Not just another report on data

- A poverty lens on the value of data
- Prioritising poor people and poor countries
Data for development: 3 pathways

- Demand / create transparency
- Data production and collection
- Hold data on individuals
- Create / use data in production process
- Development

Individuals, Civil Society and Academia

Government and International Organisations

The Private Sector
Data for development: 3 pathways

- **Demand / create transparency**
  - **Data production and collection**
  - **Hold data on individuals**
  - **Create / use data in production process**

- **Individuals, Civil Society and Academia**
  - **Reuse / combine / share**

- **Government and International Organisations**
  - **Reuse / combine / share**

- **The Private Sector**
  - **Reuse / combine / share**

- **Development**
  - **Greater accountability**
  - **Criminal activity**
  - **Dark net**
  - **Policy design / evaluation**
  - **Service delivery**
  - **Corruption**
  - **Surveillance**
  - **Business opportunities**
  - **Widening inequality**
  - **Market concentration**
Unlocking potential by reusing data

- Individuals, Civil Society and Academia
- Government and International Organisations
- The Private Sector

Reuse
Combine
Share

Reuse
Combine
Share
# Table of Contents

## Part 1: Advancing development objectives through data

- Chapter 1: Harnessing the Value of Data for the Poor
- **Chapter 2: Data as a Force for Public Good**
- Chapter 3: Data as a Resource for the Private Sector
- **Chapter 4: Creative Reuses of Data for Greater Value**

## Part 2: Aligning data governance with the social contract

- Chapter 5: Data Infrastructure Policy: Ensuring Equitable Access for Poor People and Poor Countries
- **Chapter 6: Data Policies, Laws, and Regulations: Creating a Trust Environment**
- Chapter 7: Creating Value in the Data Economy: The Role of Competition, Trade and Tax Policy
- Chapter 8: Institutions for Data Governance: Building Trust through Collective Action

## Part 3. Moving toward an integrated national data system

- Chapter 9: Creating an Integrated National Data System
Eyes in the Sky, Boots on the Ground

Integrating Surveys and Satellites to Improve Crop Area and Yield Measurement

Talip Kilic
DEC Policy Research Talk
May 25, 2021
Background

- Role of agriculture in rural livelihoods (Byerlee et al. 2007, Davis, et al. 2017)

SDG Target 2.3.
"By 2030, double the agricultural productivity and the incomes of small-scale food producers..."
Background

• Role of agriculture in rural livelihoods (Byerlee et al. 2007, Davis, et al. 2017)
• Need for accurate, crop-specific measures of area under cultivation, production and yields – not only at the national-level but with enhanced within-country disaggregation
• Surge in availability of high-resolution satellite imagery
The idea

Machine Learning for Classification & Yield Estimation

High-Resolution Seasonal Crop Area and Yield Mapping At-Scale

Satellite Data

Ground Truth Training Data
Research on satellite-based methods for agricultural monitoring in smallholder systems
Important takeaways from existing research

• Training data does have a bearing on the quality and spatial resolution of satellite-based estimates

Uganda Plot-Level Ground (Crop Cutting, Survey-Based) versus Satellite-Based Maize Yields
Key takeaways from existing research

• Training data does have a bearing on the quality and spatial resolution of satellite-based estimates

• Existing research has largely been at sub-national levels, with heterogeneity in the type of and approach to training data collection as part of surveys

• Large-scale surveys can address training data needs of earth observation applications on crop area mapping and crop yield estimation in lower-income countries, but...
  - There are no clear recommendations on survey methods and fieldwork protocols that should be adopted to generate the right training data
On-going research to address knowledge gaps

- **50x2030: Data-Smart Agriculture Initiative** is supporting research for the development of guidelines for large-scale survey data collection in low- and lower-middle income countries as to meet training data needs of high-resolution satellite-based estimation of **crop areas and crop yields**
  - Anchored in peer-reviewed research conducted over the period 2019-2022
  - Focus countries: Cambodia, Ethiopia, Malawi, Mali, and Uganda
  - Focus crops: Maize, wheat, sorghum, millet, barley, and rice
  - Important for assessing the utility of existing georeferenced survey data for earth observation research and for informing the design of future household and farm surveys
On-going research to address knowledge gaps

