The Case for Open Data in Agricultural Science

Photo Credit: A. Etizinger / CGIAR. Tanzania farm landscape [1]

A capstone project submitted in partial satisfaction of the requirements for the degree of Master of Science in

International Agricultural Development

University of California, Davis

Winter Quarter, 2019

Erin Havens
M.S. Candidate, International Agricultural Development
M.S., Agricultural & Resource Economics
University of California, Davis

[1] http://www.focusonland.com/countries/biofuels-and-land-use-in-tanzania/

Table of Contents

Acknowledgements

I would like to thank my committee chair Dr. Robert Hijmans for his help and support during this project, as well as the rest of my committee. I would also like to thank the project team: Aniruddha Ghosh, Alex Mandel and Hongfei Wang.

Acronyms

CGIAR: Consultative Group for International Agricultural Research

GARDIAN: Global Agricultural Research Data Innovation & Acceleration Network

GFC: Geospatial and Farming Systems Research Consortium

ICRISAT: International Crops Research for the Semi-Arid

CIAT: International Center for Tropical Agriculture

CIP: International Potato Center

CIMMYT: International Maize and Wheat Improvement Center

IITA: International Institute of Tropical Agriculture

Executive Summary

Researchers and practitioners in international agricultural development depend on quality data to advance the resiliency and productivity of smallholder farmers, who make up the majority of the world's poor (Food and Agriculture Organization of the United Nations, 2012). As data collection processes become more advanced, there is a need for more data sharing and better resources for data analysis across research institutions, universities, and other organizations. In an effort to address some of these needs, I have contributed to the development of an online platform to facilitate the use of open source agricultural data, including case studies that demonstrate data management techniques in R. The platform will use publicly available data from the Global Agricultural Research Data Innovation & Acceleration Network (GARDIAN) website, an open data platform developed by the CGIAR consortium. This project is not only important for agricultural studies at major universities and institutions such as CGIAR centers, but also plays a significant role in advancing international agricultural development through capacity building in resource scarce institutions, and by helping to advance research questions in topics around agricultural production and food security.

This capstone report first reviews the literature on data sharing and open data in the field of agriculture. The next section provides a background and context of the project and methodology used, followed by a summary the datasets analyzed and case studies produced, including the methodologies used in each case. The final section is a discussion of challenges and opportunities that arose from this process as it relates to data sharing and analysis in international agriculture, and recommendations moving forward.

Introduction

Researchers and practitioners in the field of international agricultural development depend on quality data to advance the resiliency and productivity of smallholder farmers, who make up the majority of the world's poor (Food and Agriculture Organization of the United Nations, 2012). This is especially important as global challenges such as climate change and food insecurity continue to threaten agricultural productivity and rural livelihoods (Altieri & Koohafkan, 2008). Data in agricultural sciences comes in many forms, ranging from surveys to agronomic samples to satellite photography. As data collection processes become more advanced, there is a need for policies and platforms to facilitate data sharing, as well as better resources for data analysis across research institutions and university campuses.

In an effort to address some of these needs, I have contributed to the development of an online platform to facilitate the use of open source agricultural data, which will include several case studies that demonstrate techniques for data management in R. The platform will use publicly available data from the Global Agricultural Research Data Innovation & Acceleration Network (GARDIAN), a data catalogue and website developed by the CGIAR consortium. The platform addresses key issues that come up when using open source data, and is targeted to meet the needs of agricultural scientists in specific. This project is not only important for agricultural studies at major universities and institutions such as CGIAR centers, but also plays a significant role in advancing international agricultural development—through capacity building in resource scarce institutions, and by helping to advance research questions in topics around agricultural production and food security. As part of this project, I produced several case studies with a geographic focus on data available in Tanzania. These case studies demonstrate how to clean and analyze the data, and will contribute to the online platform.

This capstone report first reviews the literature on data sharing and open data in the field of agriculture. The next section provides a background and context of the project and methodology used, followed by a summary the datasets analyzed and case studies produced. The final section is a discussion of challenges and opportunities that arose from this process as it relates to data sharing and data analysis in international agriculture, and recommendations moving forward.

Literature Review

Introduction

The importance of data sharing in academic research appears to be acknowledged by many, but it is often not practiced in research settings (Andreoli-Versbach & Mueller-Langer, 2013; Fecher, Friesike, & Hebing, 2015; Pasquetto, Sands, Darch, & Borgman, 2016). This discrepancy is pointed out in much of the literature on data sharing and open data. This section will review the literature around what data sharing is, how it can be done, obstacles to data sharing, and how these can be overcome.

Definitions

A few terms appear repeatedly in the literature on data sharing in research and academia and merit formal discussions of their definitions in practice. These terms include *data*, *data sharing*, and *open data*, and *cyberinfrastructure*. The main challenges in defining open data or data sharing are concerns about who the data is open to, and how the data is accessed.

Data itself is difficult to define because of the range of formats and collection methods. What is considered data also depends on the field of study and can vary widely even within specific disciplines. For this reason, it is useful to take a broad definition of data in the context of

data sharing. The National Academies of Science defines data as “facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors” (National Research Council, 1999). The major categories of data, according to the National Science Foundation, include observational, computational, experimental, and records (Borgman, 2012).

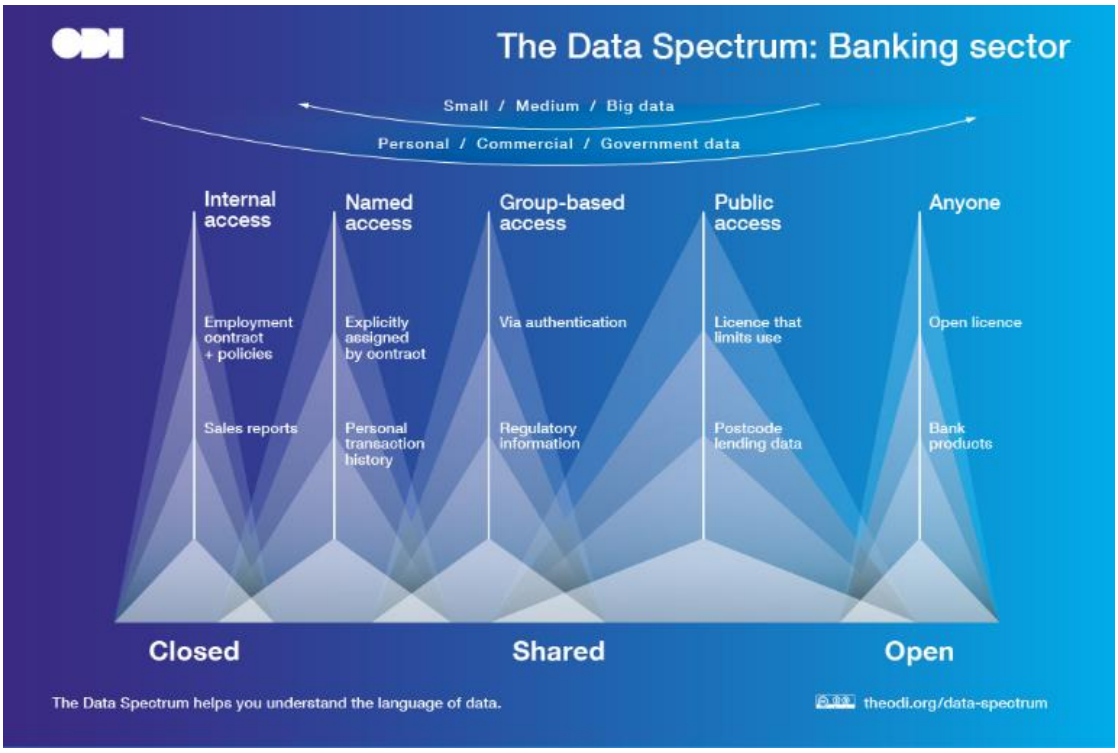


Figure 1: The Open Data Spectrum, from Closed to Shared to Open, with applications to the banking sector. In this case, closed data is only available internally, whereas open data is publicly available to anybody. Source: (https://theodi.org/about-the-odi/the-dat 1)

It is useful to distinguish closed, shared and open data (Figure 1). Open data can be defined as data that is freely accessible to anyone to validate research findings and to re-use for any purpose; but the usage may be restricted (e.g. for non-commercial purposes only) according to the license under which they have been released (Abbà et al., 2015; Betbeder, Damy, & Herrmann, 2017; Michener, 2015; Borgman, 2012). In contrast, “shared data” is distributed

privately, upon request. Shared data may be shared with specific groups of people for a specific purpose—and its use can therefore be much more restricted (ODI, 2018).

Data can also be discovered and disseminated in many ways, such as through private exchange, digital repositories, or as supplementary material in journal articles (Pasquetto et al., 2016). According to survey results from the Belmont Forum's open data survey, the most common discovery route for data is as a reference in the journal article itself, followed by online search engines and data repositories (Schmidt, Gemeinholzer, & Treloar, 2016). Cyberinfrastructure platforms, such as data repositories and data catalogues, have helped to increase discoverability of data and are gaining traction across multiple disciplines. This includes both discipline and institution specific data repositories, as well globally-scoped repositories with fewer restrictions on the data type (Wilkinson, 2016). A data repository is a collection of datasets, including metadata, that allows users to access the data (Olson & McCord, 2000). For example, The Dataverse Platform provides allows researchers to share their data, while giving the researchers control of its continued management (Crosas, 2011). Other platforms have been developed with varying capabilities and levels of success, such as Dyrad and figshare. There are also a number of sites that catalogue data (such as GARDIAN), which means that they simply link data searches to existing data repositories.

Reasons for data sharing:

The benefits of data sharing are echoed across multiple papers, the most common of which are to reproduce or verify existing research and publications, and to ask new questions from existing data (Borgman, 2012; Fecher, Friesike, & Hebing, 2015; Michener, 2015). Many also argue that sharing data increases efficiency by saving time and money that would be required for further or repetitive data collection (Michener, 2015; Pasquetto et al., 2016; Wallis,

Rolando, & Borgman, 2013). Michener et al. further discusses the hidden costs of *not* sharing, which include "contributing to higher research costs and lost opportunity costs; adding barriers to innovation; reducing the effectiveness of cooperation, education and training; suboptimal data quality; and widening the gap between developed and developing countries" (Michener, 2015).

