Applications of Data Science to Livelihood Sustainability in Africa: A Case Study

Matthew W Cooper[1,2], Cara Arizmendi[3], Tabby Kabui Njung'e[1], Krista Jones[4], and Robert Shaffer[5]

[1]Conservation International
[2]University of Maryland Geography
[3]University of North Carolina at Chapel Hill Psychology
[4]Smithsonian Institute
[5]University of Texas Political Science
Affiliation

July 28, 2017

Applications of Data Science to Livelihood Sustainability in Africa: A Case Study

Abstract

The Vital Signs program at Conservation International collects data in several African countries on agriculture, the environment and human well-being. The purpose of the project is to inform policy and ensure that agricultural intensification and economic development do not erode the natural capital and ecosystem services on which agriculture and human well-being depend. Having amassed a large dataset, Vital Signs lacked the capacity to fully explore all of it, and thus joined the University of Washington Data Science for Social Good (DSSG) program. This paper outlines the unique approach that Vital Signs took, including getting significant stakeholder input in drafting and planning the analyses to be conducted as well as committing two staff member to be based at the program full time. The paper further discusses the analyses conducted and methods used during the program.

Introduction

In 2013, the Vital Signs program was created at Conservation International in response to demands for a monitoring system for sustainable agricultural development (Sachs et al., 2010). Vital Signs is based on the premise that integrated data at the right scales is key to understanding trade-offs and synergies between multiple and competing development objectives in order to make better decisions for the environment and development. Feeding the worlds' growing population will place unprecedented demands on agriculture and natural resources, and governments and development NGOs alike are under pressure to increase agricultural productivity and improve the well-being and livelihoods of people in Africa, while ensuring environmental sustainability. To ensure the planned agricultural development is sustainable, there is still need for better data, better analytical methods and better risk management approaches to evaluate the trade-offs and optimize synergies among policies for food production, poverty alleviation and environmental outcomes of agricultural intensification. Moreover, much existing knowledge in agriculture, environment and livelihoods is sectoral, based on information assembled within disciplinary silos. Accordingly, decision makers often must make important agricultural development, food security and environmental conservation decisions without seeing the whole picture. To make smart decisions about sustainable agricultural development, holistic systems-level data and systems-level understanding are needed. Vital Signs is contributing to filling these critical gaps through providing an a global public good in the form of dataset that is integrated, co-located, quantitative, with near real-time measurements of agriculture, ecosystem services and human well-being metrics. Currently, baseline biophysical, household wellbeing and agricultural production is available for Ghana, Uganda, Rwanda and Tanzania. In each of the countries Vital Signs has established a formal memorandum of with government institutions mandated to either be providers or users of the data in policy making.

As Vital Signs creates an increasingly rich database, data is being collected faster than it can be analyzed and utilized. While the data has led to significant insights and publications around themes like agricultural intensification, yields resilience, natural resource gathering, and soil nutrient mapping (Hengl et al., 2017; Kanter et al., 2016; Nijbroek and Andelman, 2016), many other subsections of the Vital Signs database have been under-utilized and under-explored. Themes like water availability, nutrition, crop commercialization, and the role of education have significant data available for study but no one to analyze it. Key stakeholders, including national governments, funding agencies, and scientific partners had pressing questions concerning these topics, and Vital Signs had the data to answer these questions but not the bandwidth. In order to get more personnel focused directly on the under-explored parts of Vital Signs data, Vital Signs eagerly applied to join in the University of Washington's Data Science for Social Good program.

