

A High-Frequency Mobile Phone Data Collection Approach for Research in Social-Environmental Systems: Applications in Climate Variability and Food Security in Sub-Saharan Africa

Stacey A. Giroux, Inna Kouper, Lyndon D. Estes, Jacob Schumacher, Kurt B. Waldman, Joel Greenshields, Stephanie Dickinson, Kelly K. Caylor, Tom P. Evans

Stacey A. Giroux

Ostrom Workshop, Indiana University, Bloomington, Indiana

Department of Anthropology, Indiana University, Bloomington, Indiana

Inna Kouper

School of Informatics, Computing and Engineering, Indiana University, Bloomington, Indiana

Lyndon D. Estes

Graduate School of Geography, Clark University, Worcester, Massachusetts

Jacob Schumacher

Department of Geography, Indiana University, Bloomington, Indiana

Ostrom Workshop, Indiana University, Bloomington, Indiana

Kurt Waldman

Department of Geography, Indiana University, Bloomington, Indiana

Joel T. Greenshields

School of Public Health, Indiana University, Bloomington, Indiana

Stephanie L. Dickinson

School of Public Health, Epidemiology and Biostatistics, Indiana University, Bloomington, Indiana

Kelly K. Caylor

Department of Geography, University of California, Santa Barbara, California

Bren School of Environmental Science and Management, University of California, Santa Barbara, California

Tom P. Evans

School of Geography and Development, University of Arizona, Tucson, Arizona

Abstract

Collecting high-frequency social-environmental data about farming practices in sub-Saharan Africa can provide new insight into environmental changes that farmers face and how they respond within smallholder agro-ecosystems. Traditional data collection methods such as agricultural censuses are costly and not useful for understanding intra-annual and real-time

decisions. Short-message service (SMS) has the potential to transform the nature of data collection in coupled social-ecological systems. We present a system for collecting, managing, and synthesizing weekly data from farmers, including data infrastructure for management of big and heterogeneous datasets; probabilistic data quality assessment tools; and visualization and analysis tools such as mapping and regression techniques. We discuss limitations of collecting social-environmental data via SMS and data integration challenges that arise when linking these data with other social and environmental data. In combination with high-frequency environmental data, such data will help ameliorate issues of scale mismatch and build resilience in environmental systems.

Keywords: high frequency data; farming; food security; Short Message Service (SMS); sub-Saharan Africa

Highlights

Many geographic regions suffer from sparse social-environmental data resources.

Data collected from farmers via SMS can address gaps in social-environmental data.

Data gathered via SMS present unique quality, management, and use challenges.

1. Introduction

Smallholder farmers provide up to 80% of the food supply in Asia and sub-Saharan Africa (SSA) (FAO 2013), and small farms (i.e., less than two hectares) operate about 12% of agricultural land in the world (Lowder et al. 2016). These farmers live in an uncertain environment where climate variability is tightly related to the potential for agricultural decisions to ensure food security (Kotir 2011, Mendelsohn 2008). Achieving food security means understanding, among other things, the ways in which these farmers make agricultural decisions and adapt to environmental shocks (Burnham and Ma 2016, Harmer and Rahman 2014). However, researchers' and other stakeholders' ability to meet these objectives continues to be hampered by not only a lack of consistent, quality data about farming households (Carletto et al. 2013), but also by a disconnect between such socio-economic data and climatic data, and the fact that these data are generally collected at temporally coarse scales that are mismatched with the processes being investigated (Cumming et al. 2006). Many agricultural decisions such as when to plant, fertilize, and harvest are tightly linked to weather patterns. For example, for the majority of farmers in Sub-Saharan Africa, who are smallholders lacking access to irrigation (Burney et al. 2013, Debats et al. 2016), optimal planting dates tend to fall at the start of the rainy season. If the start of the rainy season is delayed or is rendered unclear by intermittent rains or storms, the risk of losing a crop to early-season floods, or planting too early, increases

significantly. Farmers must cope with both sudden weather events that have immediate impacts on crops, as well as learn to adapt to changing weather over many agricultural seasons.

Typically, annual, national-level agricultural censuses take the form of crop forecast and post-harvest surveys. Such surveys have been used to assess food security of smallholder farmers and to collect data about farmer management practices, such as planting dates, area planted, and total harvest quantities. However, these surveys may suffer from recall bias (Beegle et al. 2012, Tourangeau et al. 2000), are expensive to administer, and typically do not integrate climate or spatial data. As Carletto et al. (2013) note (specifically for Africa), agricultural data quality can also be impacted by questionable data collection standards, reliance on improperly drawn or incomplete samples, and inconsistency within measured variables over time or between measures collected in different locations.

An additional problem posed by relying on traditional survey-based methods is that such infrequent, cross-sectional data do not capture many intra-annual activities and decisions. Traditional survey data are collected at a single point in time from a cross-section of respondents who are chosen to be representative of a larger population. Sequential cross-sectional surveys can be combined to form a panel, which can be used to understand changes in practices or behavior over time. Time intervals range from relatively brief periods in the case of, for example, a treatment or experiment, to years apart, for example, in longitudinal health studies. When the topics of interest are farming and food security, where individuals make multiple choices each week or each month about how to manage their crops and provide adequate nutrition for their households, these traditional methods do not provide enough context and information.

Social-environmental data collection via mobile phones looks to be one of the more promising avenues for reaching many people with high-frequency data collection where remoteness poses obstacles to frequent in-person interviews. We organize the opportunity for mobile-phone data collection (MPDC) into the following key domains: 1) MPDC enables collection of high-frequency social data to better understand intra-annual dynamics and decision-making, 2) MPDC allows data to be collected in near-real time which can enable faster response to environmental shocks and disturbances, 3) MPDC reduces the cost of collecting data over large spatial extents and for large populations by removing dependence upon personnel and hardware resources, 4) MPDC imposes less of a burden on respondents because it is administered via a tool used in daily life.

Rates of mobile phone ownership in SSA are growing rapidly. As of 2016, there were nearly 75 mobile cellular subscriptions per 100 people in SSA (The World Bank 2016). Pew Research Center found that, for countries they surveyed in Africa, 75% of adults owned a cell phone, which they used most commonly for sending text messages (Short Message Service, or SMS) (Pew Research Center 2015). A report by Ericsson in 2014 predicted a doubling of voice call traffic and 30 million mobile subscriptions across sub-Saharan Africa by 2019 (Ericsson 2014). Alongside the increasing penetration of cell phones, research has shown a correlation between cell phone usage and livelihood gains: for example, cell phone access has helped improve farmers' agricultural outcomes (Aker and Ksoll 2016).

Several recent studies have shown the feasibility of MPDC for high-frequency data collection, in SSA and elsewhere (e.g., Bell et al. 2016, Hoogeveen et al. 2014, Garlick et al. 2016, Leo et al. 2015). Rather than focusing on SMS explicitly, these studies compare various phone-based survey modes, including interactions with a human or a computerized agent (e.g., via interactive voice response (IVR) or unstructured supplementary survey data (USSD) protocols). One well-known exception is the World Food Programme's (WFP) mobile Vulnerability Analysis Mapping (mVAM) program, which tracks food security and vulnerability in the developing world (Bauer et al. 2013; Mock et al. 2016; Morrow et al. 2016). mVAM utilizes multiple modes for data collection, including SMS, and integrates some spatial data variables, including country and administrative units such as camps and villages within each country. More often, SMS technology has been used in SSA for health-related monitoring and information sharing. These efforts include disease surveillance and morbidity estimation (Cinnamon et al. 2016; Mwingira et al. 2017), health quizzes (de Lepper et al. 2013), information-based interventions (de Tolly et al. 2012), or pushing health-related reminders to individuals (Pop-Eleches et al. 2011).