Data sharing can also be beneficial to those who collect or produce the data. For example, researchers and academics may benefit from improved reputations because sharing data is associated with more citations on scholarly articles and more acknowledgement of the work (Betbeder et al., 2017; Fecher, Friesike, Hebing, Linek, & Sauermann, 2015; Michener, 2015). Increasingly, sharing of data for research papers is becoming a requirement of funding sources such as NSF, as well as other organizations, publishers, and professional societies—which would provide a strong monetary incentive to share data as well (Borgman, 2012; Michener, 2015).

Many papers reflect the sentiment that making data available will be a benefit to the public good, accelerating scientific knowledge and new discoveries. By increasing access to data across fields, some of the most challenging questions in research can be answered more easily. For example, See et al. discuss the relationship between global cropland data and food security, noting that sharing of existing data on croplands that are currently unavailable would help accelerate this link (See et al., 2015). Additionally, sharing of data may provide benefits beyond just the scientific community, such as educational tools and public engagement, as well as capacity building in developing countries (Pasquetto et al., 2016).

Barriers to Data Sharing:

While there are many benefits to sharing data, the drawbacks and challenges cited in the literature are just as many. These include technological, social, organizational, financial, and

other barriers (McLure, Level, Cranston, Oehlerts, & Culbertson, 2014). Some relate directly to the researchers own context or perceptions, while others are more institutional in nature. It is important to understand these barriers so that appropriate solutions and necessary resources can begin to support better data sharing and management practices.

While on the one hand open data may increase transparency and allow for verification or reproduction of research, many researchers also fear opening their data to the public for these same reasons. Papers cite worries about misuse of data, scrutiny, discovery of errors or challenges against the conclusions of papers (Abbà et al., 2015; Betbeder et al., 2017; Michener, 2015). Additionally, because of the competitive nature of academia and the funding and resources required in many data collection processes, researchers are hesitant to share data due to fear that competitors will take advantage of the data, fail to properly cite it, and lose intellectual property (Abbà et al., 2015). Where researchers do not have these concerns, there may be a lack of incentive to voluntarily share data if it is not required (Wallis et al., 2013). Finally, many researchers note that they would be willing to share data, but only under certain conditions—such as knowledge about who is using the data and why, a certain amount of time between data collection and sharing, and support from employers or institutions (Fernandez, Eaker, Swauger, & Steiner Davis, 2016; Linek, Fecher, Friesike, & Hebing, 2017).

If the researcher is willing or required to share, many report difficulty in deciding which data should be shared, and in what format (Abbà et al., 2015; Wallis et al., 2013). The extra steps required of sharing data—such as curating the data to an appropriate format, developing the metadata, and documenting the data—often create financial constraints. Additionally, many report a lack of expertise in data management—as well as a lack of known forums, tools, standards, support, and resources on which to share the data. There is a noted lack of variety and

integration of cyberinfrastructure with interoperability and flexibility in data management (Abbà et al., 2015; Betbeder et al., 2017). If platforms do exist, standards and policies may be unfamiliar, or may be lacking and lead to further challenges. Wilkinson notes that because so many repositories have been created, the network of data sharing platforms has become decentralized and creates further challenges for discoverability and reusability by both humans and computational platforms (Wilkinson, 2016). Legal reasons are another common barrier to sharing data, including concerns about licensing, confidentiality, and intellectual property regimes (Michener, 2015).

Best Practices:

In light of the previous challenges and barriers to data sharing, many researchers have stressed the importance of increasing policies, infrastructure, and resources for researchers to be able to share and manage their data (Betbeder et al., 2017). This has included several attempts at defining formal standards, including the five-star deployment scheme and the FAIR Guiding Principles. These standards are important, because it is often not enough to just make the data open source; if the data is not properly formatted according to the following standards, then the data cannot be reused and replication of research studies is infeasible. For example, if data is shared as a pdf, or in a proprietary format, the data is not truly accessible to many users. If the data does not contain a codebook or a dictionary, its use will further be limited. The following protocols have attempted to address some of these issues.

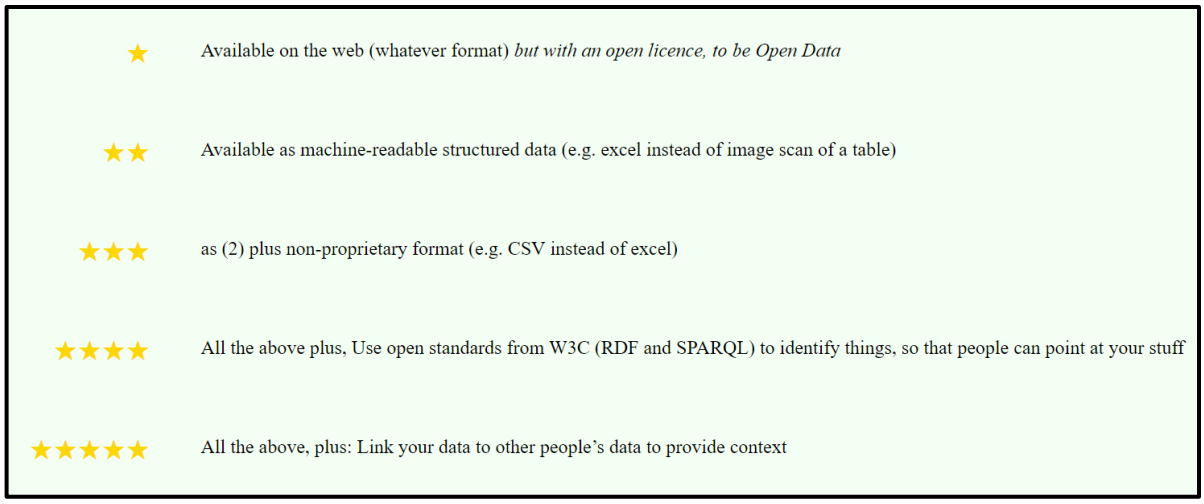


Figure 2: Tim Berners-Lee's five Stars of Open Data, established in 2006. The necessary requirement is that the data has an open license. Data receives five stars if the data is machine readable, non-proprietary, uses open standards, and can be linked to other data sources. Source: https://www.w3.org/DesignIssues/LinkedData.html

The five-star open data deployment scheme (Figure 2) was established by Tim Berners-Lee in 2006 to promote open and linked data. The minimum qualification for open data according to this standard is that the data has an open license. Data receives five stars if the data is machine readable, non-proprietary, uses open standards, and can be linked to other data sources. In 2014, the FAIR Guiding Principles (findability, accessibility, interoperability, and reusability) were established by a number of key stakeholders from academia, industry, and funding agencies (Wilkinson, 2016). The FAIR principles were created with the goal of supporting both individuals' and machines' ability to find and reuse data. The FAIR principles address several of the issues that have come up with regards to data sharing and reusability. Major components are that metadata should be widely available and documented according to standards, that the data should be easily automated for computer use, and that terms and conditions, such as licenses, are clearly laid out. In principle, these principals would facilitate the discoverability and reuse of data, though they cannot address every challenge in data sharing (Wilkinson, 2016).

Many authors have built upon the importance of establishing clear ownership and licensing policies in order to increase data sharing and reuse (Leone, 2017). Common policies include Creative Commons (CC), Open Data Commons (ODC), and licenses by individual government agencies (GODAN, 2017). With these policies in place, researchers and institutions can better ensure that they receive credit for their data, and thus may be more willing to share the data.

Funders, journals, and professional societies can also play a large role in changing trends, as researchers are required to comply with their standards in order to receive funding or to be able to publish (Michener, 2015). In recent years, these institutions have already begun to implement policies that require data sharing, or data management plans. For example, the National Science Foundation has required data management plans for grant applications, which appears to have been a tipping point in data sharing practices in research and academia due to the extensive reach of this funding institution (Borgman, 2012).

Application to agricultural science

In agriculture sciences, a common theme cited in the literature on data sharing is the diversity of data types and sources. Agricultural data is collected at multiple spatial and temporal scales, and from a wide range of sources including secondary research, weather stations, surveys, and agronomic records (Diekmann, 2012; Williams, 2013). Additionally, different fields within agriculture (for example, field vs genetic research) may use a range of data collection methods. This poses an additional challenge to developing standards for data management that can accommodate this diversity.

Scholars also discuss the continued need for data management standards in agriculture, where current practices are not meeting the generally accepted best practices mentioned previously (Fernandez et al., 2016). Where standards do exist, they are still in very early stages (Diekmann, 2012). Many of the same recommendations and best practices discussed above apply to the field of agricultural sciences. For example, in assessing nutrient management research in agriculture, Eagle et al. recommend common protocols for measurements in research collection, consistently defined controls and treatments, better recording and reporting of data, publishing of data as supplemental material or in data repositories, and requirements from funders and journal articles (Eagle et al., 2017).

These articles suggest a need for continued effort in agricultural sciences to increase resources in research and academia for data sharing and analysis. The CGIAR Platform for Big Data in Agriculture, including GARDIAN, is an important step to achieving these goals. However, beyond the platform for data sharing, another important need is further understanding of data management and analysis—especially of existing datasets by second-hand users.

Project Background and Methodology

A common challenge that was cited in the literature was a lack of data management policies, platforms, and resources in agricultural sciences (Carlson & Stowell-bracke, 2013). As standards for sharing data become more common and data becomes more publicly available, there is a need for better understanding of how to use this data for research purposes. A case study at Purdue found that researchers in agricultural sciences primarily use Microsoft Excel for data analysis, followed by R and SAS (Pouchard & Bracke, 2016). For this reason, providing resources for use of these data analysis software would be beneficial.

In an effort to address some of these needs, I have contributed to the development of an online platform to facilitate the use of open source agricultural data, which will include several case studies that demonstrate techniques for data management in R using publicly available data from the GARDIAN website. The platform addresses key issues that come up when using open source data, and is targeted to meet the needs of agricultural scientists in specific. The target audience is initially for researchers and academia –but can expand to institutions globally and private individuals with an interest.