Vital Signs and the DSSG Program

Program Goals

Vital Signs operates with an extensive network of stakeholders and partners in each of the countries where data is collected. Before beginning the DSSG program, Vital Signs began coordinating with partners and policymakers across various disciplines to collect a list of policy-relevant and answerable questions that the data could address. These questions were generated during in-person stakeholder workshops and refined using input from field data collection teams, policy stakeholders, and data managers to ensure that the questions generated were pertinent to policymakers, could be addressed by data that Vital Signs had, and were statistically feasible. Thus, Vital Signs began the DSSG program with a significant and ambitious framework to structure the work that would be done over the course of the summer. Over the course of the summer, the questions were refined to reflect fellows' skill sets, backgrounds, and personal interests, as well as temporal constraints. These final questions answered were:

1. Are the effects of extension services on crop productivity moderated by farmers' educational attainment?

2. How does intensification of agricultural practices relate to landscape-level distribution of resources and equitable outcomes?

3. What is the effect of household natural food collection on household food expenditures?

4. How does crop commodification affect food security and childhood nutrition?

5. How does household use of different sources of water relate to the odds of experiencing water insecurity events?

6. What are the households' per capita water use in the dry season compared to WHO-recommended standards, and how do they vary by country, water source, and gender of the household head?

7. What are the predicted household agricultural yields? The model output from this question goes against the traditional approaches of predicting yields that have relied solely on agricultural inputs such as land sizes and input. Rather, this model goes further to include not only agricultural yields but also social and ecological covariates.

8. How do climatic, social, and infrastructural factors impact risk of water insecurity?

While most organizations that participate in DSSG programs are locally based with staff able to travel to work with DSSG programs a few days a week, Vital Signs staff was based in Nairobi and Washington, DC. Thus, Vital Signs sent two staff to spend the summer in Seattle at the UW campus. The Vital Signs database is complex and occasionally messy, so having staff based in Seattle full time ensured that questions and roadblocks could be addressed quickly. Furthermore, because Vital Signs was making a financial investment in the program by supporting two staff in Seattle, the Vital Signs put significant consideration into the analyses that were proposed and had a vested interest in getting usable results.

Program Deliverables: What we did and how it accomplished our goals

For each proposed analysis, Vital Signs sought to have a blog post, as well as a code base for the analysis. Blog posts are widely read and shared by professionals in international development and are readily understandable by non-scientific audiences. They are also easy to produce, giving fellows the time to address many questions and analyses over the course of ten weeks. At the same time, having a final, published product ensures that the fellows' work is communicated and something exists that can be picked up and built upon at a later date. Because blog posts are not peer-reviewed, Vital Signs made sure to give the appropriate caveats and qualifications in the blog posts produced.

In addition to blog posts, the fellows wrote more detailed analyses using R Markdown, showing the steps that they took, including data cleaning, summarizing, and model validation. These R Markdown documents are designed for data scientists and other analysts to be able to quickly pick up where the fellows left off, and to fully vet the work that was done.