More recently, researchers have used other technology to help gather high-frequency social or behavior data related to the environment. These efforts include using smart meters to measure water use and quantity (Horsburgh et al. 2017) and energy use (Raimi and Carrico 2016). Researchers have also harnessed the power of crowdsourcing data to ameliorate coarse or missing datasets, for example, to create gap-free, daily snow cover maps (Kadlec and Ames 2017) and improve land cover information (Fritz et al. 2012; Estes et al. 2016). Yu et al. (2017) developed a smartphone application to collect geotagged agricultural land system information from citizens, thereby allowing for improved understanding of changes in agricultural land systems.

The objectives of this study are to (1) describe a software and data infrastructure for MPDC of social and environmental dynamics in smallholder agroecosystems; (2) organize and evaluate a classification of data types that can be effectively collected via SMS; (3) outline challenges and limitations of MPDC for social-environmental systems analysis, in terms of both our data collection process and the data themselves, including data quality assessment; and to (4) illustrate the power and value of SMS-based social-environmental monitoring. We present this work within the context of ongoing surveys of smallholder farmers in Zambia and Kenya, with a focus on data from Zambia. Our findings contribute to the development of a set of best practices for implementing large-scale surveys using mass, mobile communication technologies, and can be used by other research groups to work toward building their own projects in places such as SSA.

2. Data collection methods and infrastructure for high-frequency social data

2.1 Study context

Zambia has a humid, subtropical climate, with annual rainfall ranging from 500-1400 mm. The entire country, however, is vulnerable to drought and dry spells (Estes et al. 2014). The country experiences one major rainy season, with the onset of rains and maize planting typically occurring in October or November. However, farmers' perceptions are that rainy season onset is occurring increasingly later in the year (see Figure 1) (Waldman et al., forthcoming). Maize harvesting takes place in May and June, depending on factors such as the planting date, maize variety, moisture of the crop, and presence of pests.

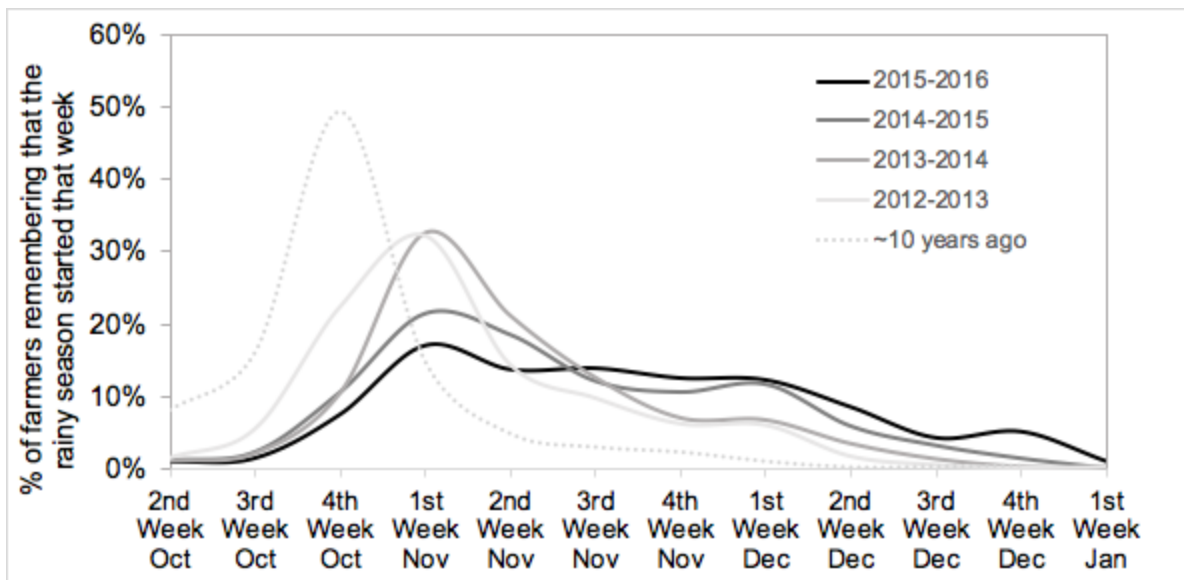


Figure 1. Density plots showing percent of farmers answering which week they estimated the rainy season began for each of a number of seasons. During in-person household interviews (not via SMS), we asked farmers (N=1021-1177) across six provinces in Zambia (Central, Copperbelt, Eastern, Northern, Northwestern, Southern; results presented here are pooled across provinces): To the best of your memory, when did the rains begin in the 2015-2016 season? 2014-2015? 2013-2014? 2012-2013? About ten years ago?

Our research team has been conducting surveys via SMS with farmers in Zambia since late 2013, and as of March 2018, just under 800 farmer households were enrolled in the SMS survey program in Zambia. SMS program households in Zambia are located across eight provinces, with a denser concentration around the city of Choma. Figure 2 shows the location of households enrolled in the SMS survey program in Zambia.

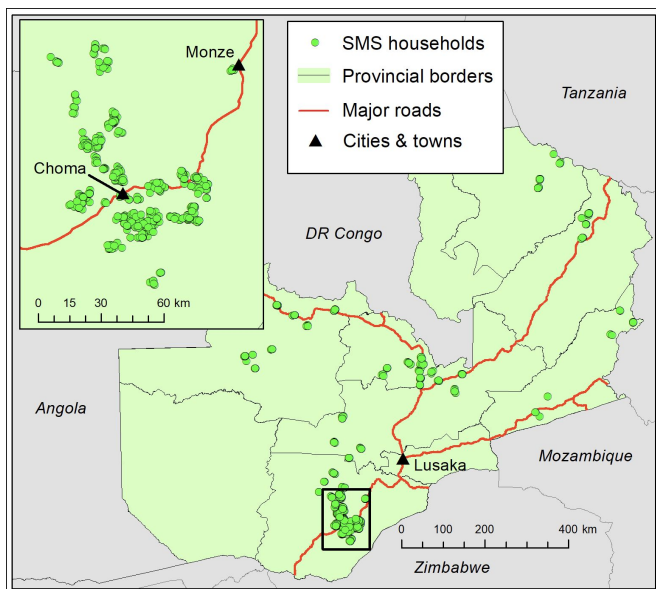


Figure 2. Location of enrolled households.

Farmers were recruited into our SMS survey system through two methods. Our initial recruitment enrolled farmers who were interviewed as part of a large-scale extensive in-person survey conducted in 2015, 2016 and 2017. The sampling design of this survey involved 40 market nodes around which a spatially stratified random sample of 30 households were surveyed. Market nodes were selected based on identification of major market locations within a set of districts representing different agroecological zones in the country. Districts constitute level 2 in the United Nations administrative boundary system (Provinces are Level 1). Four markets were identified in each of 10 districts and in most cases this constituted all the daily markets within a district. Households were selected by sampling along a series of roads (or

transects) emanating from the market node. In cases where the household settlement pattern around the market node was not uniform, oversampling in populated areas was used. The result was a spatial cloud of households concentrated around each market node location where the size of the cloud varied as a function of population density, but a majority of households were located within 10km of the market node location. The total sample population was approximately 1,200 households in 2015. A larger number of market nodes were selected as focal sample points in the Southern Province so that analyses requiring high sample density (i.e. heterogeneity of perceptions of rainfall within 5 km x 5 km grid cells) was possible. At the end of this in-person household survey documenting the demographic structure of the household, labor and farming practices, and perceptions of climate change, we asked farmers if they owned a mobile phone and were willing to participate in weekly SMS data collection. In 2015, we enrolled about 760 farmers in the SMS survey program, 310 farmers in 2016, and 230 farmers in 2017.