- Recent work by Azzari et al. 2021 aimed to answer several operational research questions in the context of satellite-based maize area mapping across Malawi and Ethiopia
  - How much training data do we need to reach an acceptable level of accuracy of a crop classification algorithm?
  - How does the approach to georeferencing plot locations in surveys impact algorithmic accuracy?
  - How do the type of satellite data and exclusion of plots under specific area thresholds affect algorithmic accuracy?
Survey data

- We use georeferenced plot-level survey data nationally-representative, multi-topic surveys that were implemented by the Malawi NSO and the CSA of Ethiopia under the World Bank LSMS-ISA Initiative
  - **Malawi Integrated Household Panel Survey (IHPS) 2019**
    - Longitudinal sample, dating back to 2010
    - Reference season: 2018/19
    - Plot-level georeferenced information: Single plot corner point + plot boundaries
  - **Malawi Fifth Integrated Household Survey (IHS5) 2019/20**
    - Cross-sectional sample
    - Reference season: 2017/18 or 2018/19
    - Plot-level georeferenced information: Single plot corner point + plot boundaries
  - **Ethiopia Socioeconomic Survey (ESS) 2018/19**
    - Baseline for a new longitudinal sample
    - Reference season: 2018 meher season
    - Plot-level georeferenced information: Single plot corner point
Data collection scenarios

Geolocation

Plot boundary
- 6 approaches to sample full boundaries

Corner point
- Pixel value at one corner point (used as

Centroid
- Mean of all pixels in boundary

Plot mean
- Mean of all pixels in simplified boundary constructed from corner points

Plot points
- Values at a random 20% of pixels in boundary

Boundary points
- Values at 4-8 corner points

Convex hull
- Mean of all pixels in simplified boundary constructed from corner points

Field Size

Data Pool Size

Map data ©2020 Google
Data collection scenarios and modeling

In Malawi, we tested a total of 26,250 scenarios:
- 7 geolocation methods - boundary points, centroid, convex hull, corner, hull mean, plot points, and plot mean.
- 50 data subsets - 2% to 100% subsets of training data, at an increment of 2% points.
- 5 area thresholds - 0, 0.05, 0.1, 0.15, and 0.2 ha.
- 3 feature types - optical only, radar only, both optical and radar.
- 5 replications to capture variability due to random sampling

In Ethiopia, we tested a total of 250 scenarios – based on findings from Malawi and availability of only corner points:
- 1 geolocation method – corner point
- 50 data subsets - 2% to 100% subsets of training data, at an increment of 2% points.
- No area threshold, with optical data only
- 5 replications to capture variability due to random sampling
Headline findings

- **< 1,000 plots**: multi-point approaches perform better

- **2,000+ plots**: aggregation approaches - plot mean (based on plot boundary) & hull mean (based on all corner points) - outperform all

- Need -7,000 plots with a single corner point to reach performance with -3,000 plots under aggregation approaches

- Aggregation approaches had the fastest learning

- Peak performance can be achieved with ~ 60% of training data (~4,000 plots) under plot mean (preferred) and hull mean (second best).

- Centroid method outperforms the single corner point method. If only a single georeferenced point can be collected, it should be near the center of the plot (third-best).
Small differences, large consequences

- Small differences in model performance may lead to large differences in estimated areas.
- Aggregated methods yielded most conservative estimates.
- Multi-points methods tended to “overclassify” the most.
- There is value in achieving small performance gains with better training data.
Key takeaways

- Collecting a **complete plot boundary** is preferable to competing approaches to georeferencing plot locations in large-scale household surveys.

- Seemingly-small erosion in maize classification accuracy under less preferable approaches to georeferencing plot locations **consistently results in total area under maize cultivation to be overestimated** - in the range of 0.16 to 0.47 million hectares (8 to 24 percent).