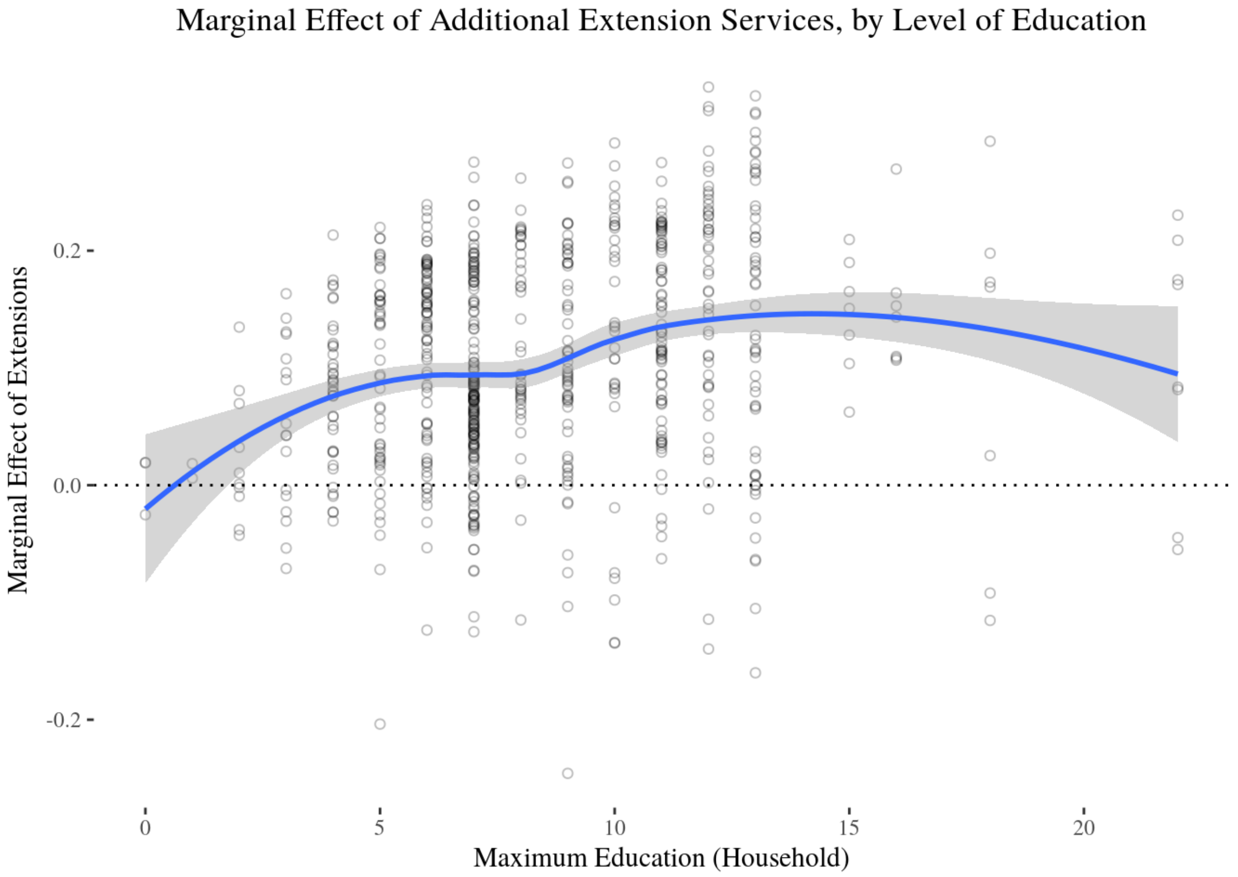
Listed below is a sample of the analyses conducted, the methods used, and the results obtained.

Extension Services, Farmers' Education, and Yields

This analysis explored whether trainings for farmers provided by governments or nonprofits, also known as extension services, had an effect on crop productivity and whether this effect was moderated by farmers' educational attainment. Existing studies (e.g., Davis, 2008; Davis et al., 2012) associate extension services with negligible effects on higher-education households. We examined whether a similar effect can be observed when examining a broader set of extension services, and found that the effect of extension services is estimated to be small and positive, but roughly equivalent amount across all levels of educational attainment.

For the purposes of the analysis, our two primary predictor variables of interest were household-level receipt of extension services and household-level educational attainment. The types of extension services used in this study were trainings on Agricultural Production, Agro-processing, Marketing, and Livestock Production. For education, we used the maximum individual-level educational attainment in years of schooling. This is because older household members often receive fewer opportunities to obtain educational than their younger counterparts, who may be able to "translate" advice. For other variables, we controlled for country, total area farmed, household size, age of household head, gender of household head, and median household-level field distance to road and market. We also included variables corresponding to field ownership and shared field usage, discretized into "All Owned"/"Some Owned" /"None Owned" and "All Shared"/"Some Shared"/"None Shared", respectively.

To model the relationships between the dependent variable (logged total crop value by household) and the predictor variables described above, we used Hainmueller and Hazlett (2013)âĂŹs Kernel-Regularized Least Squares (KRLS) approach. KRLS is a flexible regression model which treats each observation as a weighted combination of the values of all other observations in the dataset, weighted by their similarity to the given value. This approach allowed us to easily explore mediating relationships like the one outlined in the in this paper, without making strong assumptions regarding the shape of the relationship between the variables under consideration.