This project is not only important for agricultural studies at major universities and institutions such as CGIAR centers, but also plays a significant role in advancing international agricultural development. Increasing access to data and its subsequent use through data management resources on public online platforms can enhance capacity building (Pasquetto et al., 2016). In resource-scare research institutions, open source data may provide an important research tool that would not otherwise be available. With the addition of resources for using R, an open source platform, advances could be made in institutions that would not ordinarily have the resources to do so.

In agricultural development, many of the advances come from extension workers who are tied to research from universities and institutions. Without proper access to data many of the research questions with direct links to farmer wellbeing and broader global issues, such as climate change and food security, are impaired. For example, in a study by See et al. a lack of access to existing data on global cropland was a significant impediment to moving forward with research that had implications for food security (See et al., 2015). Additionally, specific agricultural questions in developing countries, such as the impact of conservation agriculture on

yield, require survey and agronomic data to be made available for more rapid understanding and implementation. (Brouder & Gomez-Macpherson, 2014).

This project is working closely with the Global Agricultural Research Data Innovation & Acceleration Network (GARDIAN), a CGIAR platform. The purpose of GARDIAN is to "[enable] discovery of agricultural data and publications across the CGIAR system and beyond", and as of November 2018 contains 93,841 publications and 2,376 datasets. (http://gardian.bigdata.cgiar.org). This capstone project assessed the current status of GARDIAN in providing public access to data by focusing specifically on Tanzania. To do this, I searched the GARDIAN website for all available data in Tanzania, and downloaded this data where possible. Then, for several of the publicly available datasets I produced case studies that demonstrate how to clean and analyze the data, which contribute to the online platform. The final case study demonstrates how multiple datasets from different sources can be combined and analyzed together, which strengthens the power of the research.

Summary of datasets and case studies

Overall, a total of ten case studies were produced from data collected in Tanzania, from seven different projects affiliated with the CGIAR consortium. The case studies are produced as R markdown files, with the code and output both included (see appendix 2 for an example). The purpose of the case studies is to demonstrate data management techniques in R, with a specific focus on using publicly available datasets from the GARDIAN website. The case studies include data cleaning and organization, summary statistics and analysis, and replication of published articles. This section will provide a brief overview of each of the projects and datasets analyzed, as well as the lessons and key methods of analysis for each case study. The following section

will discuss some of the trends in challenges that were encountered with respect to data cleaning and organization, re-use, and replication.

Project	Institution	Topic/Purpose	Location	Type of data:	Supplementary Materials
Africa Research in Sustainable Intensification for the Next Generation (Africa RISING): 2014	IITA	Sustainable intensification of maize-legume-livestock integrated systems	Tanzania, Malawi, and Zambia; Three districts in Tanzania: Babati, Kiteto and Kongwa between	Community level data from 810 households and 25 villages	The dataset includes codebook, survey instrument, field report, site selection report
Adoption Pathways Project: 2010, 2013, and 2015	CIMMYT	Long-term adoption and adaptation of sustainable intensification practices as part of a broader theme of sustainable intensification research	Eastern and Southern Africa: Ethiopia, Kenya, Tanzania, Mozambique, and Malawi	Panel dataset which was collected across the five countries and three years.	Many of the datasets includes a questionnaire for many, and final report.
TAMASA Tanzania Agronomy Panel Survey: 2016 and 2017	CIMMYT	To understand variability in management and yield, farmer decision making, and as a baseline for nutrient management tools.	Ethiopia, Tanzania, and Nigeria	Household and community level panel dataset. The 2016 dataset consists of 607 households, while the second year contains 580 observations.	Questionnaire and codebook.
More Milk in Tanzania (MoreMilkiT): 2014	ILRI	Secure income for marginalized dairy producing communities through dairy market hubs	Tanzania; regions of Morogoro and Tanga and districts of Kilosa, Mvomero, Handeni, and Lushoto.	Baseline survey interviewed 932 households in 2012, and the monitoring survey, interviewed 461 households.	Questionnaire and survey protocol were publicly available separately.
Marando Bora: 2011	CIP	Seeks to strengthen quality planting material for sweet potatoes	Tanzania districts of Bunda, Musoma, Misungwi, Nyamagana, Ilemela, Magu, Sengerema, Geita, and Ukerewe.	Project baseline and endline surveys. The baseline survey targeted 621 households. The endline survey targeted 732 households, with an overlap of 434 households.	Questionnaire and data dictionary.
Transforming Key Production Systems: Maize Mixed East and Southern Africa: 2016	IITA	To link agronomic data with soil data to better understand how the soil affects yield	Three sites in the Babati District of Tanzania	Soil sample analysis	Field form, PowerPoint presentation, and location map.
Tropical Legumes II (TL II): Nov-Dec 2008	ICRISAT, CIAT, and IITA	Seeks to improve livelihoods of smallholders through increasing productivity in legumes	10 countries in sub-Saharan Africa, including Tanzania	Tanzania project baseline survey; household survey of 613 households	Survey instrument.

Chart 1: Description of datasets analyzed, including the project and CGIAR institution each is associated with, as well as information on the project purpose and data type. Each of these datasets were downloaded from the GARDIAN platform, and used to create the case studies described below.

Title	Data Source	Methods
Africa Rising Baseline Evaluation Survey	Africa RISING	The case study conducts basic summary statistics, cross tabulations, ANOVA methodology, and Principal Components Analysis to create an index for wealth
Adoption Pathways Project, Tanzania	Adoption Pathways Project	The case study analysis of data from the three years in Tanzania. The case study conducted basic summary statistics from each of the three years, as well as compared these summary statistics over the three years.
Pathways to Intensification: Cross-Country Report 2013	Adoption Pathways Project	The case study compares demographic information across the five countries, and creates a list of the datasets to compare adoption of sustainable intensification practices
Analyzing survey data on technology adoption	Adoption Pathways Project	The case study replicates the results by Kassie et al. in their article, "Understanding the Adoption of a Portfolio of Sustainable Intensification Practices in Eastern and Southern Africa". The paper assesses how various plot and household-level variables affect the adoption of six sustainable intensification practices, using a multivariate probit model with plot-level data.
Taking Maize Agronomy to Scale in Africa: Tanzania	TAMASA Tanzania Agronomy Panel Survey	The case study creates a map of the geographic distribution of the communities and calculates basic summary statistics from the dataset available over the two years. This includes plot characteristics, household demographics, and fertilizer use and farm management practices.
More Milk in Tanzania (MoreMilkiT) Baseline survey	MoreMilkiT	The case study assesses and summarizes the baseline data to assess the progress to date in establishing dairy market hubs.
Marando Bora: A case study of biofortified sweetpotato adoption	Marando Bora	This case study is a replication of a study by Shikuku et al. titled “Effect of Farmers’ Multidimensional Beliefs on Adoption of Biofortified Crops: Evidence from Sweetpotato in Tanzania”. The article estimates how various beliefs held by farmers about orange-fleshed sweet potatoes affect the adoption of the variety. They use inverse probability weighting and difference-in-differences (IPW-DID) to first estimate propensity scores, then find the marginal effect of each belief on adoption. This methodology is replicated in the case study.
Soil and Land Health	Transforming Key Production Systems: Maize Mixed East and Southern Africa	This case study creates charts that compare the three sites, and maps the spatial distribution of a few key variables.
Tropical Legumes II	Tropical Legumes II (TL II)	The case study focuses on basic data cleaning and summary statistics.
Comparison of Datasets from Tanzania	All	This case study is a comparison of all of the datasets in Tanzania, to analyze patterns and trends between the various datasets available within a single region. We show the geographical distribution of the data, conduct summary statistics that compare the demographics across datasets, compare educational attainment across the datasets, and analyze information provided about fertilizer to see if there are any links between fertilizer use and household level demographic information.

Chart 2: Description of case studies produced from the above datasets, with a brief overview of the key methodologies and lessons taught in each.

Discussion:

Findability and Accessibility

A number of challenges were encountered in using the GARDIAN catalogue to find agricultural data from Tanzania. First, there was a lack of data available, with disproportionately high number of survey datasets relative to other data types. Where data was available, there were often challenges downloading the data, as many datasets were locked and had unclear policies around who was permitted to request the data, and how long the process would take to receive data. Additionally, there were portions of the data missing for many of the datasets that could be immediately downloaded—such as geographic coordinates or entire sections of the questionnaire. In some cases, the metadata and accessory materials were minimal as well, with limited information on the project origins, and a lack of information on variable names and codes, or survey instruments that would be needed to interpret the data. This created significant time investments of the data re-user to sort through questionnaires and try to understand the meaning of specific variables. Where codebooks were provided, such as in the Africa RISING dataset, the process of pulling out specific variables was much simpler for a data re-user.

Many of these challenges were exemplified in the dataset titled “Transforming Key Production Systems”. In this dataset, metadata information, such as author and other notes descriptive information, were included as values in the excel tables. This information had to be removed in order to further analyze the data. Another challenge in this dataset was that the files were uploaded in several different formats—.csv, .tab, .xls, and .xlsx—creating extra work to read in and combine the files in R. Finally, the data did not come with a summary of variable

names or description of the files. Thus, in order to identify key variables the tables had to be reviewed one-by-one, and certain assumptions had to be made about variable names.

Data Organization and Management

Further, the datasets themselves were often found to be unorganized and posed significant challenges to a second-hand user of the data. Even datasets within the same project had wildly different structures and formats, leading to additional time investment to understand the organizational structure of each. Another issue stemmed from the survey instruments themselves, where similar questions were asked in a number of contradictory ways across different projects. An example is the survey question on educational background of participants, asked in most surveys. The same question (level of education) was asked in 5 formats over five different surveys (see appendix 1). Researchers appear to disagree even on the levels of education that exist with Tanzania, as two of the surveys provided the same question type but provided different options for educational level. While it is undoubtedly important to ask survey questions as they fit to the specific research question and context, there appears to be a lack of standardization in survey design which would greatly benefit researchers who hope to analyze trends from a range of datasets. Similar issues have been encountered in the literature with other types of data, such as agronomic data.

The Adoption Pathways Project dataset exemplifies many of the challenges that are common to several of the datasets uploaded on GARDIAN. First, in looking at the panel data across three years in Tanzania, we can see that the same question was organized in three different ways across the three years of panel data (figure 3). The question on the survey was the following:

"Taking all means into consideration (own food production + food purchase + help from different sources + food hunted from forest and lakes, etc), how would you define your family's food consumption in the last year?