Our second recruitment method was through village focus groups composed of one-third farmers with < 0.5 ha in land holdings, one-third female-headed households, and one-third farmers from the village selected by village leaders and agricultural extension agents. The first two groups we intentionally recruited because of the potential for samples drawn with input from village leaders tend to undersample those two important groups. We then trained all of these respondents to use the phones to participate in the program and we chose a subset of them to participate in an annual household survey. In addition, farmers who were not responding to surveys were periodically purged from the sample, as were any farmers who requested to be removed from the program.

We use TextIt (<https://textit.in/>), a low-cost messaging platform that allows users to create SMS or voice applications for data collection. Survey question sets consist of 4-8 questions in English, designed to take no longer than three minutes to answer. These question sets are referred to as “flows,” and are built in TextIt and disseminated to farmers each week. Farmers receive a series of questions related to what time of year it is: planting, growing, harvesting, or interseason. Flows can be constructed to include skip logic or branching depending on an answer to a prior question, and can also shift from one flow to another. For example, if a farmer responds that he or she has finished harvesting all of his or her maize, he or she will automatically be switched to the interseason flow. An Android smartphone maintained in Zambia runs the TextIt application that is used to remotely send and receive the questions and answers. Over time, some questions have been adjusted slightly for clarity, or alternated with others as the research program has progressed. Farmers receive a small

payment to compensate them for participating in the survey each week in the form of talk time (value approximately \$0.20 USD), which is provided directly to their phones.

Each week a portion of the sample responds, with some farmers responding regularly and other farmers responding sporadically. The overall response rate therefore varies, as does the weekly overlap between respondents as farmers drop in and out of the response pool. Thus, unlike traditional survey data, panel data, or even data coming from meteorological stations, which are structured and lend themselves well to organization and retrieval with relational databases, data from SMS-based surveys are characterized by a highly variable structure and frequent missing values.

2.2 Broader data infrastructure

While the TextIt platform provides tools for low-cost and relatively easy data collection, its functionality does not adequately support analysis, replication, and preservation of data that are crucial for scientific research (National Science Foundation, 2007). With rapid, frequent data collection, manual methods and limited automation provided by data collection platforms is slow, cumbersome, and prone to errors, duplications, and fragmentation. To realize opportunities that are emerging from collecting new forms of data such as those discussed in this paper, we established and implemented the following principles in our approach to data infrastructure:

- **Unified storage.** A unified database allows us to store large amounts of data in one place, easily manage updates, additions, and cleaning, increase accuracy and security, and establish links to other datasets for more complicated analyses.
- **Preservation and replication.** Downloading and storing data on the servers managed by the project team provide additional backup and security. If the cloud platform becomes corrupt or unavailable, the risk of losing data is minimized as data can be easily restored from the database, which also has a backup copy.
- **Improved querying and retrieval capabilities.** A database allows data to be organized according to research needs, i.e., in anticipation of the most common queries. It also allows the addition of more metadata to enable filtering and subsetting of data for various types of analysis.

2.3 Data pipeline

The data infrastructure we developed will also facilitate long-term curation and management of these data. While data analysis techniques are at the core of many discussions on climate variability and food security, reliable data use depends equally, if not more, on what happens before and after the analysis. In our research, we incorporate the concepts of *data*

lifecycle and *data pipeline* to take better care of data and to work with big and heterogeneous data (Plale and Kouper 2017). A fully developed infrastructure supports data throughout its lifecycle, from collection to sharing and re-use. Viewing data through the lens of the lifecycle framework helps to maximize the benefit of data, minimize its cost, and improve data quality. The data pipeline is an abstraction that describes software tools and services that are applied to data objects as they go through the lifecycle.

Our data preservation and analysis pipeline combines our own innovative methods of ingesting, processing, and visualizing the data with the existing solutions for storage. This pipeline allows us to automatically direct data from its cloud collection to locations of storage and processing (Figure 3).

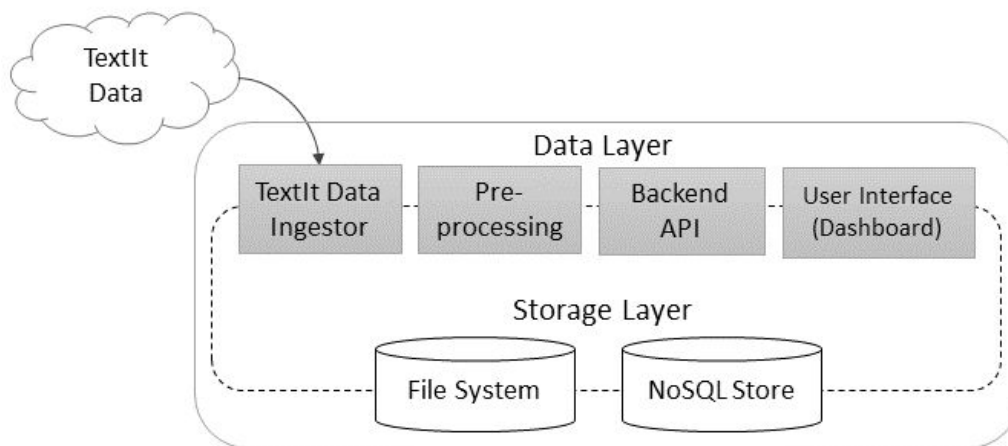


Figure 3. Data preservation and analysis pipeline.

The pipeline consists of Java and shell scripts developed by our team (TextIt Ingestor, n.d.) and is organized into modules within the data layer and a noSQL store that comprises the storage layer. We use TextIt's RESTful JSON API that has endpoints to perform bulk operations on contacts, runs, flows, and events (see <https://textit.in/api/v2/>). The data layer includes the ingestor, pre-processing, backend API, and user interface modules. The module "TextIt Data Ingestor" runs in intervals that can be defined in a configuration file. As flows are sent out at different times for Zambia and Kenya, we schedule retrievals accordingly – on Mondays at 7 AM local time (11 AM GMT) to collect Zambia data and on Saturdays at 2 AM local time (6 AM GMT) to collect Kenya data. The module retrieves available data for the last week from TextIt and saves it on disk (File System) as multiple JSON files. The pre-processing module performs

several checks: it makes sure each file was retrieved correctly, removes duplicates, and merges multiple files that belong to one flow.

The pre-processing module also contains scripts for metadata management. As the TextIt platform provides a limited number of metadata fields that can be added to describe flows and contacts, we developed tools to add additional metadata to flows and contacts. The tools allow us to pull information from the database, add additional values through automatic population of the fields or manual edits via a website, and write to the database again. The following metadata variables are added to flows: flow type (test / regular), season (planting, growing, harvest, interseason), country (Zambia or Kenya), creator, run start date, run end date. Contact information is enhanced with “Date last responded” information to provide summaries of non-responsive contacts over time.

2.4 Database solutions

The data from the disk are then inserted into a noSQL database. We use the open-source solution MongoDB as a noSQL Store. This part of the storage layer is designed as three databases per country that store raw, processed (split), and integrated documents (Figure 4).

