tools. Technical issues undermine user trust, making the adoption process harder. As the climate-smart agriculture team found, a tool must offer sufficient services to engage users and be easy to operate, otherwise it will not be used.

Human input still matters

- Combine human and computational power for optimum results. The sentiment analysis in Brazil showed that alongside technological methods, successful analysis of complex socio-political issues using online social data requires human interpretation.
- Teach how to fish instead of giving fish. Projects such as the RoadLab app in Belarus showed the importance of building clients' technical capacity and skills, so innovative approaches can be sustained in the long run and have real impact on communities.

- Enhance other research approaches with big data analytics. Big data makes observations possible from a distance, both in terms of space and time, and offers naturalistic insight into people's true attitudes. However, big data analytics does not always offer a representative sample of views, and is therefore best used to complement other approaches to understanding development issues.
- Enhance big data analytics with other research approaches. Despite the current excitement about big data, it is ultimately 'only' a new (albeit very rich) data source, and does not make other data obsolete. As the financial inclusion research showed, talking to people remains a powerful source of information. In the Philippines, OpenRoads delivers on-line transparency, but institutions that link transparency to accountability are still needed to improve service delivery.



Big data does not make other data obsolete, so combine human and computational power for optimum results

The power of partnerships

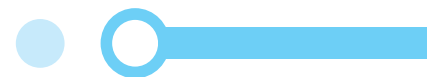
- Partnerships are crucial for success. In Latin America, working with agricultural organizations was key to scaling up data-driven agronomy. The development and use of the sentiment analysis tool in Brazil required close collaboration with experts in fields such as society, history and government.
- Gain credibility with partners, to assuage legitimate concerns about how information they share will be used.
- Create communities of practitioners – networking will be powerful in promoting uptake and refinement of big data methods. In Latin America, CIAT hopes to support a user community for data-mining techniques in agriculture.
- Involve local stakeholders to bring valuable insight from new perspectives. In Bogotá, local stakeholders helped identify potentially plausible explanations for crime. The OpenRoads platform relies on stakeholder input, from government agencies to ordinary citizens.
- Empower local communities through big data tools. Although resources and capabilities are needed to process and manage big data, projects such as the drone mapping in Kosovo demonstrate how new technology can help communities and local government make informed decisions.
- Use big data to promote dialog. Platforms such as India.Nightlights, OpenRoads and RoadLab promote a multi-stakeholder conversation. The Bogotá crime research and Kosovar mapping work also brought together stakeholders to work jointly towards progress.

Back new approaches with reliable evidence

- Rigorous validation or ‘ground-truthing’ is essential to establish confidence in novel data sources and methods. Earlier work establishing the strong correlation between lights in satellite imagery and electricity supply on the ground generated confidence in the India.Nightlights project approach.
- The Sri Lanka poverty mapping illustrates the broad scope for fruitful collaboration between poverty economists and geo-spatial image experts – but hard evidence is needed for the benefits that big data outputs can provide.
- Validation of big data models is more complicated than survey-based techniques. The data involved is not collected for the purpose to which it is being put, so inspection may not be revealing and pre-testing not possible. But, at a minimum, in-sample validation techniques (such as a holdout test set) should be used.
- Persevere. Several projects’ success rests on the importance of persevering when faced with obstacles. The overall India.Nightlights process took five years to reach the stage of publicly launching the web platform, overcoming hurdles such as data processing requirements en route.



Partnerships are crucial for success in pioneering big data projects



Artificial neural networks

In machine learning, artificial neural networks (ANNs) are models inspired by biological neural networks and used to estimate functions that depend on a large number of inputs and which are generally unknown. ANNs comprise complex systems of interconnected 'neurons' which exchange messages and are able to model a system by means of a training algorithm. The connections have numeric weights that can be adjusted, making neural networks adaptive to inputs and capable of learning.

Big data

Big data is an umbrella term used to describe the constantly increasing flows of data emitted from connected individuals and things, as well as a new generation of approaches being used to deliver insight and value from these data flows. Sources include technologies such as the internet, mobile phones, ground sensors and satellites.