1. Food shortage through the year, 2. Occasional food shortage, 3. No food shortage but no surplus, 4. Food surplus."

2010 Data

Household ID	Food Security
1	1
2	2
3	2
4	4
5	3

2013 Data

Household ID	Chronic	Transitory	Breakeven	Surplus
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	0	1
5	0	0	1	0

2015 Data

Household ID	Food Security
1	Chronic
2	Transitory
3	Transitory
4	Surplus
5	Breakeven

Figure 3: Food Security Status, as represented in the three years of panel data. In each year of the survey, the data was recorded in a different way.

Each year, the question was recorded slightly differently. In 2010, a numeric code was used from 1-4 to distinguish between the four levels of food security. In 2013, separate columns were used for each level of food security, and each was a 0-1 dummy variable indicating whether the household had experienced the level of food security. The organization of the question in 2015 was similar to 2013, but instead of the numeric code the individual food security status was used in a single column for food security. By having more consistent data organization within a project dataset, the data analysis would be less complex—especially important as the questions are more advanced and difficult to discern for a data re-user.

Another example from the Adoption Pathways Project looks at data from multiple countries within a single year. While each country used a nearly identical survey instrument, there is no system in place to ensure that variable names or organization of the data is consistent across the datasets (Figure 4). This creates challenges for the data analyst to find the variables of interest—especially when there is not a codebook. While many of these challenges can be

overcome individually with simple code, extremely messy datasets can add significant time barriers and prevent researchers from reusing the data.

Country	**Variable Names**					
Tanzania	hhldid	sex	age	martstat	education	mainoccup
Mozambique	hhid	b103	b105	b107	b108	b109
Kenya	hhldid	sex	age	martstat	education	mainoccup2
Malawi	HHID	Sex	Age	Marital_Status	Education	Occupation
Ethiopia	HHID	A2	A4	A5	A6	A7

Figure 4: Educational Attainment variable names across the five countries. Even though each country used the same survey for the same project, no two countries had the same system for assigning variable names.

Replication and Reuse

Beyond organizational management of the data, a few of the case studies attempted to replicate existing publications. However, there were significant difficulties in doing so. The first major barrier in replication was that often, portions of the data were missing. For example, in the Marando Bora case studies, only two of the three years of data existed. While it was possible to use the same model, it was not possible to obtain the same results—therefore preventing true verification of the research findings. Additionally, there were challenges in certain models, where different software was used. In the Adoption Pathways Project replication case study, the paper uses a multivariate probit model with plot-level data. This was a challenge to replicate because the original code was in Stata, and the function was more difficult to perform in R (especially given that the author did not provide their code).

Another challenge in using the data to replicate research results was the difficulty in translating the raw data provided into the actual variables used in the research analysis. For example, in the Adoption Pathways Project replication, one of the independent variables was a

dummy variable for practice of minimum tillage. Two of the questions in the survey instrument corresponded to minimum tillage, and it was not clear which of the questions was used in the final analysis. While the rationale used by the author may have been sound, failure to provide an explanation leaves the second-hand data user to guess which survey questions correspond to the key variables of interest. In other instances, the number of observations differed slightly throughout the dataset, and it was not clear which observations had been omitted by the author. Even a small change in the observations could affect the ability to replicate results. This type of challenge was encountered repeatedly in case studies that attempted replication of articles, as decisions around how to clean and interpret the data were poorly documented.

On the flip side, in certain instances the significant level of processing of data prevented reusability or verification. This was exemplified in a replication dataset provided on GARDIAN by CIAT for a paper titled “Climate smart agriculture rapid appraisal (CSA-RA): A tool for prioritizing context-specific climate smart agriculture technologies” (Mwongera et al., 2017). The data had been cleaned and organized to such a level that it would have been impossible to achieve contradictory results. The possibility to re-use this data for other uses was also limited by the narrow information provided in the dataset.

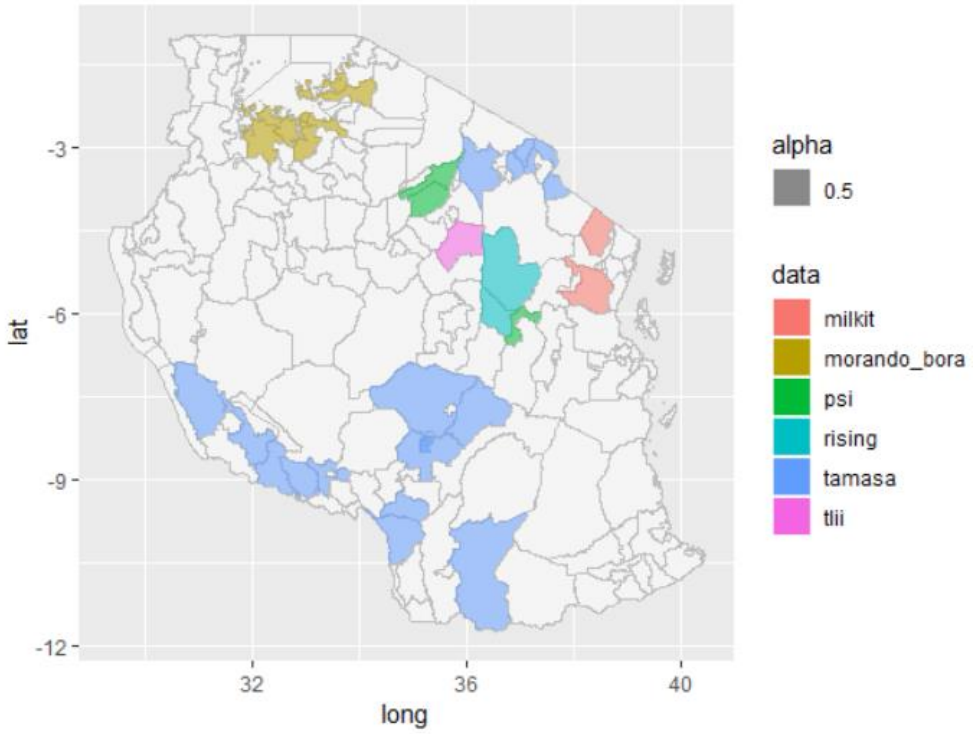


Figure 5: While combining the data from different sources increased the geographic spread across Tanzania, there were few variables that could be compared across all of the datasets. Collaboration in survey protocols could increase the reach of a single data collection incident, at little cost to the researcher.

The final case study combines several of the datasets, in attempt to see how larger questions might be answered by combining data to increase the number and geographic spread of observations (Figure 5). In each of the surveys, similar questions were asked around household demographics, input use, and other agricultural practices. Despite these similarities, it was difficult to find variables that could be compared exactly across the datasets. Slight differences in the way survey questions were phrased, or the omission of a single question, greatly reduced the ability to answer meaningful questions from a combined dataset.

Recommendations:

In light of the challenges discussed above, we recommend a number of best practices for researchers and institutions to implement in order to ensure better accessibility and usability of agricultural data. Regarding data reuse and data management, more resources and open source

tools on data must be made available to researchers and institutions. The case studies and online platform presented here for data management in R contribute to the availability of resources, and will become available freely. However, this is just one step. This platform and other resources for data management should be incorporated into curriculum for agricultural science professionals, and into institutional training at research centers and other organizations.

We also recommend that cyberinfrastructure platforms provide additional resources and guidelines for those uploading their data, especially with regard to the data structure and format. While these guidelines exist, they were not followed in most of the datasets found on GARDIAN. As part of this, cyberinfrastructure platforms should require supplementary materials such as survey instruments or variable definitions, as well as the appropriate metadata information. It should be the responsibility of the researcher to not only upload the raw data, but to provide enough supplemental information that the data can be reused by a third party. Regarding replicability, it would be beneficial to provide the code used by the author in analysis so that the pathway from raw to processed data can be better traced.

Reliable data repositories are necessary so that researchers and institutions who have an interest or are required to share data can more readily do so, and so that those searching for available data can easily find it. Because there are many existing platforms for sharing data in agricultural development, we recommend greater collaboration between these institutions in order to link these platforms for easier access to data. The GARDIAN platform is already linked to a number of repositories such as dataverse and the ILRI datasets portal, but could be further integrated into the existing network of agricultural data. Additionally, while GARDIAN is useful for finding survey data published by CG centers, there are a vast number of surveys conducted through other institutions, governments, and organizations, including small-scale grassroots

organizations. Greater collaboration between the various stakeholders would increase findability and usability of the data.

Additionally, within the field of agricultural development there should be a set of guidelines for researchers to follow as they develop survey instruments to ensure that data from multiple sources can be compared. That was noted previously by Brouder & Gomez-Macpherson in the context of conservation agriculture in Sub-Saharan Africa and South America:

> "Surveys represent an important approach to understanding the socio-economic modifiers of the potential of [conservation agriculture] to improve yields; facilitating reuse of survey data in [systemic reviews] with on-station and on-farm results is critical to advancing understanding of outcomes…we suggest an interdisciplinary team with representation from sociology, agronomy, economics and policy be tasked with developing a consensus document on minimum data and best practices for agriculture technology surveys of smallholder farmers in SSA and SA" (Brouder & Gomez-Macpherson, 2014).

The above analysis of survey data across projects in Tanzania provides further evidence of this need. For example, best practices for survey questions about educational attainment or input use would make it easier to compare these metrics across datasets in Tanzania. This could also be applied to other types of agricultural data, such as agronomic data, where standard measurements can be taken and variables used (Hunt, White, & Hoogenboom, 2001). This relates closely to the issue of semantics in data, or of common vocabularies to ensure that data is more interoperable. Currently, semantics in agricultural data is experimental and standardization has not become widespread (GODAN, 2018).

Finally, while this report only looks at data from the GARDIAN platform, it is important to consider the broader field of data sharing in agricultural sciences and in agriculture more broadly. As “big data” and computer processing power becomes more advanced, equity issues become more challenging, and there is a greater need to ensure that data can both be provided by and accessible to grassroots organizations and smallholder farmers in addition to large institutions such as CG centers. This appears to be a gap in the literature on accessibility in data sharing. Thus, as conversations continue around open data in agricultural sciences, it is important that the voices of smallholders and grassroots organizations are incorporated into the conversation.