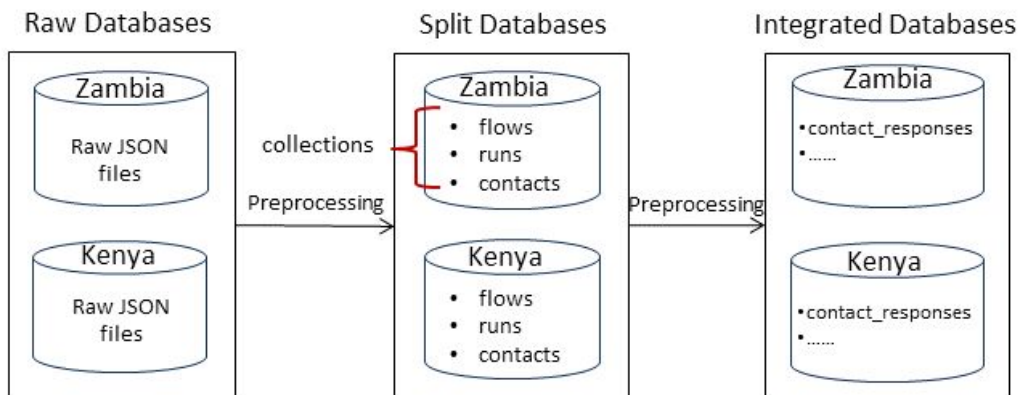


Figure 4. Database design.

In addition to being open-source, MongoDB offers benefits as a noSQL solution:

- **Conformity to the native data collection format.** Many current data collection platforms, including TextIt, store data in a document-oriented format using JSON (or other) encoding. Documents are not required to adhere to a standardized structure, i.e., they may have different sections or fields. Preserving data in a raw JSON format allows us to maintain a link between our storage and TextIt. For example, in the case of data corruption, selected documents can be uploaded back to the TextIt platform.

- Flexibility in structure.** As described above, flows change from season to season and sometimes questions are modified. Therefore, we cannot expect the survey data to have a fixed schema and design a relational database. While the data could have been transformed from JSON to a relational database, having flexibility in structure allows us to accommodate changes in questions and survey structures over time without compromising previous data or requiring change in the database design. Data storage is being separated from the application (research design and implementation) logic.
- Big data management.** As the amount of data collected will grow tremendously over time, we need a solution that allows us to manage data efficiently. NoSQL databases are known to be highly scalable for managing “big data” in a distributed environment, without compromising performance (Nayak et al. 2013).

A raw database saves data as they come from TextIt in JSON format. This is storage for preservation. The split database allows us to save data in a more logical and structured manner and avoid arbitrary partitions of data done by TextIt, such as 250 responses per file for each week that are embedded within a flow with its metadata. We extract runs (responses) from each file and organize the database into three logically consistent collections: (1) *contacts*, which contains information about respondents, (2) *flows*, which contains information about each survey, and (3) *runs*, which contains information about each response within the survey. Such structure also helps with efficiency in queries. See Figure 5 for the schema of the split database.

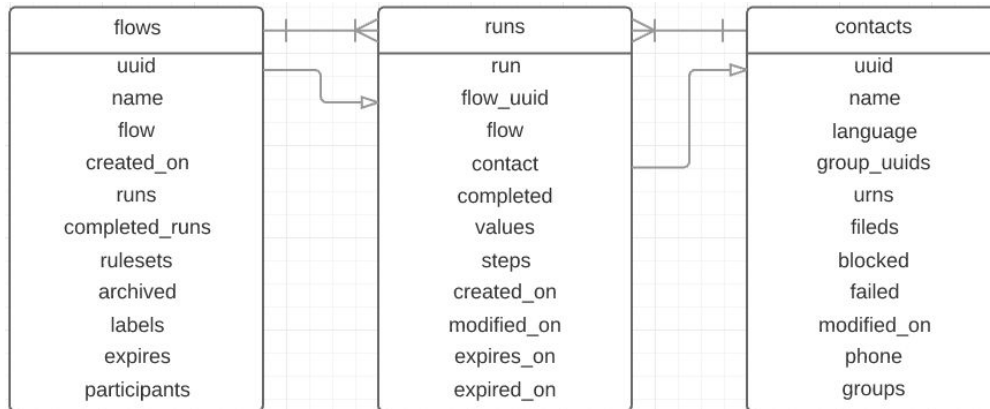


Figure 5. Split database schema.

In addition to the contact, flow, and run data, this database also contains information about the status of data retrieval from the TextIt API, which is subsequently used in monitoring both the data and the server’s health and completeness.

The integrated database consists of data that was converted into a form suitable for further analysis; that is, all responses per contact are gathered together and reformatted into a flat row-column representation rather than a hierarchical key-value pair representation. Such preprocessing minimizes repeated calculation overhead when the database is queried multiple times with different requests. Essentially, this third database is designed to accommodate the most frequent queries that researchers use on the database.

The backend API module exposes data from MongoDB to the front-end services so that data can be displayed and explored on a website or downloaded for further cleaning and analysis. The backend API is an additional layer that provides standardized access to data and at the same time prevents direct manipulation within the database. This module can be further expanded with services that facilitate computational analysis, modeling, and visualizations.

Metadata management, data monitoring, and data exploration can also be done through our user interface. A website has been developed that can serve data out of the databases and present various data products to stakeholders. Currently, the user interface provides access to flow and run completion rates as well as to summary statistics of respondents (Figure 6).



Figure 6. Web-based dashboard for data monitoring and exploration.

2.5 Automated cleaning

Data cleaning and analysis are currently done outside of the automated pipeline, using such tools as Open Refine for cleaning and R, SPSS, ArcGIS, and Tableau for analysis and visualizations. Open Refine (<http://openrefine.org/>) is a visual open-source tool for cleaning and

transforming messy data. It allows for data exploration, identifying outliers and potential errors, normalizing spelling, and correcting typos and mistakes. In Figure 7, for example, we show how Open Refine helps to identify similar answers with different spellings (e.g., FIVE / Five) and bring them to the same standard form or to change them from text to numerical form.

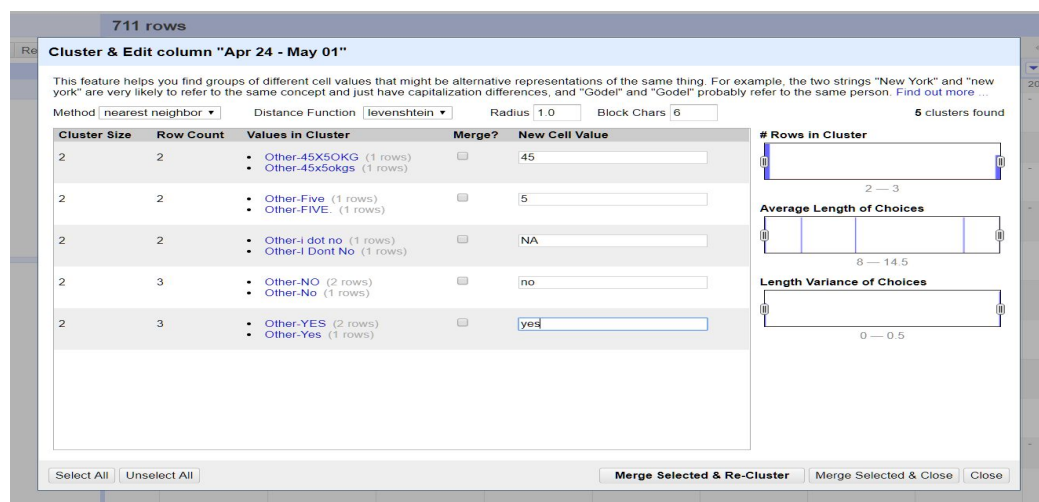


Figure 7. Open Refine for data cleaning.

Once the similar values are clustered and their correctness is verified, a standardized new value can replace all other values. However, because respondents are allowed to answer any question with open text (rather than selecting from closed-ended response options), we must take further steps to clean the data and attempt to salvage as many valid responses as possible. Our team is currently in the process of systematically implementing more intelligent methods of cleaning using syntactic similarities in values. In the absence of these more intelligent methods, our research team has a choice to either forgo data that cannot be cleaned using the methods we already employ, or clean the unique answers manually.