- Collecting **GPS coordinates of the complete set of plot corners** is a second-best strategy, can approximate full plot boundaries and can in turn train models with comparable performance.

- **Classification performance peaks with ~60% of the training data** under preferred and second-best approaches to georeferencing plot locations.

- If only a single GPS point can be collected, it should be near the plot center rather than at the plot corner. With large datasets, performance could be comparable to that of complete plot boundaries.

- No plot observations should be excluded from model training based on a minimum plot area threshold.

- Optical features alone can provide sufficient signal to maximize prediction quality.
Looking forward

- Continuing 50x2030-supported research to formulate the guidelines for large-scale surveys to fulfill training data needs of satellite-based crop area mapping and yield estimation in smallholder farming systems
  - Leverage additional existing large-scale survey data from Mali, Malawi and Uganda – with georeferenced plot outlines and objectives measures of yields based on crop cutting – to:
    - Expand crop classification to new countries & new cereals: barley, millet, rice, sorghum & wheat
    - Identify training data requirements for high-resolution yield estimation for maize & new cereals
  - Gauge sensitivity of recommendations to the choice of model (e.g., RF vs. CNN) & geospatial covariates
  - Document accuracy of out-of-season predictions & inter-temporal decay in model accuracy
  - Conduct research on object-based classification and automated detection of plot boundaries
  - Depending on the COVID-19 pandemic, conduct additional survey experiments in 2022 in non-African settings
Open access data assets

10-m resolution crop area and maize area maps for Malawi and Ethiopia for each agricultural season from 2016 to 2019 on World Bank Development Data Hub


The social contract for data
The three elements of a social contract for data: Value, Equity and Trust
Working towards an integrated national data system (INDS)
Participants in the INDS create and share data

Private sector → Government entities → Civil society and individuals

International and regional organisations → Institutions → Academia
Five foundations sustain the INDS
Trust

1. Trust that the information created from data is credible.
   => test new methods and data (data quality)
   => monitor data systems (e.g. transparency, access, SPI)

2. Pillars (Infrastructure, Policies, Institutions, Governance)
   => participants need to trust that they won’t be harmed from sharing data and that they will share in some of the benefits

3. INDS
   => steps to create trust
Within-subject design: Two measures of plot size. Malawi, Uganda, Tanzania, Niger. (Self-report – GPS) / (GPS)

<table>
<thead>
<tr>
<th>ha</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.5</td>
<td>103%</td>
</tr>
<tr>
<td>0.5-1</td>
<td>23%</td>
</tr>
<tr>
<td>1-2</td>
<td>6%</td>
</tr>
<tr>
<td>2-5</td>
<td>-1%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>-33%</td>
</tr>
<tr>
<td>Tot</td>
<td>-0.04%</td>
</tr>
</tbody>
</table>

Uganda: GPS (density) & Self reports (histogram)

Data experiment example, Crop Yields

- Ethiopia Socioeconomic Survey
- Within-subject design: Two independent measures of crop yields for the same plot (n=5,000).
- External validity: Nationally representative, two points in time, 23 crops.
- Plot size measured with GPS (denominator)
- 5248 plots with both crop cuts and self-reported production

Data and analysis

Download and explore the data behind the WDR 2021.

- **Statistical Performance Indicators (SPI)**
  The indicators measure statistical performance across 174 countries. They are grouped into five dimensions, which capture whether foundations such as financing, skills, and governance needed for a strong statistical system are in place. Under each dimension is a set of indicators to measure performance providing a time series extending at least from 2016 to 2019 in all cases, with some indicators going back to 2004.

- **Global Data Regulation Survey**
  The Global Data Regulation diagnostic is a comprehensive assessment of laws and regulations on data governance. It covers both the enabler and safeguard regulatory practices in 80 countries. The questionnaire comprises 37 questions designed to determine if a country has adopted good regulatory practice on data governance.