The final case study in this capstone project was only the first step in attempting to understand the power of combining multiple datasets to increase observations and extend the reach of existing data. This case study demonstrates that it may be too early to achieve significant success in agricultural survey data, as the data were not sufficiently interoperable and thus the ability to combine multiple resources was limited. Further research should continue to assess the feasibly of combining datasets in survey and other types of agricultural data, and provide suggestions for best practices as recommended by Brouder et al.

Conclusion

This capstone report has made clear the case for better practices around sharing of data in agricultural sciences, and in research and academia more broadly. The literature review demonstrated the benefits of sharing data, including to verify research, produce new findings out of existing data, reduce time and money, and increase access to information across all institutions. While many are hesitant to share data, a number of institutional polices and measures can be put in place to incentivize data sharing, as well as to increase the reuse of this

data by others. This includes a more sophisticated and collaborative network of cyberinfrastructure platforms, addressing licensing and copyright concerns, more access to data management tools and software, and better data management practices to increase interoperability.

In producing ten case studies from data publicly available through the GARDIAN website, we were able to simultaneously assess the current status of open source agricultural data, as well as to create resources for researchers who need better support in data cleaning and management. Through this process, we presented a number of recommendations for institutions and individuals—such as improving resources and guidelines for those sharing data, increasing efforts toward creating a shared ontology in agricultural sciences and in surveys, and better collaboration between cyberinfrastructure platforms to increase searchability of datasets. Finally, it is essential to ensure that data sharing platforms include data from all sources and are accessible by everybody, not just major institutions and government agencies—especially as "big data" becomes more prominent in agricultural sciences.

Bibliography

Abbà, S., Birello, G., Vallino, M., Perin, A., Ghignone, S., & Caciagli, P. (2015). Shall we share? A repository for Open Research Data in agriculture and environmental sciences. *EPPO Bulletin*, *45*(2), 311–316. https://doi.org/10.1111/epp.12212

Altieri, M., & Koohafkan, P. (2008). *Enduring Farms: Climate Change, Smallholders and Traditional Farming Communites*. *Third World Network Penang, Malaysia*. https://doi.org/10.1046/j.1365-2540.1998.00414.x

Andreoli-Versbach, P., & Mueller-Langer, F. (2013). Open access to data: An ideal professed but not practised. *Research Policy*, *43*(9), 1621–1633. https://doi.org/10.1016/j.respol.2014.04.008

Betbeder, M. L., Damy, S., & Herrmann, B. (2017). Changes in Data Sharing Reuse Practices and Perceptions among Scientists Worldwide. *CEUR Workshop Proceedings*, *1860*, 26–40. https://doi.org/10.1371/journal.pone.0134826

Borgman, C. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi

Brouder, S. M., & Gomez-Macpherson, H. (2014). The impact of conservation agriculture on smallholder agricultural yields: A scoping review of the evidence. *"Agriculture, Ecosystems and Environment."* https://doi.org/10.1016/j.agee.2013.08.010

Carlson, J., & Stowell-bracke, M. (2013). Data Management and Sharing from the Perspective of Graduate Students: An Examination of the Culture and Practice at the Water Quality Field Station. *Portal: Libraries and the Academy*, *13*(4), 343–361.

Crosas, M. (2011). The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data.

Diekmann, F. (2012). Data practices of agricultural scientists: Results from an exploratory study. *Journal of Agricultural and Food Information*, *13*(1), 14–34. https://doi.org/10.1080/10496505.2012.636005

Eagle, A. J., Christianson, L. E., Cook, R. L., Harmel, R. D., Miguez, F. E., Qian, S. S., & Diaz, D. A. R. (2017). Meta-Analysis Constrained by Data: Recommendations to Improve Relevance of Nutrient Management Research. *Agronomy Journal*, *109*(6), 2441–2449. https://doi.org/10.2134/agronj2017.04.0215

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, *10*(2), 1–25. https://doi.org/10.1371/journal.pone.0118053

Fecher, B., Friesike, S., Hebing, M., Linek, S., & Sauermann, A. (2015). A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing. *DIW Berlin*, (February), 1–26. https://doi.org/10.2139/ssrn.2568693

Fernandez, P., Eaker, C., Swauger, S., & Steiner Davis, M. (2016). Public Progress, Data Management and the Land Grant Mission: A Survey of Agriculture Researchers' Practices and Attitudes at Two Land-Grant Institutions. *Issues in Science and Technology*

Librarianship.

Food and Agriculture Organization of the United Nations. (2012). Smallholders and Family Farmers. *Sustainability Pathways*. Retrieved from http://www.fao.org/fileadmin/templates/nr/sustainability_pathways/docs/Factsheet_SMALLHOLDERS.pdf

GODAN. (2017). LESSON 5.2: LICENSING OPEN DATA, (December), 1–4.

GODAN. (2018). LESSON 4.3: SEMANTIC INTEROPERABILITY, (December), 1–4.

Hunt, L. A., White, J. W., & Hoogenboom, G. (2001). Agronomic data: Advances in documentation and protocols for exchange and use. *Agricultural Systems*, *70*(2–3), 477–492. https://doi.org/10.1016/S0308-521X(01)00056-7

Leone, L. (2017). Addressing big data in EU and US agriculture: A legal focus. *European Food and Feed Law Review*, *12*(6), 507–518.

Linek, S. B., Fecher, B., Friesike, S., & Hebing, M. (2017). Data sharing as social dilemma: Influence of the researcher's personality. *PLoS ONE*, *12*(8), 1–24. https://doi.org/10.1371/journal.pone.0183216

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (2014). Data Curation: A Study of Researcher Practices and Needs. *Portal: Libraries and the Academy*, *14*(2), 139–164. https://doi.org/10.1353/pla.2014.0009

Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*(P1), 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

Mwongera, C., Shikuku, K. M., Twyman, J., Läderach, P., Ampaire, E., Asten, P. Van, … Winowiecki, L. A. (2017). Climate smart agriculture rapid appraisal (CSA-RA): A tool for prioritizing context-speci fi c climate smart agriculture technologies. *AGSY*, *151*, 192–203. https://doi.org/10.1016/j.agsy.2016.05.009

Pasquetto, I. V., Sands, A. E., Darch, P. T., & Borgman, C. L. (2016). Open Data in Scientific Settings. *CHI '16*, 1585–1596. https://doi.org/10.1145/2858036.2858543

Pouchard, L., & Bracke, M. S. (2016). An Analysis of Selected Data Practices: A Case Study of the Purdue College of Agriculture. *Issues in Science and Technology Librarianship*.

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLoS ONE*, 1–29. https://doi.org/10.5281/zenodo.16384

See, L., Fritz, S., You, L., Ramankutty, N., Herrero, M., Justice, C., … Obersteiner, M. (2015). Improved global cropland data as an essential ingredient for food security. *Global Food Security*, *4*, 37–45. https://doi.org/10.1016/j.gfs.2014.10.004

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0067332

Wilkinson, M. D. (2016). Comment: The FAIR guiding principles for scientific data

management and stewardship. *Scientific Data*, 1–9. https://doi.org/10.1038/sdata.2016.18

Williams, S. (2013). Data sharing Interviews with Crop Sciences Faculty: Why They Share Data and How the Library Can Help. *Wikipedia, the Free Encyclopedia*. https://doi.org/10.1126/science.aaf4545

Appendix 1: Educational Level Question

PSI

0	None/Illiterate
100	Religious Education
1	Adult Education or 1 year of education
*****	Give other education in years

Africa RISING

1	Pre-Primary
2	Adult/Vocational
2:10	Standard
11	Primary + Course
12-19	Forms
20	Ordinary Diploma
21:25	University
-95	None
-99	Don't know

MoreMilkiT
Highest Level of Education

0	No formal and illiterate
1	No formal but literate
2	Primary School
3	High/Secondary School
4	College
5	University
6	Child below school age (6 years)
7	Other (specify)

TAMASA

-99	don't know
-9	none
0	pre-school
1	standard 1
8	Standard 8
9	form 1
14	form 6
15	college 1
19	college 4
20	univ 1
23	univ 5
24	postgrad

Marando Bora

0	Pre school
0	No formal education
1	Std 1
7	Std 7
8	Form 1
13	Form 6
14	College 1
15	College 2
15	Diploma year 1
15	certificate year 1
16	College 3
16	Diploma year 2
16	Certificate year 2
17	College 4
18	graduate

Appendix 2: Example of R Markdown File

Africa Rising Baseline Evaluation Survey

Introduction

This example is an analysis of household and plot-level data from the Tanzania Africa RISING (Africa Research in Sustainable Intensification for the Next Generation) Baseline Evaluation Survey (TARBES). Following the results from the summary report, this case study shows basic data summary statistics, cross tabulations and ANOVA methodology. This follows some of the work presented in this report.

This dataset contains results from structured questionnaires in 810 households and 25 communities in Tanzania. The data are stored in multiple files with filename extension ".tab". We can get a vector with the filenames using the `list.files` function.

```
datapath <- "../data/Africa RISING/Tanzania/Baseline/data"
ff <- list.files(datapath, pattern='\\.tab$', full=TRUE)
length(ff)

## [1] 49
```

This usually refers to tab delimited text files. These can be read on with functions `read.table` or, more conveniently, with `read.delim`. But since we have 49 files, it is more convenient to read in multiple tables in one step with the function `lapply`.

```
x <- lapply(ff, read.delim)
```

To keep track of the individual tables we can name each list element based on the filenames. The following code allows us to do that.

```
z <- basename(ff)
z <- gsub(".tab$", "", z)
z <- substr(z, 5, nchar(z))
names(x) <- z
```

Household Survey

The dataset is split into two major sections: household level, and community level. We will begin with the household level data in the `interview` table.

```
d <- x$interview
```

The survey divides the households into four groups depending on specific categories designated by the study: Africa RISING households, experiment households, members of the community that are not direct beneficiaries, and control households.

Unfortunately, the data provided uses integer codes instead of values. This is an obsolete, but unfortunately still frequently practised approach. Fortunately, the data came with a code-book, so that we can fix some of this.