Controlling for the other predictor variables described previously, we estimated that the average marginal effect of the total extension services variable is positive and significant. Under this approach, each additional extension service received is associated with approximately a 12 percent increase in total crop value. In contrast to previous studies, the effect of extension services does not appear to be moderated by education.

Natural Resources and Food Expenditures

Many Vital Signs communities rely on natural resources in their daily lives, including both food (e.g., fish) and nonfood items (e.g., building materials). These resources can potentially improve multiple aspects of wellbeing, including nutrition, finances, health, and housing (Ahenkan & Boon, 2011; Shackleton & Shackleton, 2004). We examined self-reported collection of natural resources across Vital Signs communities in Ghana, Rwanda, Tanzania, and Uganda.

We were interested in whether collection of natural foods saved people money, allowing them to spend more on nonfood items. Due to the low volume of collection reported in Rwanda and Tanzania, this analysis focused exclusively on Ghana and Uganda. We looked at the relationship between the estimated value of collected natural foods (in USD) and the proportion of a household's budget that went towards food. Because the outcome variable was a proportion, we used a generalized linear model with a beta regression for the outcome. The data showed a significant relationship, with households that collected a greater value of natural foods spending a smaller proportion of their budget on food. However, this correlation does not indicate causality; it could be that households that are able to gather more from the forests do not need to spend as much on food, but it is also possible that households that do not have as much money to spend on food seek out and are more dependent on natural resources.