3. Data types

As described earlier, we survey farmers throughout the year, asking questions about planting, growing, and harvesting maize, in addition to questions outside of the maize season. We organize our questions into three categories: spatio-environmental survey questions, temporally linked questions, and event-based questions. Thinking about questions in different ways matters; a question that has less to do with the farmer's personal characteristics or farming practices and is instead related to the physical environment around the farmer (e.g., Did it rain on your fields in the last 7 days?) presents different data challenges and analysis opportunities than an event-based question related to the timing of certain crop management practices. A sample of questions we have asked farmers regularly is provided in Appendix A. In this section

we use four questions from the project to describe this typology. The question texts and administration timeframes are provided in Table 1. In section 4 of the paper (Data quality and usability), we reference a sample of the list of questions in Appendix A to illustrate data quality issues. Finally, in section 5 (Multi-temporal analysis and visualization of high-frequency data), we also discuss the questions presented in Table 1.

Question label	Question text	Question type	Administration time frame
<i>Rain</i>	Did it rain on your fields in the last 7 days?	spatio-environmental	nearly every week during the period under consideration
<i>Storage</i>	How many 50 kg bags of maize do you have in storage now?	temporally linked	every week during the period under consideration
<i>Maize buying</i>	Did you buy any maize for your household usage this week?	event-based	harvest portion of the season and the period before the next growing season
<i>Maize selling</i>	Did you sell any maize this week?	event-based	harvest portion of the season and the period before the next growing season

Table 1. Questions used as examples throughout the text

3.1 Spatio-environmental questions

We consider spatio-environmental questions to be those that ask farmers about their immediate environment and are not dependent on responses either to other questions we ask or to having answered the same question in prior weeks. The *Rain* question is of particular interest because of its potential to serve as an alternative to meteorological data in remote areas, or to be combined with satellite-based precipitation data to capture finer scale dynamics. The data from this question, coupled with mobile meteorological stations that we continue to install in the study areas, provide a more accurate representation of microclimatic variation in rainfall than the few meteorological stations that are spread across the country. The *Rain* question is less sensitive to the issue of different respondents dropping in and out each week than questions that ask directly about farmer behavior, because the quantity we seek to

calculate from the answers is not affected by values given in prior weeks (unlike planting or harvest questions). To ascertain whether it rained in a particular week in a given location, we can aggregate farmers' responses into grids and calculate the proportion of farmers responding that week who answered "yes" to the *Rain* question. The size of the grid cell chosen for this analysis is based on the desired minimum average number of farmers responding per week per cell, which is a function of enrolled farmer density and weekly response rate. Although aggregation loses spatial precision, it removes dependence on the response rate of individual farmers, while minimizing the noise caused by incorrect responses or between-farmer differences in interpreting how much rain is enough to justify answering yes to the rainfall question.

3.2 Temporally linked questions

Storage is an example of a question whose answer displays direct temporal dependency. With *Storage*, we know when households are harvesting their maize, and we know that maize storage declines over time, thus we should expect to see such a pattern of decline over the course of weeks. It could be fair to estimate that maize storage declines linearly from the point of harvest, and therefore even if there are data gaps for some weeks for some farmers, one could impute those missing values. However, the rate of change week to week and month to month in number of bags of maize in storage will differ from household to household, depending on what time of the year it is, how many people live in the household, how well off the household is in general, and their maize availability, among other things. For example, households may wish to sell some of their maize, which is one source of loss of maize in storage. Some households must sell maize as soon as or even before it is harvested, at lower prices to so-called "briefcase buyers" (local, small-scale, private buyers who enter the market early and tend to buy at lower prices), whereas those who are able to wait until later in the season to sell will find larger private buyers or the Food Reserve Agency (FRA), who generally offer higher prices. Thus, this type of question is more sensitive to missing data from individual farmers.

3.3 Event-based questions

Event-based questions attempt to collect data on discrete events that are not spatially linked nor tied to prior responses to the question over time. In our example, the events in question are buying and selling maize, which we have structured into discrete events by asking about these activities each week. *Maize buying* and *Maize selling* therefore represent the question type with the greatest potential for problematic data gaps. If a household does not

respond to a question like this for a given week, we completely miss the reporting for that event, have little recourse for recovering that information, and do not have a strong rationale for imputation. To fill in some of these data we may be able to rely on recall data from the annual, in-person household surveys we conduct, which include questions about buying and selling maize in the three months prior to the survey. This covers a portion of the same time period as that covered by the SMS questions, but not the entire period, and the household survey data are less precise and more subject to recall error. However, the household survey data are useful for characterizing the relationship between other household-level variables and the act of buying or selling maize.

4. Data quality and usability

4.1 Sample size and bias

One of the primary considerations for our research was to grow the SMS survey program large enough to have a sufficient sample size for understanding trends over time. Addressing this issue is to some degree simple: the more respondents, the better. However, because the issue of overall sample size is compounded by respondents' inconsistent participation, we find it instructive to examine what the data look like over a range of sample sizes. To illustrate, we take the *Rain* question and plot the trend over time for rainfall (i.e., the proportion of respondents saying that it rained last week over the number of respondents replying that week) at sample sizes of 10, 50, 100, 250, and 500 farmers to visually show the range of observations that could have been observed with a smaller number of farmers versus a larger number (Figure 8). We choose *Rain* as the example, but we could choose any question we ask farmers to serve as the illustration for sample trends. A Monte Carlo simulation was performed with the bootstrap method to first select n farmers from the data, with replacement, and then calculate the proportion that reported rainfall each week (noting that not all farmers responded each week). The farmers were then resampled 10,000 times, with the proportion reporting rainfall recalculated each time. The variability of responses that could have been received is described by calculating the 2.5th and 97.5th percentile of the estimated proportions of rainfall each week (95% bootstrap confidence interval) and is represented by the shaded grey region in Figure 8. The five black trendlines for each plot in Figure 8 represent five individual draws out of the 10,000 simulations to show examples of the data that might have been observed if we only sampled n farmers. Overall then, we get a good sense of how smooth the dataset becomes as we approach a sample size of 500 or larger. Recall that our farmers are located across eight provinces in Zambia, and we did not take geography into account in producing these graphs, as

they are meant to illustrate the methodological issue of sample size. More substantive analyses would take geographic location of households into account.