First make a table with the integer code "id", and its corresponding text code and description.

```
id <- 1:9
description <- c("AR", "IE, no coupon", "IE, one coupon", "IE, multiple coupons",
                 "AR + IE, no coupon", "AR + IE, one coupon", "AR + IE, multiple coupons",
                 "non-benificary in benificiary village", "control")

code <- c("AR", "no coupon", "coupon", "coupon", "no coupon", "coupon", "coupon", "IB", "cont.
")

codetab <- data.frame(id=id, code=code, desc=description)
codetab
```

```
##   id      code                                  desc
## 1  1        AR                                    AR
## 2  2 no coupon                         IE, no coupon
## 3  3    coupon                        IE, one coupon
## 4  4    coupon                 IE, multiple coupons
## 5  5 no coupon                   AR + IE, no coupon
## 6  6    coupon                  AR + IE, one coupon
## 7  7    coupon           AR + IE, multiple coupons
## 8  8        IB non-benificary in benificiary village
## 9  9     cont.                                control
```

Note that four groups used are in fact aggregations of the orginal nine groups. Now we can make a new variable *tgroup*, with the text code that we can understand.

```
d$tgroup <- code[codetab$desc]
```

The below code shows how many from each group are in each district, similar to the second half of Table 2 from the survey report.

```
group <- as.factor(x$interview$group)
#There are many more groups in the survey than displayed in the summary statistics; some of th
e groups are combined, which the code below allows us to do.
levels(group) <- c("1", "2", "3", "3", "2", "3", "3", "8", "9")
levels(group) <- c("AR", "no.coupon", "coupon", "IB", "control")
district <- as.factor(x$interview$a2)
levels(district) <- c("Bab", "Kon", "Kit")
#Generates a table for number of each group in each district
table(district, group)
```

```
##         group
## district  AR no.coupon coupon  IB control
##      Bab  90       142    186  45     135
##      Kon  14         0      0  45     105
##      Kit   3         0      0  15      30
```

The next lines of code create a dataframe with summary information on household characteristics in order to replicate the household demographic information on the household survey report (table 3). The location of each variable can be found by searching the surveys themselves, as well as the household level codebook that was downloaded as part of the dataset.

```
#First create a dataframe for household information
HH <- as.data.frame(x$interview$hhid)
colnames(HH)<-"hhid"

#To calcuate the household size, we make a table that counts number of individuals per household.
size <- as.data.frame(table(x$sectionB$hhid))
colnames(size) <- c("hhid", "size")

#This takes information from just the household head.
head <- x$sectionB[x$sectionB$b2 == 1,]

#Additional household variables
HH$group <- group
HH$distt <- district
HH$size <- size$size
HH$sex <-ifelse(head$b3 == 1, 0, 1)
HH$age <- head$b4a
#To calculate dependency rate, which is the number of people in household under 15 or over 65 divided by total household population
age <- x$sectionB[, c("hhid", "b4a")]
age$b4a[age$b4a<0] <- NA
age$b4a <- ifelse(age$b4a < 15| age$b4a > 65, 1, 0)

depend <- aggregate(age$b4a, list(age$hhid), sum)
depend <-as.data.frame(depend)
depend$count <- HH$size
HH$dependency <- depend$x/depend$count
```

The survey and codebook explain what each value for education level corresponds to, and were used to determine whether individuals had attended primary or sechondary school.

```
HH$no.school <-ifelse(head$b6 == "-95", 1, 0)
HH$prim.school <- ifelse(head$b6 >= 3 & head$b6<=11, 1, 0)
HH$sec.school <- ifelse(head$b6 >= 12, 1, 0)
HH$not.lit <- ifelse(head$b7 == "-95", 1, 0)
HH$kiswahili <- ifelse(head$b7 == "1", 1, 0)
HH$eng.kis <- ifelse(head$b7 == "3", 1, 0)
HH$ag <- ifelse(head$b8 =="1" | head$b8 == "2"| head$b8 == "4", 1, 0)
HH$married <- ifelse(head$b10 != "6", 1, 0)
HH$christian <- ifelse(x$interview$a19=="1", 1, 0)
HH$muslim <- ifelse(x$interview$a19 == "2", 1, 0)
```

Finally, we create a dataframe of the summary statistics. This includes the mean and standard deviation of all variables, and then demonstrates the this data split up by group, district, and gender of household head. The final result is table 3 from the survey report.

```
mean <- apply(na.omit(HH[,4:17]), 2, mean)
sd <- apply(na.omit(HH[,4:17]), 2, sd)
dem <- data.frame(mean, sd)

#To get statistics by group
bygroup <- aggregate(HH[,4:17], list(HH$group),mean, na.rm=T)
#Save the names of each group, which is the first column
ngroup <- bygroup[,1]
#Transpose the table and create a dataframe
bygroup <- as.data.frame(t(bygroup[,-1]))
#Add back the name of each group
colnames(bygroup) <- ngroup
```

```
#Combine the summary information by group to the demographic dataframe
dem <- data.frame(dem, bygroup)
```

To get statistics by district, we follow the same process as above, using the `aggregate` function. We then present the reults in a table.

```
bydistrict <- aggregate(HH[,4:17], list(HH$distt),mean, na.rm=T)
n <- bydistrict[,1]
bydistrict <- as.data.frame(t(bydistrict[,-1]))
colnames(bydistrict) <- n
dem <- data.frame(dem, bydistrict)

#Finally, get summary statistics by gender
bygender <- aggregate(HH[,4:17], list(HH$sex), mean, na.rm=T)
bygender <- as.data.frame(t(bygender[,-1]))
colnames(bygender) <- c("male", "female")
dem <- data.frame(dem, bygender)
dem <- round(dem, 2)

library(knitr)
kable(dem, caption ="Household Demographics")
```

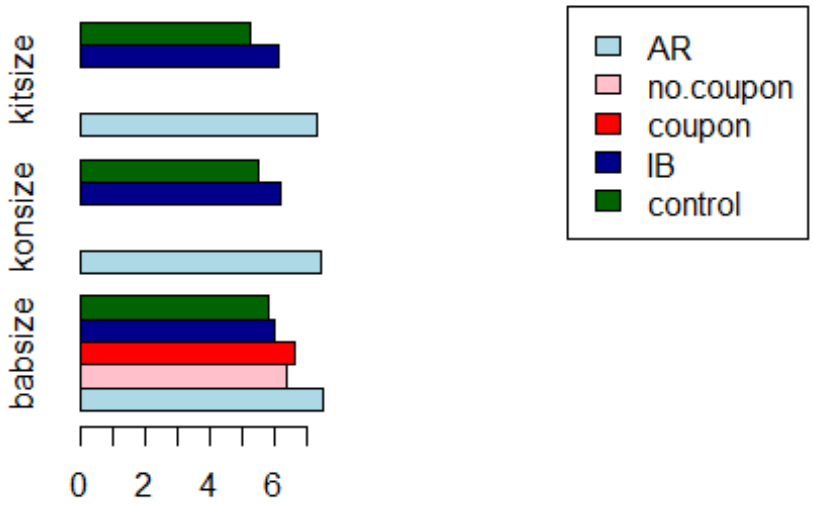
Household Demographics

	mean	sd	AR	no.coupon	coupon	IB	control	Bab	Kon	Kit	male	female
size	6.31	2.79	7.50	6.39	6.62	6.12	5.64	6.48	5.88	5.65	6.41	5.66
sex	0.14	0.34	0.10	0.09	0.13	0.12	0.17	0.12	0.17	0.17	0.00	1.00
age	47.09	14.52	51.18	46.10	47.34	43.38	47.29	47.58	46.05	44.73	46.29	52.31
dependency	0.46	0.23	0.47	0.41	0.46	0.47	0.48	0.45	0.50	0.49	0.46	0.47
no.school	0.23	0.42	0.16	0.15	0.16	0.17	0.39	0.17	0.39	0.48	0.20	0.46
prim.school	0.70	0.46	0.78	0.77	0.78	0.74	0.56	0.76	0.56	0.48	0.73	0.50
sec.school	0.05	0.23	0.05	0.07	0.06	0.09	0.03	0.06	0.06	0.02	0.06	0.03
not.lit	0.24	0.43	0.15	0.18	0.18	0.17	0.38	0.18	0.39	0.46	0.20	0.49
kiswahili	0.71	0.46	0.80	0.75	0.75	0.75	0.60	0.75	0.60	0.50	0.74	0.49
eng.kis	0.05	0.22	0.05	0.06	0.06	0.09	0.03	0.06	0.02	0.04	0.06	0.03

ag	0.88	0.32	0.83	0.89	0.90	0.84	0.91	0.88	0.93	0.79	0.89	0.87
married	0.95	0.21	1.00	0.90	0.95	0.96	0.97	0.96	0.94	0.98	0.96	0.90
christian	0.92	0.28	0.94	0.97	0.96	0.85	0.87	0.94	0.93	0.54	0.92	0.90
muslim	0.05	0.23	0.05	0.01	0.01	0.14	0.07	0.04	0.01	0.40	0.05	0.07

The summary statistics also demonstrate the household size in each district with a barplot. It is important for future research to understand if the different districts, as well as the different groups, are significantly different from eachother before the start of the program.

```
#This code generates different tables for each distrcit
Babati <- HH[district=="Bab",]
Kongwa <- HH[district=="Kon",]
Kiteto <- HH[district=="Kit",]

#Household size for each district
babsize <- tapply(Babati$size, Babati$group, mean)
konsize <- tapply(Kongwa$size, Kongwa$group, mean)
kitsize <- tapply(Kiteto$size, Kiteto$group, mean)

#Combine all district information
all <- cbind(babsize, konsize, kitsize)

#Create a barplot with the information on household sizes in each district and group
colors <-c("lightblue", "pink", "red", "darkblue", "darkgreen")
par(mfrow=c(1,2) )

barplot(all, beside=T, horiz=T, col=colors)
plot.new()
legend("topright", rownames(all), fill=colors, xpd=T)
```




```
#TODO: Adjust the legend to not cover graph...(sort of done with par and plot.new, but now the plot itself is extra small)
```