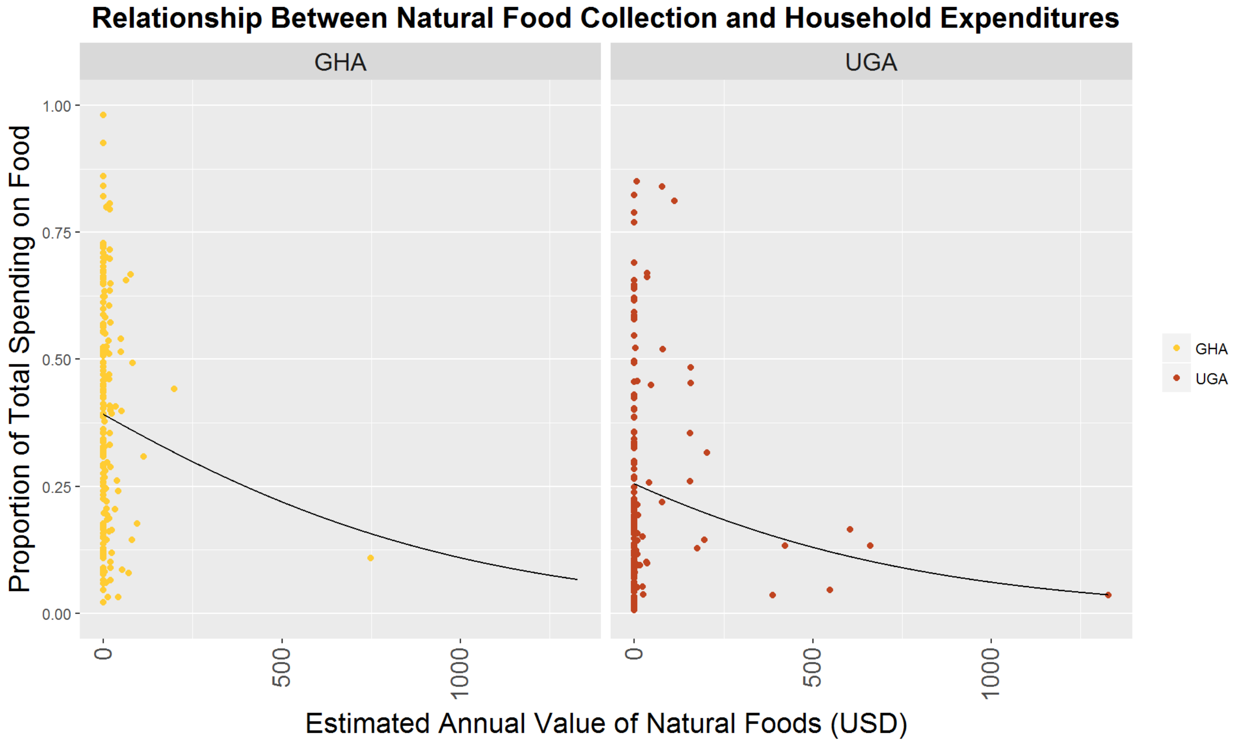


Figure 1. As households gather a greater value of natural foods from nearby forests and savannas, less of their total spending is devoted to food.

These results were heavily influenced by a small number of households that gathered a particularly high volume and variety of natural foods. Use also varied greatly by location, with one Ugandan community collectively collecting more than 5000USD annually in natural foods, nearly 3.5 times the value of the next highest community. This particular community abuts Mount Kei Forest Reserve, a protected area of more than 400km^2 that allows for sustainable use of natural resources. Thus, the presence of such a natural area clearly proves to be an extremely valuable resource for the people living nearby.

Crop Commodification and Childhood Nutrition

When subsistence farmers begin growing cash crops, they can generate income, and this can lead to positive outcomes in terms of health and education. However, producing cash crops can compete with food crop production, within households and countries. While cash cropping is a major vehicle for economic development, cash crops have been implicated in situations with worsened nutrition and food security. One example is the "Sikasso Paradox" in Mali, where the most agriculturally fertile region of Mali is a major producer of cotton, and also the country's malnutrition hotspot (Cooper & West, 2017). The earliest research on crop commercialization and nutrition found that on the whole most cases lead to increased incomes and improved nutrition outcomes, although there were cases of cash cropping leading to worsened nutrition outcomes, (Von Braun, 1995). Lately, most of the conversation in development focuses on the latter cases, and the notion that agricultural commercialization necessarily leads to improved nutrition outcomes has become a common myth (Christiaensen, 2017). Nevertheless, the latest science on the issue confirms that cash cropping does not always lead to improved nutrition outcomes, and can even have ambiguous or negative effects (Carletto, Corral, & Guelfi, 2017).

For these initial analyses, we were interested in how crop commodification is related to food security. Ultimately, we wondered if increased levels of crop commodification impacts nutrition in children under the age of five, a time sensitive development period. Looking at data from 489 households in Rwanda, Ghana, Uganda, and Tanzania, we first ran a structural equasion model relating crop commodification to food security and food security to child nutrition. We measured food security as a function of the amount of food consumed per person and months of insecurity reported by households and child nutrition as height and weight in relation to predetermined standards. For crop commodification, we use the Crop Commercialization Index, or the percent of the total value of a household's agricultural output that was sold (Carletto et al., 2017). Ultimately, we observed relationships between crop commodification and food security in addition to a relationship between food security and child nutrition. However, we observed no direct relationship between child nutrition and crop commodification.

This initial exploration raised interesting questions. We could not attribute causal relationships to the relationships between commodification, food security, and childhood nutrition. Additionally, there may be latency between the time crop commodification increases and the time we begin to see changes in children's health. Ultimately, continuing to collect these data will provide a breadth and depth of knowledge for understanding what types and levels of commodification are beneficial, neutral, or potentially detrimental to food security and child development.

Water Sources and Water Insecurity

In our initial exploration of water security, we were interested in how dependence on different sources of water (i.e., ground, surface, rain) relate to water security. Participants in the survey identified their primary source of water during each season (rainy and dry), and sources were coded as ground, surface, rain, or multiple sources. Additionally, participants were asked to report whether they experienced a water insecurity event in the past 12 months. A water insecurity event was defined as being faced with a situation when the household did not have enough water to meet their family needs.