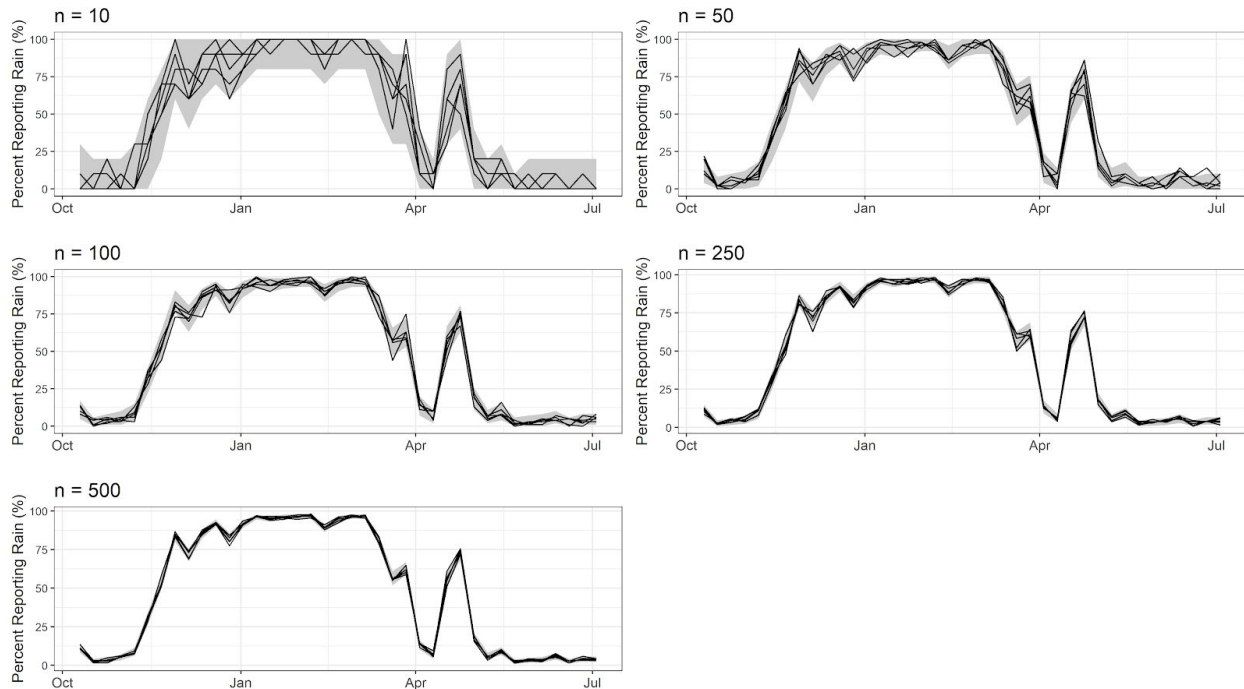


Figure 8. The five black lines in each plot represent five random simulations of the percentages that might have been observed with sample sizes of $n=10$, 50, 100, 250, or 500 farmers. The shaded grey region represents the range of possible observations, based on 10,000 replicates of a Monte Carlo simulation (95% bootstrap confidence interval). Data represent the proportion of farmers stating that it rained on their fields that week.

We recognize that our sampling design may be affected by some degree of bias. Most obviously, we are missing farmers or farming households who do not have cell phones, and we may miss some farmers who are illiterate or who do not speak English. However, we try to overcome the latter two limitations by ensuring that the farmers we recruit have household members who are able to read English and are willing to help the farmer respond to the survey.

We also performed a regression analysis to check for potential nonresponse bias with the current sample in Zambia (results not shown). We tested whether any of the following variables significantly predicted the number of weeks responding to the SMS surveys for the period from October 3, 2016 through July 10, 2017: age of household head; sex of household head; number of people living in the household; highest education level of anyone in the household; total area of farming for the last growing season; maize plot size for the last growing

season; off-farm income; and a count of the number of food security measures employed in the week prior (Coates et al. 2007). No variables were significant in the model and $R^2 = 0.022$.

4.2 Nonresponse

There is a potential tradeoff between data quality and the amount of information that can be collected. Although it seems intuitive that the longer a survey lasts, rates of survey attrition and satisficing would increase, and the likelihood of participating in future surveys would decrease, the little research that has been done on these issues has mixed results (e.g., Bogen 1996, Lynn 2014). In our project, respondents tend to drop in and out of the answer pool week to week. That is, while we have a stable response rate that tends to range from approximately 40-65%, virtually none of our farmers respond every single week. This could be due to any number of factors, from technology infrastructure and hardware issues, to respondents simply forgetting. It is difficult to estimate to what degree the fact that we make a weekly survey request accounts for respondents' spottiness, and we do not have follow-up, qualitative nonresponse data for Zambian farmers. However, we do have such data for Kenyan farmers, which we collected through follow-up phone calls with nonresponding farmers from October 2015 through April 2016. This survey provided enough data for us to understand the main reasons for nonresponse, which we believe are also likely to apply to Zambian farmers. Table 2 lists the fifteen reasons we have coded based on the qualitative data. The top five reasons for nonresponse over this period include the respondent forgetting; being too busy; not receiving the SMS; the questions no longer coming through mid-survey; the lack of cellular network.

Respondent forgot
Battery/phone charging problems
Respondent too busy
Respondent didn't get SMS
Respondent didn't get talktime
Respondent sent texts from second SIM
Next question didn't come
Accidentally deleted SMS

No network
Phone spoiled/can't type
Respondent was sick
Respondent did not have anyone to assist
Received 2 sets of questions
Respondent had traveled
Respondent had used talktime

Table 2. All reasons coded for SMS nonresponse

4.3 Usability

SMS-administered surveys pose unique challenges that must be factored into the survey design and data analysis pipeline. Screen space presents a particular consideration in terms of question wording, and the fact that most farmers in SSA do not use smartphones (as opposed to feature phones, or so-called “dumb” phones; Pew Research Center 2018) limits choices for presenting questions (Callegaro et al. 2015). Our questions require either yes/no (binary), numeric, or text responses. We provide no predetermined response options for selection. We ask very few questions that require a text answer, and we have found that answers to this type of question to be of lower quality for analysis. For example, we have attempted to ask farmers what seed varieties they have planted. Some farmers simply reply with the manufacturer name such as “SeedCo,” others will only provide the variety number “513,” while others will specify the manufacturer and the variety name, “SeedCo 513.” Numeric answers fare better. These include questions such as “How many 50kg bags of maize do you have in storage now?” Binary questions, such as “Did it rain on your fields in the last 7 days?” provide the cleanest data. We analyzed the proportion of usable answers versus any answers provided for a subset of 12 questions. By “usable,” we mean responses that are not outside the bounds of possible answers and are fully comprehensible. For example, in response to a question such as “How many 50kg bags of maize do you have in storage now?” unusable answers include “depends,” “yes,” “100,000,” and the like. We compared a set of seven yes/no questions with five text or numeric type questions, over a period from November 2015 to February 2016, and found less data loss in terms of usability for binary questions than for text or numeric (Figure 9).