Next, we compare the means for each of the household variables across districts (included in table 4 in the survey report). In order to compare means difference, we use a t-test to generate the p-values, lower p-values indicating that the difference in means is likely to be statistically signifcant.

```
vars <- colnames(Kongwa)[4:17]

d <- data.frame(bab.vs.kit=rep(as.numeric(NA), length(vars)), bab.vs.kong=NA, kong.vs.kit=NA)
rownames(d) <- vars


for (i in seq_along(vars)){
  j <- vars[i]
  d[i,1] <- t.test(Babati[[j]], Kiteto[[j]])$p.value
  d[i,2] <- t.test(Babati[[j]], Kongwa[[j]])$p.value
  d[i,3]<- t.test(Kongwa[[j]], Kiteto[[j]])$p.value
}

round(d,3)

##             bab.vs.kit bab.vs.kong kong.vs.kit
## size             0.029       0.011       0.571
## sex              0.429       0.134       0.948
## age              0.471       0.255       0.746
## dependency       0.246       0.003       0.585
## no.school        0.000       0.000       0.257
## prim.school      0.001       0.000       0.338
## sec.school       0.130       0.999       0.207
```

```
## not.lit          0.000       0.000       0.376
## kiswahili        0.001       0.000       0.248
## eng.kis          0.497       0.003       0.461
## ag               0.153       0.065       0.037
## married          0.305       0.555       0.207
## christian        0.000       0.472       0.000
## muslim           0.000       0.001       0.000
```

We create a function that converts the p-values to stars which indicate significance level.

```
p_to_star <- function(p) {
  ifelse(p <= .01, "***", ifelse(p < .05, "**", ifelse(p <= .1, "*", "")))
}

d <- p_to_star(d)

kable(t(d))
```

	size	sex	age	dependency	no.school	prim.school	sec.school	not.lit	kiswahili	eng.kis	ag	married	christian	muslim
bab.vs.kit	* *				***	***		** *	***				***	***
bab.vs.kong	* *			***	***	***		** *	***	***	*			***
kong.vs.kit											* *		***	***

Many of the following summary statistics divide the data by wealth quintile, an index that was constructed in the survey report by using Principal Components Analysis on a number of assets reported by the household. The code below replicates this process, which first creates the index and then divides the households into quintiles of wealth.

First, all assets for the index had to be compiled into a single dataframe. Two of the tables were reshaped from wide to long so that each household was a single observation. Then, assets from three different tables were combined into a single table called “allasset”. Most of the variables were then converted to 0-1 dummy variables, depending on whether the household owned the asset.

```
#First I remove two variables that are not necessary in the reshape to count number of assets.
Group is the same for each household, and year is also not important in this process.
wide <- x$sectionO2[, -c(2,5)]
#Reshape the data from wide to long so that there are as many observations as households.
long <- reshape(wide, idvar="hhid", timevar = "assetid", direction = "wide")
#This code creates a dummy variable, so that households that have the asset are given a value
of 1, regardless of how many they own.
long[,2:38] <-ifelse(long[,2:38] >= 1, 1, 0)

#Livestock assets are also found on another table. We similarly need to reshape soo that only
livestock <- x$sectionJ1[,c(1,3,4)]
llivestock <- reshape(livestock, idvar="hhid", timevar = "j1_2", direction = "wide")
```

Next we combine all variables to a single dataframe

```
allasset <- cbind(long, llivestock)
#Remove duplicate household ID.
allasset <- allasset[,-39]
#Here we add a few variables from a different table to this dataframe, and convert them to 0 1
variables.
allasset$wall <- ifelse( x$section01$o1 == 4, 1, 0)
allasset$floor <- ifelse(x$section01$o2 == 4, 1, 0)
allasset$roof <- ifelse(x$section01$o3 == 4, 1, 0)
allasset$water <- ifelse( x$section01$o7 == 1, 1, 0)
allasset$light <- ifelse(x$section01$o9 == 1, 1, 0)
#This changes the livestock columns from 1-2 values to 0-1.
allasset[,39:59] <- ifelse(allasset[,39:59] == 2, 0, 1)

#We can compare the means from all assets to table 8 of the summary report, and see that the t
wo are the same.
means <-apply(allasset[,-1], 2, mean, na.rm=T)
means<- round(means, digits = 3)
#Two of the columns have values of all 0, and this lack of variation is not useful for our ana
lysis.
allasset <- allasset[, -c(1,14,43)]
#To preserve the information on household ID, I change the rownames to match household ID.
rownames(allasset) <- long$hhid
```

Now that the data on assets has been compiled and cleaned, we can do the principal components analysis. PCA is a statistical technique that reduces the number of variables, and each principal component that is extracted is a weighted linear combination of data from all of the variables. The first principal component explains the most variation in the data, and following the mehodology in the report, is used as the index for wealth. The final output summarizes table 8 in the survey report, breaking down the mean index for each group.

```
#This function does PCA on the table of all assets.
pr.out =prcomp (na.omit(allasset) , scale=TRUE)
#The first column in the output is the first principal component. We make this into a datafram
e, and add a column for household ID. PC1 is used as the wealth index.
PC1 <- pr.out$x[,1]
PC1 <- as.data.frame(PC1)
PC1$hhid <- rownames(PC1)

#Now, we want to add a column that tells us the quintile of wealth.
PC1.sorted <- PC1[order(PC1$PC1), ]
PC1.sorted$quintile <-1
PC1.sorted[160:321,3] <-2
PC1.sorted[322:483,3] <-3
PC1.sorted[484:645,3] <-4
PC1.sorted[646:809,3] <-5

#To replicate the chart, we take the mean and standard deviation and combine them into a singl
e dataframe.
quint.mean <- aggregate(PC1.sorted$PC1, list(PC1.sorted$quintile), mean)
quint.sd <- aggregate(PC1.sorted$PC1, list(PC1.sorted$quintile), sd)
quintiles <- data.frame(quint.mean, quint.sd)
quintiles <- quintiles[,-3]
colnames(quintiles) <- c("quintile", "mean", "SD")
kable(quintiles, caption= "Wealth Index by Quintile")
```

Wealth Index by Quintile

quintile	mean	SD

1	-2.6644612	0.4351458
2	-1.5396402	0.2863733
3	-0.4930006	0.3507865
4	0.8191148	0.4363631
5	3.7819546	2.4314999

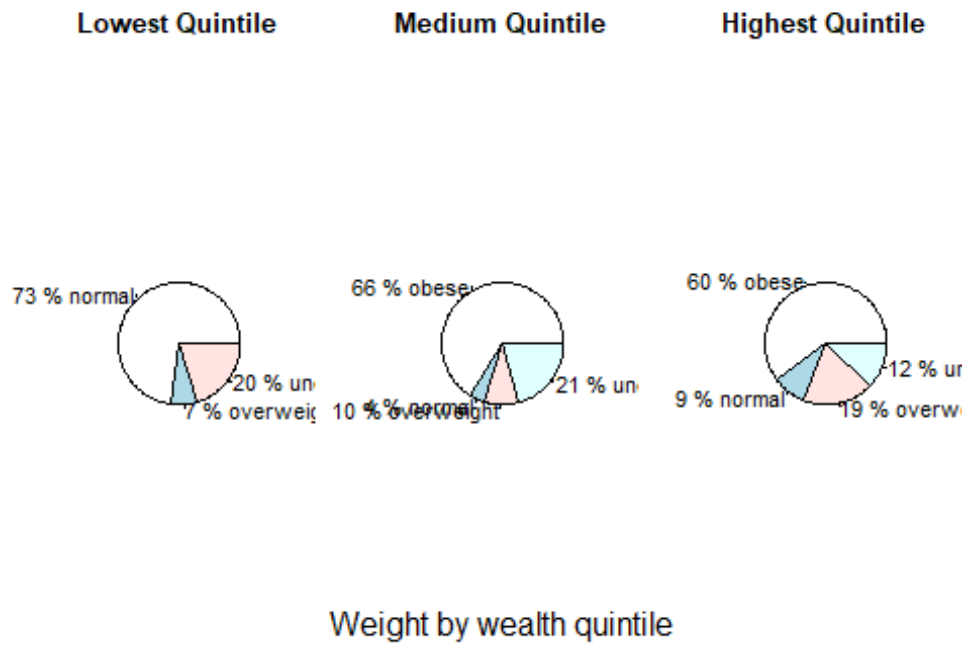
Part 4.3 of the report assess the health and nutrition status of women and children. As part of this assessment, BMI is calculated and the individuals are categorized as underweight, normal weight, overweight and obese. BMI is calulated by divided weight by height in meters squared, which is done in the following chunk of code.

```
#First create a table with just height and weight
BMI <- x$sectionT[,c("hhid", "t7a", "t8a")]
names(BMI) <- c("hhid", "weight", "height")
#The next codes first convert from cm to m, then square height and divide weight by height
BMI$height <- BMI$height/100
BMI$height <- BMI$height*BMI$height
BMI$BMI <- BMI$weight/BMI$height

#Here we merge the BMI data with the quintile data by household ID.
BMIquint <- merge(BMI, PC1.sorted, by="hhid")
#Next, we create a column that labels each of the BMI according to the distinction of overweig
ht, normal, underweight, or obese.
BMIquint$dist <-"underweight"
BMIquint$dist[BMIquint$BMI<25 & BMIquint$BMI>18.5] <- "normal"
BMIquint$dist[BMIquint$BMI<30 & BMIquint$BMI>=25] <- "overweight"
BMIquint$dist[BMIquint$BMI>=30] <- "obese"
BMIquint$dist[is.na(BMIquint$BMI)] <-NA
```

Once we have the information gathered on each individual, including their BMI distinction and wealth quintile, we create three pie charts that show the distribution of BMI for the lowest, middle, and highest wealth households.

```
#Create a table for just the poorest households
poor <- BMIquint[BMIquint$quintile=="1",]
#The table function counts the number of individuals in each category
poortable <-table(poor$dist)
#To generate a percentage
pct <- round(poortable/sum(poortable)*100)
lbs <- c("normal", "overweight", "underweight")
#Labels of the pie chart
label<- paste(pct, "%", lbs)

#The same code is used for the medium and highest quintile individuals.
med<-BMIquint[BMIquint$quintile=="3",]
medtable <- table(med$dist)
pct2 <- round(medtable/sum(medtable)*100)
lbs2 <- c("obese", "normal", "overweight", "underweight")
label2<- paste(pct2, "%", lbs2)

rich <-BMIquint[BMIquint$quintile=="5",]
richtable <- table(rich$dist)
pct3 <- round(richtable/sum(richtable)*100)
lbs3 <- c("obese", "normal", "overweight", "underweight")
label3<- paste(pct3, "%", lbs3)
```

```r
#Finally, we create all the pie charts in a single output.
par(mfrow=c(1,3) )
pie(poortable, labels=label, main = "Lowest Quintile")
pie(medtable, label=label2, main = "Medium Quintile")
mtext(side=1, text="Weight by wealth quintile")
pie(richtable, label=label3, main = "Highest Quintile")
```