Big data analytics

Big data analytics is the emerging set of tools and methods to manage and analyze the explosive growth of digital information. It includes visualization, machine learning techniques and algorithms.

Cluster analysis

Cluster analysis or clustering is used in machine learning to group (or cluster) a set of objects so that they are more similar to each other than to those in other groups. It can be achieved by various algorithms that differ in their notion

of what constitutes a cluster and how best to find them. Popular notions of clusters include groups with small distances between members, dense areas of data space or particular statistical distributions.

Decision tree analysis

A decision tree is a machine learning tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs and utility. It is one way to display an algorithm.

Hotspot mapping

Hotspot mapping visualizes the geographic incidence of socioeconomic data, such as crime. One of the most widely used techniques for generating hotspot maps as smooth continuous surfaces is kernel density estimation (see below). Instead of mapping the location of individual events, hotspot mapping highlights areas with above average incidence of events. These are known as 'hotspots'.

Kernel density estimation

In statistics, kernel density estimation is a method for estimating the probability density function of a random variable. It makes inferences about the statistical population, based on a finite data sample.

Least absolute shrinkage and selection operator (Lasso)

In machine learning, Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance

the prediction accuracy and interpretability of the statistical model it produces. It selects variables by imposing a penalty on the selection of additional variables when fitting the model.

Machine learning techniques

Machine learning involves the construction of algorithms that can learn from and make predictions about data. Rather than following static program instructions, these algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions.

Manual selection prediction models

Manual selection models are those built by researchers (as opposed to derived from machine learning) for statistical prediction.

Natural language processing

Natural language processing is a field of computer science related to human-computer interaction. It links computational linguistics and human (natural) languages, and aims to enable computers to derive meaning from human or natural language input.

Nearest neighbor analysis

Nearest neighbor analysis (also known as proximity search, similarity search or closest point search), is an optimization problem for finding closest (or most similar) points. It attempts to measure distributions according to whether they are clustered, random or regular.

Random forest technique

Random forest is a machine learning tool for classification, regression and other tasks. It constructs a multitude of decision trees to

leverage the power of multiple alternative analyses, randomization strategies and ensemble learning to produce accurate models, variable ranking and detailed data reporting. It can spot outliers and data anomalies, display clusters, predict outcomes, identify predictors, discover data patterns and replace missing values.

Regression analysis

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

- ***Count regression***

Count regression analysis involves modeling using count data, a statistical data type in which the observations can take only non-negative integer values and where these integers arise from counting rather than ranking.

- ***Stepwise regression***

Stepwise regression is the step-by-step construction of a regression model that involves automatic selection of predictive variables. It can involve trying out one independent variable at a time and including it in the regression model if it is statistically significant, or including all potential independent variables in the model and eliminating those that are not statistically significant (or a combination of both methods).

Acknowledgments

With warm appreciation to the **Innovation Labs' Big Data Team**, for their hard work and commitment in organizing the Big Data Innovation Challenge and supporting the winning projects:



Adarsh Desai
Program Manager



Trevor Monroe
Operations Officer



Andrew Whitby
Data Scientist



Bruno Sanchez Nuno
Data Scientist



Luda Bujoreanu
Operations Advisor



Kiwako Sakamoto
Data Analyst

With thanks also to the team's **Collaborators** for their ongoing support:

Amparo Ballivan

Lead Economist, Development Data Group (DECDG)

Isabelle Huynh

Senior Operations Officer, Transport and ICT Global Practice (GTIDR)

Malar Veerappan

Senior Data Scientist, DECDG

Rajan Bhardvaj

Lead IT Officer, Information and Technology Solutions (ITS)



Special thanks to *Publication Coordinator* **Norma Garza**
Knowledge and Learning / Open Contracting and Extractives Governance



“ These case stories demonstrate that big data can improve development effectiveness and help World Bank operations achieve results through better evidence, efficiency, awareness, understanding and forecasting... Ultimately, big data analytics can be an accelerator for ending poverty and boosting shared prosperity ”



WORLD BANK GROUP