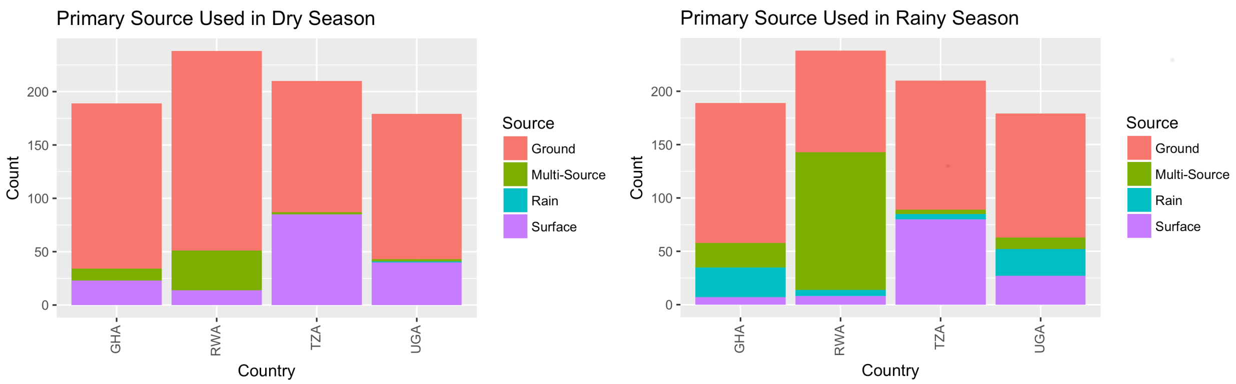


Figure 2. Household sources of water during the wet season and the dry season, by country.

Looking at the source of water used, we saw that regardless of country, most participants rely primarily on ground water, especially during the dry season. While all countries show an increase in use of rain water or multiple sources in the rainy season, Rwanda relies on rain water and multiple sources to a greater degree than other countries surveyed. For most, rain water is one of the multiple sources used. Tanzania showed the lowest proportion of participants relying on rain water, as surface water is more abundant there. We also found that households in Uganda and Ghana were more likely to experience water insecurity events, while households in Rwanda and Tanzania were less likely.

We fit a logistic regression model to each country to see if reliance on a particular water source during the dry seasons and rainy seasons was related to an increased probability of experiencing a water insecurity event. We found that in Rwanda, Ghana, and Uganda, controlling for source in the dry season, reliance on rainwater in the rainy season was associated with an increased probability of experiencing a water insecurity event in the past 12 months. In Rwanda, those who use multiple sources of water during the rainy season, and in Ghana, those who rely on surface water during the rainy season are also more likely to experience an event. However, in Uganda, reliance on surface water during the dry season, but not the rainy season, is related to a water insecurity event. In Tanzania, where surface water is more available, we found no relationships between source of water used and likelihood of an event. In many countries, relying on rainwater during the rainy season may become problematic during times when rainwater is not available. Other aspects of the relationship between source and water insecurity are going to be country specific.

Per Capita Household Water Use

Water is a critical resource for communities, as it is necessary for a range of activities, including hydration (for both people and animals), food preparation, hygiene, and crop production. Indeed, the United Nations declared access to safe, clean drinking water and sanitation a human right "essential to the full enjoyment of life and all other human rights" (General Assembly “Resolution No. A/RES/64/292,” 2010). In this post, we focus specifically on household water use for drinking, cooking, and cleaning/bathing. The World Health Organization (WHO) recommends between 50-100L per person per day in order to fully meet both consumption and hygiene needs (Howard & Bartram, 2003). However, actual use in many communities in sub-Saharan Africa fall below these levels.