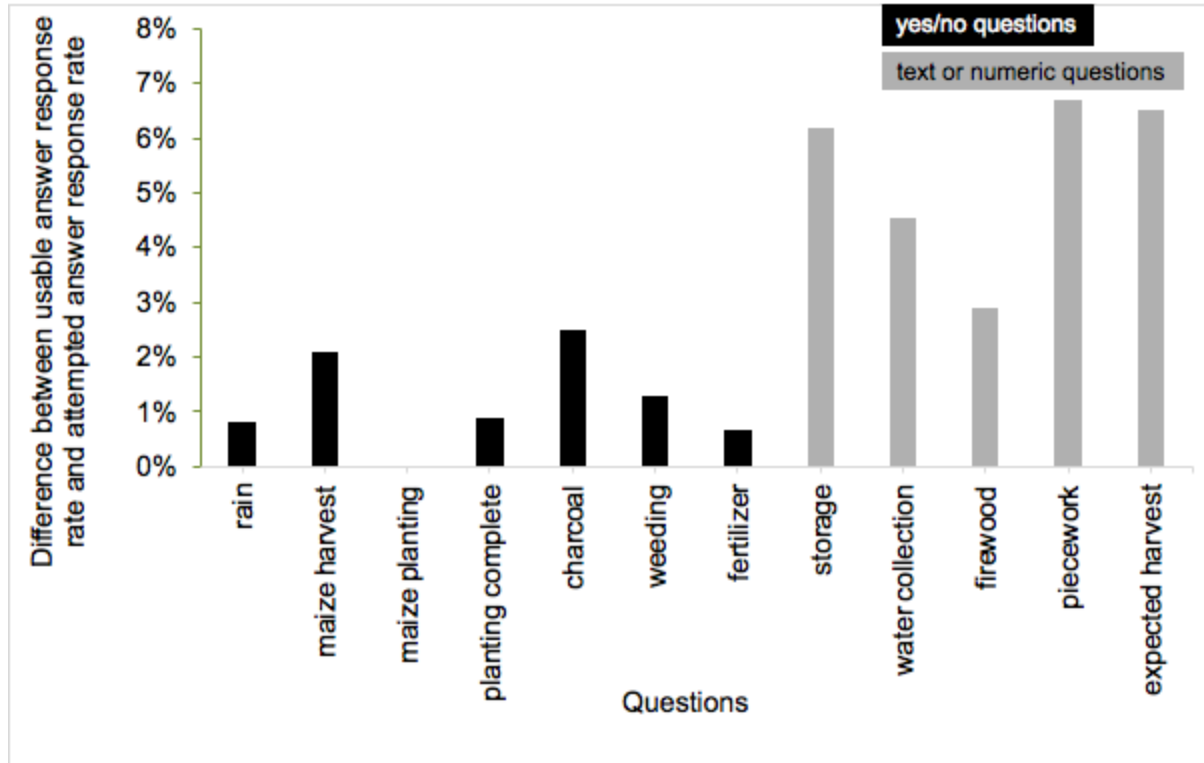


Figure 9. Data loss by question, expressed as the difference response rates between the rate adjusted for unusable data and the total rate.

5. Multi-temporal analysis and visualization of high-frequency data

In addition to challenges, the type of question asked presents diverse opportunities for recognizing patterns in the data and characterizing trends over time. We focus again on four of our survey questions, *Rain*, *Storage*, and *Maize buying/Maize selling* to illustrate the possibilities for visualizing and analyzing such high-frequency data. We examine these questions over parts of the period from the weeks beginning October 3, 2016 through October 30, 2017, which covers just over one year, or one entire growing season, for Zambia.

5.1 Spatio-temporal data

As an example of data visualization for the *Rain* questions, we chose five weeks of the study period to illustrate this data signal in the area around the city of Choma. The upper left panel of Figure 10 shows the location within the country, and the other panels, taken together, provide a snapshot of weekly rainfall events in this area between October 2016 and May 2017. Compared to precipitation data coming from the nearest meteorological station in Mochipapa, which represents a huge physical area in terms of meteorological data for the country, the data

from the *Rain* question show how much local variation there is in the region (Figure 11; Southern province only, the portion of the study area associated with Mochipapa station). That is, at first glance it may appear that farmers are reporting rain for the same periods as rainfall recorded at the Mochipapa station, however, during the rainy season (roughly November - March or April), heterogeneity in farmers' responses allows us to identify drier or wetter spots. For example, for the week of March 27, 2017, no rain was recorded at Mochipapa, but around 60% of farmers reported that it rained that week, with drier areas to the north and west of Choma.

As with the rainfall question, spatial aggregation also allows us to extract useful information from the binary responses to the questions related to weekly planting or harvesting events. These three variables--spatially aggregated rainfall, planting, and harvest proportions, can provide valuable information regarding crop management and how it varies in response to regional variations in rainfall onset. This in turn can be estimated by applying change point detection techniques (e.g. a Pettit test; Pettit 1979) to the gridded rainfall proportions, thereby incorporating the fine spatial heterogeneity in rainfall that would be masked if one were to rely on weather station data alone.

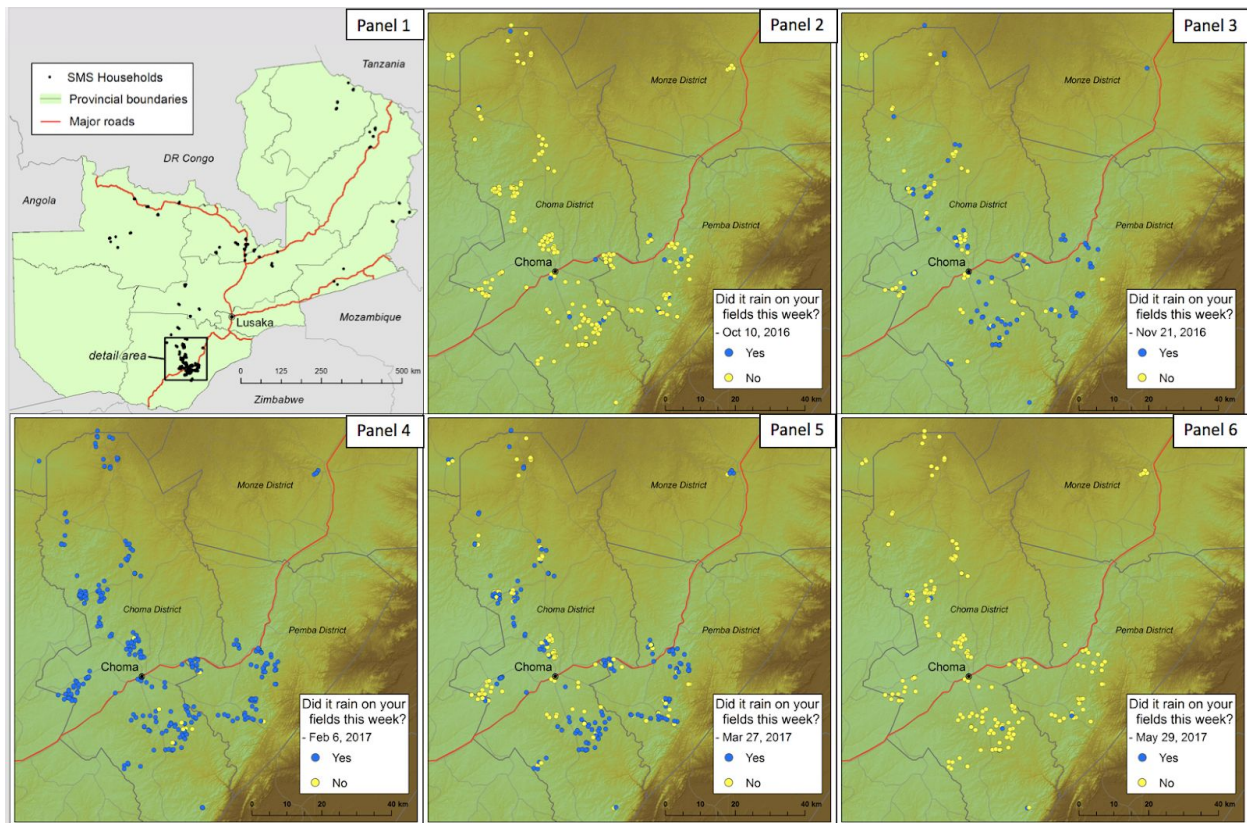


Figure 10. Visualization of data from the *Rain* question over time.