We can use t.test to see if the mean BMI differs across different districts, and the results show that they do not.

```r
BMIdist <- merge(BMIquint, HH, by="hhid")
womendist <- aggregate(BMIdist$BMI, list(BMIdist$distt), mean, na.rm=T)

dist1 <- BMIdist$BMI[BMIdist$distt=="Bab"]
dist2 <- BMIdist$BMI[BMIdist$distt=="Kon"]
dist3 <- BMIdist$BMI[BMIdist$distt=="Kit"]
t.test(dist1, dist2, alternative="two.sided", conf.level=0.95)

##
##  Welch Two Sample t-test
##
## data:  dist1 and dist2
## t = 1.6453, df = 194.35, p-value = 0.1015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1064509  1.1777849
## sample estimates:
## mean of x mean of y
##  21.69188  21.15622

t.test(dist2, dist3, alternative="two.sided", conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  dist2 and dist3
## t = -1.227, df = 37.129, p-value = 0.2275
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.0416341  0.7470479
## sample estimates:
## mean of x mean of y
##  21.15622  22.30351
```

```
t.test(dist1, dist3, alternative="two.sided", conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  dist1 and dist3
## t = -0.67245, df = 33.332, p-value = 0.5059
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.461427  1.238175
## sample estimates:
## mean of x mean of y
##  21.69188  22.30351
```

```
#TODO: I tried to replicate table 14 but am confused about what is being compared in this tabl
e. The mean BMI for each category of weight, or the percent of each district that falls into e
ach weight category.
```

Community Survey

In addition to the household survey, key informants from 25 communities were surveyed about conditions within the community as a whole. These communities consisted of seven intervention villages, and 18 control villages that are meant to be comparable. This baseline survey helps to identify whether the intervention and control villages are significantly different frome each other based on a number of key indicators, discussed below.

In this section, we will conduct summary statistics on the community level data. We begin be compiling community information about the community as a whole, the chairperson of the community, and the key informants that participated in the interviews.

```
#Begin community dataframe with type and population
community <- x$sectionCF[,c(3, 4)]
colnames(community) <- c("type", "population")
community$type <- ifelse(community$type=="1", "intervention", "control")
community$type <- as.factor(community$type)
community$elevation <- x$sectionCA$ca5c

#variables for the chairperson of the community
chairperson <- x$sectionCB[x$sectionCB$cb4 =="1",]

#Add variables for chair head
community$gender.chair <- ifelse(chairperson$cb2=="2", 1, 0)
community$age.chair <- chairperson$cb3
community$years.chair <- chairperson$cb5

#Add variables for the informants in each community.
```

```
count <- as.data.frame(table(x$sectionCB$villageid))
community$no.inform <- count$Freq
x$sectionCB$villageid <-as.factor(x$sectionCB$villageid)

age.inform <-aggregate(x$sectionCB$cb3, list(x$sectionCB$villageid), mean)
community$age.inform <- age.inform$x

years.inform <- aggregate(x$sectionCB$cb5, list(x$sectionCB$villageid), mean)
community$years.inform <- years.inform$x

gender <- x$sectionCB$cb2
gender <- ifelse(x$sectionCB$cb2 == "2", 1, 0)
gender.inform <-aggregate(gender, list(x$sectionCB$villageid), mean)
community$gender.inform <- gender.inform$x
```

Next, we create a dataframe that compares the average values for the treatment, control, and all communities as a whole.

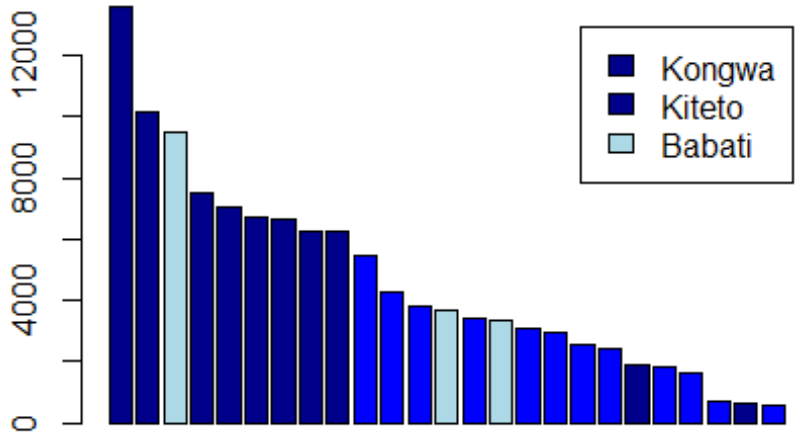
```
mean.overall <- colMeans(community[,2:10])

means <-aggregate(community[,2:10], list(community$type), mean)
means <- t(means)
means <- means[-1,]
colnames(means) <- c("intervention", "control")
means <- as.data.frame(means)

means$mean.overall <- mean.overall
kable(means, Caption = "Village, chairperson and informant characteristics")
```

	intervention	control	mean.overall
population	3797.556	6778.286	4632.1600000
elevation	1384.661	1576.414	1438.3520000
gender.chair	0.0000000	0.1428571	0.0400000
age.chair	45.72222	49.57143	46.8000000
years.chair	39.27778	38.42857	39.0400000
no.inform	4.944444	5.000000	4.9600000
age.inform	42.13175	46.19388	43.2691429
years.inform	25.00159	29.47415	26.2539048
gender.inform	0.1824074	0.3217687	0.2214286

Figure 8 of the report shows community size by district as a barplot. This next chunk demonstrates the code to do this.

```
pop <- data.frame(community$population)
pop$com <- rownames(pop)
#we can separate the observations into the three different districts
pop$distt<- x$sectionCA$ca2
pop$distt <- as.factor(pop$distt)
levels(pop$distt) <- c("Kongwa", "Kiteto", "Babati")

#Order the communities by population, from large to small
pop2 <-pop[order(-pop$community.population),]
```

```
mycols <- c("blue", "darkblue", "lightblue")
barplot(pop2$community.population, col = mycols[pop2$distt], legend=levels(pop2$distt))
```

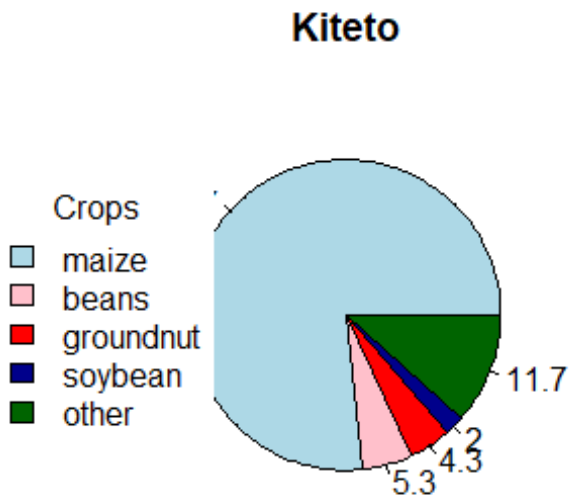



```
#TODO: Legend for barplot is messed up, and only shows two colors
```

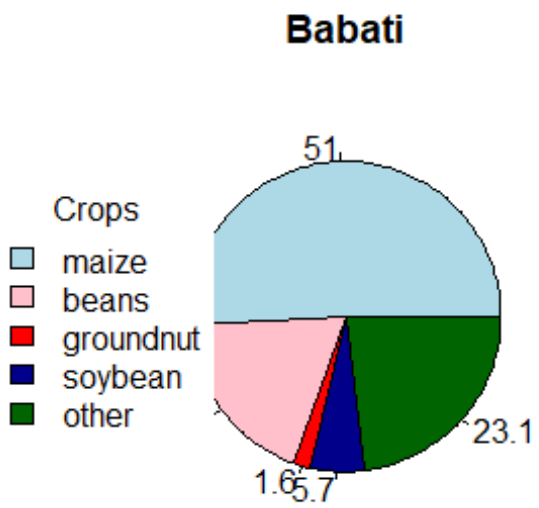
Next, we can display the main crops cultivated within communities in each of the three distrcits with a pie chart. The percentage of cultivated land area dedicated to each crop was asked to key informants in the questionnaire.

```
#Extract the four crops that were asked on the survey
crops <- x$sectionCF[,c( "cf6a", "cf6b", "cf6c", "cf6d")]
colnames(crops)<- c("maize", "beans", "groundnut", "soybean")
#The remaining percentage of land area is labeled "other"
crops$other <- 100-crops$maize - crops$beans - crops$groundnut - crops$soybean
#Next, we add information on district and find the mean percentage dedicated to each crop in e
ach district
crops$district <- x$sectionCA$ca2
crops2 <- aggregate(crops[,1:5], list(crops$district), mean)

#Next, we create  dataframes for each of the three districts.
cropBab <- crops2[1, 2:6]
cropBab <- t(cropBab)
cropBab <- round(cropBab, 1)

cropKong <- crops2[2, 2:6]
cropKong <-t(cropKong)

cropKit <- crops2[3, 2:6]
cropKit <- t(cropKit)
cropKit <-round(cropKit, 1)

#Finally, we display each district as a separate dataframe.
```

```
pie(cropKit[,1], col=colors, labels=cropKit[,1], main="Kiteto")
legend("left",legend=c("maize", "beans", "groundnut", "soybean", "other"), fill=colors, box.lt
y=0, title="Crops")
```




```
pie(cropBab[,1], col=colors, labels=cropBab[,1], main="Babati")
legend("left",legend=c("maize", "beans", "groundnut", "soybean", "other"), fill=colors, box.lt
y=0, title="Crops")
```




```
pie(cropKong[,1], col=colors, labels=cropKong[,1], main="Kongwa")
legend("left",legend=c("maize", "beans", "groundnut", "soybean", "other"), fill=colors, box.lt
y=0, title="Crops")
```