- **Data Governance Indicators**
  The DGI dataset is a subset of the Digital Government/GovTech Systems and Services (DGSS) dataset. The indicators are categorized based on WDR2021 report’s four pillars of data governance framework (infrastructure, laws & regulations, policies, and institutions). They aim to inform a maturity model (MM) approach to data governance and contain both qualitative and quantitative metrics to monitor the state of the data governance practices within the public sector.

---


This data catalog collection provides access to all available source data and R/STATA replication files (for specific figures) necessary to create the figures and maps that appear in the report.

---

**Background research**

The World Bank working papers on WDR 2021 topics.

- **Measuring the Statistical Performance of Countries: An Overview of Updates to the World Bank Statistical Capacity Index**
  Hai-Arih Dong, Mustafa Dinc, Jaderica Díaz, Hikko Maeda, John Puhliger, Limar Serapuddin, Brian Stacy, and Denisse Woldu.
  National statistical systems are facing significant challenges. These challenges arise from increasing demands for high quality and trustworthy data to guide decision making, coupled with the rapidly changing landscape of the data revolution. To emphasize the urgent need for transformation and to help create a mechanism for learning amongst national statistical systems, the World Bank has developed new Statistical Performance Indicators (SPI) to monitor the statistical performance of countries.

- **Policy Choices Can Help Keep 4G and 5G Universal Broadband Affordable**
  Edward J. Oughton, Niccolo Comini, Vivien Foster, and Jen W. Hall.
  The United Nations Broadband Commission has committed the international community to accelerate universal broadband, but the costs of meeting these objectives in the context of rapid technological change are not well understood. This paper compares the global cost-effectiveness of different infrastructure strategies for the developing world to achieve universal 4G or 5G mobile broadband.
Change in global data landscape requires new measurement and benchmarking tools
Statistical Performance Indicators

Forward looking

Measures less advanced systems, and more advanced systems

Open Data + Open Code

Gives countries incentives to build modern data and statistical systems
## 5 pillars, 22 dimensions and 51 indicators

<table>
<thead>
<tr>
<th>Pillars</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Use (User Types)</strong></td>
<td>Legislature, Executive, Civil Society, Academia, International Bodies</td>
</tr>
<tr>
<td><strong>Data Services (Service Types)</strong></td>
<td>Quality of Data Releases, Richness and Openness of Online Access, Effectiveness of Advisory and Analytical Services Related to Statistics, Availability and Use of Data Access Services</td>
</tr>
<tr>
<td><strong>Data Products (Topics)</strong></td>
<td>Social, Economic, Environmental, Institutional</td>
</tr>
<tr>
<td><strong>Data Sources</strong></td>
<td>Statistical Office (Censuses and Surveys), Administrative Data, Geospatial Data, Private Sector Data/Citizen Generated Data</td>
</tr>
<tr>
<td><strong>Data Infrastructure</strong></td>
<td>Legislation and Governance, Standards and Methods, Skills, Partnership, Finance (Domestically and from Donors)</td>
</tr>
</tbody>
</table>
Uses Indicators from various sources

Publicly available data + Open code

Partners
Index Methodology

- SPI score = sum of statistical performance indicators
- Index can be broken into components to see where a country leads and lags
- Nested structure allows for symmetry, monotonicity, and subgroup decomposability
Regional scores

Source: World Bank Statistical Performance Indicators
Country performance varies tremendously within regions

Source: World Bank Statistical Performance Indicators
Low- and Middle-Income Countries have some way to go

- High income: 79
- Upper middle income: 62
- Lower middle income: 57
- Low income: 47

Source: World Bank Statistical Performance Indicators
Over/Underperformers
Higher NSO independence score correlates with higher statistical performance

Source: WDR 2021 team calculations based on Ibrahim Index of African Governance and the World Bank's Statistical Performance Indicators.
Note: The values in the chart are jittered to make them more readable.
Greater press freedom correlates with higher statistical performance

Source: WDR 2021 team calculations based on World Press Freedom Index and the World Bank's Statistical Performance Indicators.

Download the report and explore Data Stories at

Use #WDR2021 to follow us online for communication events