Households in Vital Signs communities were surveyed as to the frequency of their water collection and the number of containers filled each time. For the purposes of this analysis, we assumed that all containers were 20L (the volume of the often-used jerry cans). In the wet season, a number of households rely on rainwater collection for household use, but we did not have sufficient information on catchment surface area and storage volume to estimate use in those cases. Thus, we focused specifically on volume of use in the dry season, when there is also more likely to be insufficiencies. In addition, this analysis excluded households with water piped directly into their homes (1 household in Uganda, 6 in Ghana, and 5 in Tanzania) and those with large storage tanks allowing use of rainwater in the dry season (1 household in Uganda and 2 in Ghana), as we were similarly unable to estimate their volume of use.

The mean household water use per capita in the dry season fell below recommended WHO levels but varied by country, ranging from 11.7L in Rwanda to 28.2L in Ghana. Volume of water use was influenced by a variety of factors, such as the age and gender of the head of the household, with homes having older and female heads using more water. In Uganda and Rwanda, homes utilizing both surface and ground water used more water overall than homes using just one source type. Thus, it appears that having a diversity of water sources available may be more important for communities in those countries that are more water-limited. This finding emphasizes that, even in the presence of sources of ground water (e.g., wells or bore holes), surface water sources (e.g., springs, rivers, lakes) remain important and must be protected. Future work will further investigate how these patterns of use relate to water security and broader climatic patterns.

RVitalSigns: Reproducible Data Management and Analyses

One of the major goals for this summer was to structure Vital Signs data processing and analysis code in a reproducible, modular fashion. As part of this goal, we developed *rvitalsigns*, an R package designed to assist with common data analysis and processing tasks. The core function of the package is an S4 class that inherits querying and database interfaces from dplyr, with additional slots used to carry frequently-used metadata (e.g. landscape geometry, landscape names and descriptions, household identifiers). Additional functions contained within the package use leaflet and ggplot2 to streamline common visualization tasks. Other features of the package include normalizing currencies to 2009 US dollars to compare financial values across countries. Following the conclusion of the summer, maintenance of the *rvitalsigns* package will be transferred to the Vital Signs data science team for use in their day-to-day workflow, including automating database querying and summarizing for several data visualization platforms.

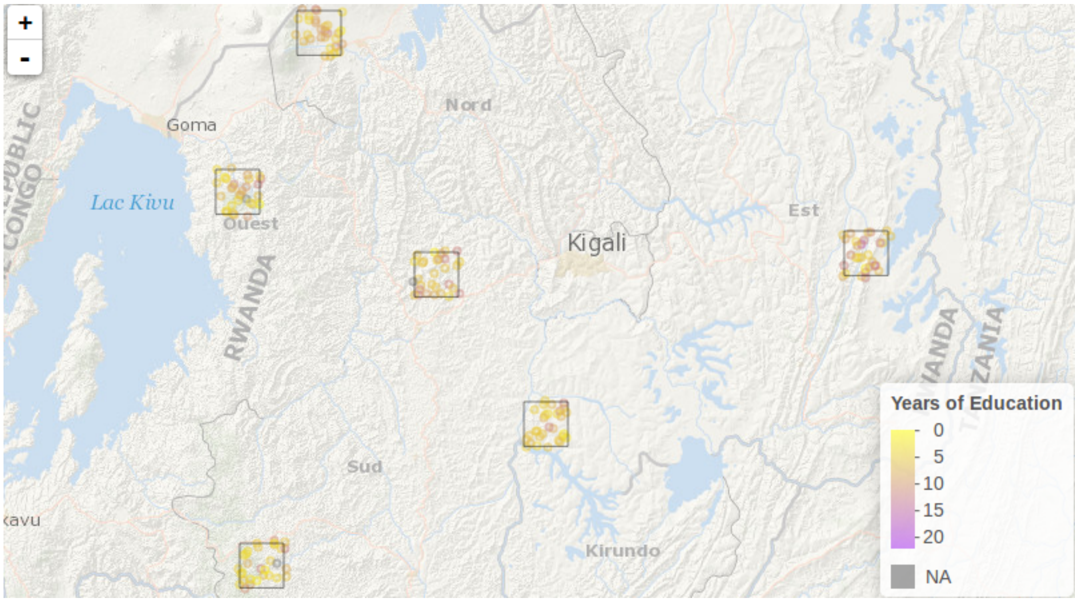


Figure 3. Household-level maximum education values from the Vital Signs dataset, plotted using leaflet and the `map_households` function in *rvitalsigns*.

Future Directions

In the final weeks of the DSSG program we will build upon current analyses of Vital Signs data using independent data from NASA's Land Data Assimilation Systems (LDAS). With this data, we will use a temporal-spatial framework to model the relationships between climatological patterns, human behavior, and likelihood of water insecurity. From our model, we will create an interactive web application where users can explore the impact of a variety of factors, including rainfall patterns, temperature trends, and water source on water insecurity. An interactive web application will meet the goals of the DSSG program by providing answers to the policy-relevant questions outlined in our goals and facilitating public engagement with Vital Signs data.

Beyond the DSSG program, Vital Signs will share the results with a broader audience including policy relevant stakeholders in the countries of operation. These results will be presented by Tabby Njung'e, one of the team leads, to partners and policymakers in Nairobi. Similarly, Vital Signs will also continue to build upon some of the research utilizing a second round of data collection with an aim of identifying trends in different phenomena.

References

Ahenkan, A. & Boon, E. (2011). Improving nutrition and health through non-timber forest products in ghana. *Journal of health, population, and nutrition*, *29*(2), 141.

Carletto, C., Corral, P., & Guelfi, A. (2017). Agricultural commercialization and nutrition revisited: empirical evidence from three african countries. *Food Policy*, *67*, 106–118.

Christiaensen, L. (2017). Agriculture in africa–telling myths from facts: a synthesis. Elsevier.

Cooper, M. W. & West, C. T. (2017). Unraveling the sikasso paradox: agricultural change and malnutrition in sikasso, mali. *Ecology of food and nutrition*, *56*(2), 101–123.

Davis, K. (2008). Extension in sub-saharan africa: overview and assessment of past and current models and future prospects. *Journal of International Agricultural and Extension Education*, *15*(3), 15–28.

Davis, K., Nkonya, E., Kato, E., Mekonnen, D. A., Odendo, M., Miiro, R., & Nkuba, J. (2012). Impact of farmer field schools on agricultural productivity and poverty in east africa. *World Development*, *40*(2), 402–413.

Hainmueller, J. & Hazlett, C. (2013). Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, *22*(2), 143–168.

Hengl, T., Leenars, J., Shepherd, K. D., Walk, M. G., Gerard, H., Mamo, T., . . . Kwabena, N. A. (2017). Soil nutrient maps of sub-saharan africa: assessment of soil nutrient content at 250 mspatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*.

Howard, G. & Bartram, J. (2003). Domestic water quantity, service level and health.

Kanter, D. R., Musumba, M., Wood, S. L., Palm, C., Antle, J., Balvanera, P., . . . Scholes, R., et al. (2016). Evaluating agricultural trade-offs in the age of sustainable development. *Agricultural Systems*.

Nijbroek, R. P. & Andelman, S. J. (2016). Regional suitability for agricultural intensification: a spatial analysis of the southern agricultural growth corridor of tanzania. *International journal of agricultural sustainability*, *14*(2), 231–247.

Resolution no. a/res/64/292. (2010). Retrieved from http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/64/292

Sachs, J., Remans, R., Smukler, S., Winowiecki, L., Andelman, S. J., Cassman, K. G., . . . Fanzo, J., et al. (2010). Monitoring the world's agriculture. *Nature*, *466*(7306), 558–560.

Shackleton, C. & Shackleton, S. (2004). The importance of non-timber forest products in rural livelihood security and as safety nets: a review of evidence from south africa. *South African Journal of Science*, *100*(11-12), 658–664.

Von Braun, J. (1995). Agricultural commercialization: impacts on income and nutrition and implications for policy. *Food policy*, *20*(3), 187–202.