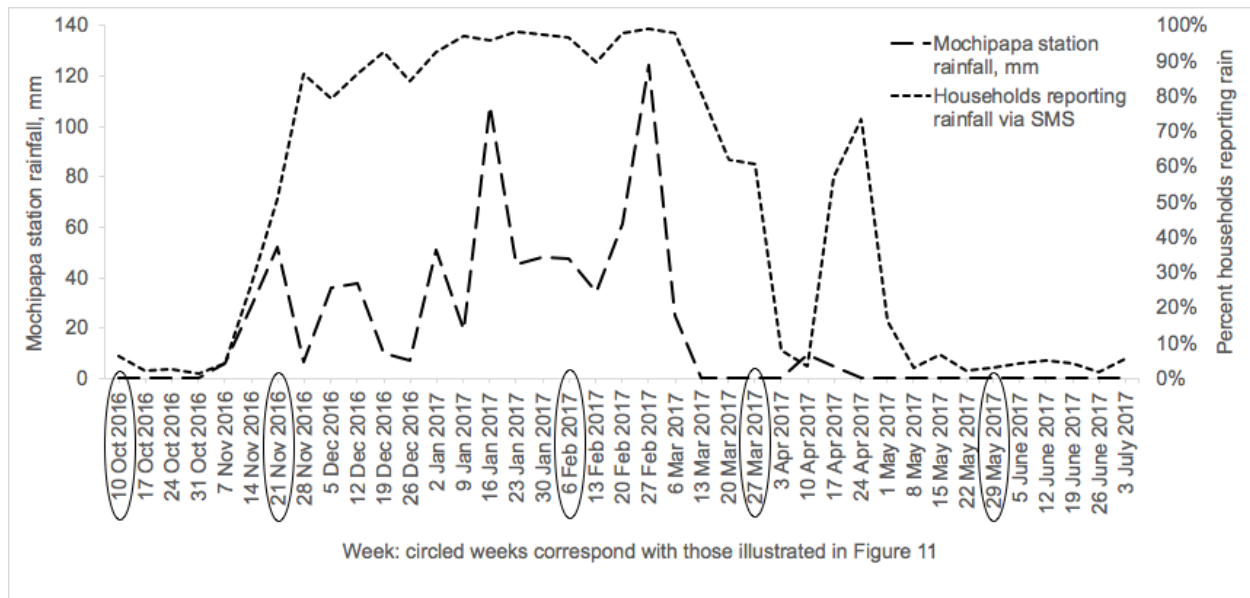


Figure 11. Percent of households in Southern province reporting rainfall plotted against rainfall amounts from Mochipapa station data.

5.2 Temporally linked data

Maize storage in Zambia is crucial, ensuring some degree of food security for smallholder households. With the *Storage* question, even a single visualization of the average number of bags in storage for households over time, broken into tertiles (Figure 12), is a powerful and useful measure of general food security among households in the study area. We first removed outlier responses for each week, calculated as those falling outside the 1.5 interquartile range. The percent of outlier responses each week ranged from zero, during a week with few responses, to 16.4, which was a week during the height of harvest season. In Figure 12, which shows the number of bags of maize households have in storage on average each month over the period of decline during the growing season, it is immediately apparent that some households are far less secure than others, and because of the temporal nature of the data, we can identify the period when storage reserves become critically low. Annual per capita maize consumption has been estimated at anywhere between 105 kg and 175 kg (Hotz et al. 2011, Kumar 1994, Prasanna 2016, Shiferaw et al. 2016).

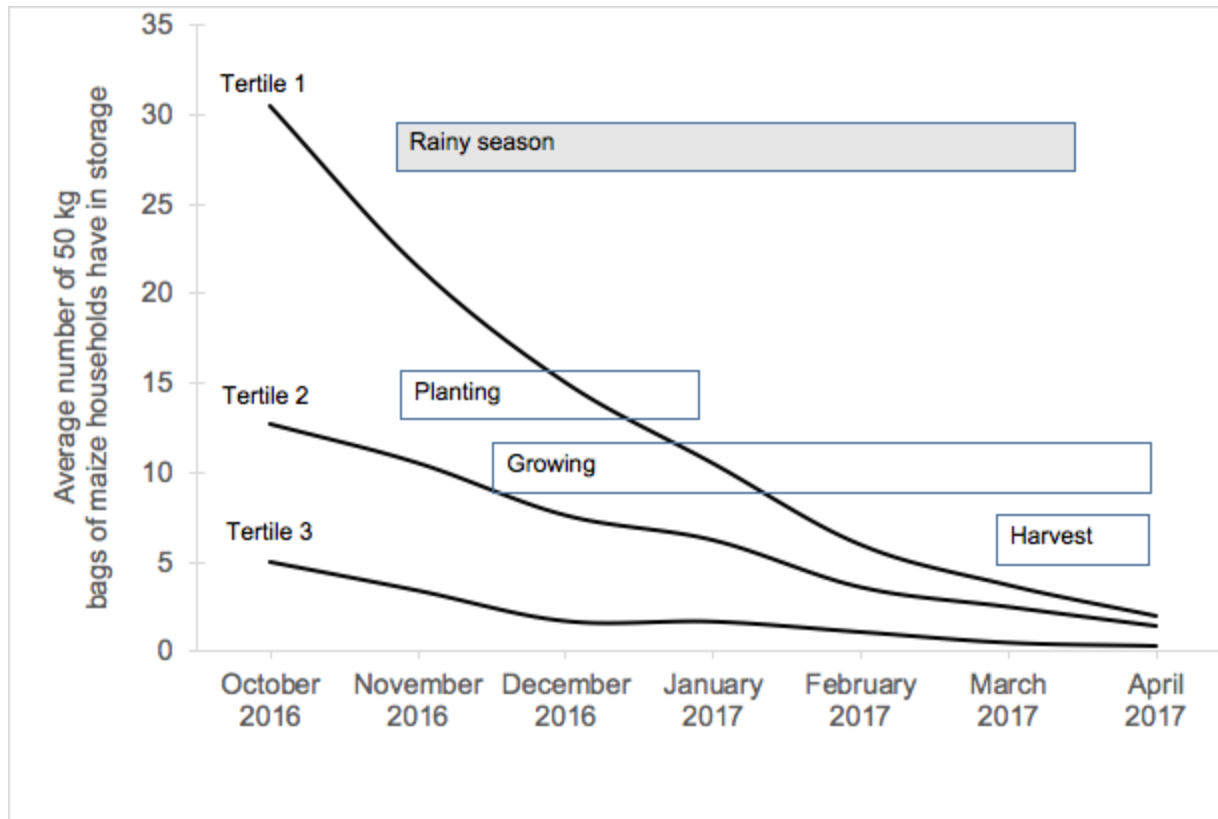


Figure 12. Average number of bags in storage for households over time, broken into tertiles, with indicators for the stages of the maize season and rainy season. Households responded to the question, How many 50 kg bags of maize do you have in storage now?

If we want to know more about the characteristics of households that are more likely to be food insecure, we can look at the underlying attributes of households at each food storage level, which are gathered via an annual survey from a subset of households in the study area. We have survey data for 340 households who answered the *Storage* question and after removal of outliers still had responses for at least one week over the period under consideration. One descriptive analytical method to explore these data is a recursive partitioning technique known as a classification or regression tree (CART). This method helps us explore the structure of the data and provides a visualization of the decision rules used to predict, in this case, the categorical outcome of food storage tertile. One may also, of course, use regression modeling for these data, but our intent here is to show an additional example of how to analyze and visualize these data. We use this method (R package rpart: R Development Core Team 2011; Therneau et al. 2012) to understand which variables might be most useful for predicting food storage from among a set of possible pertinent variables from the household survey: whether

the household's primary language is Tonga (the dominant ethnic group in this region); age of household head; sex of household head; highest education level achieved by anyone living in the household; number of people living in the household; the household's total maize plot area (ha); household's total off-farm income; number of food security measures employed in the week prior (from the Household Food Insecurity Access Scale, or HFIAS (Coates et al. 2007)); did the household give away any maize from their harvest. (Figure 13). To interpret Figure 13, for example, the node that is in the lowest left of the figure can be understood first to mean that 36% of all farmers (n=122) have total maize plot area of less than 1.2 hectares. Then, of those 122, 52% did in fact fall in the lowest tertile of food storage, 30% were in the middle tertile, and 18% were in the highest. Thus based on maize plot size alone (<1.2 hectares), about half of the farmers (52% of 122) would be correctly predicted to be in the lowest tertile group.

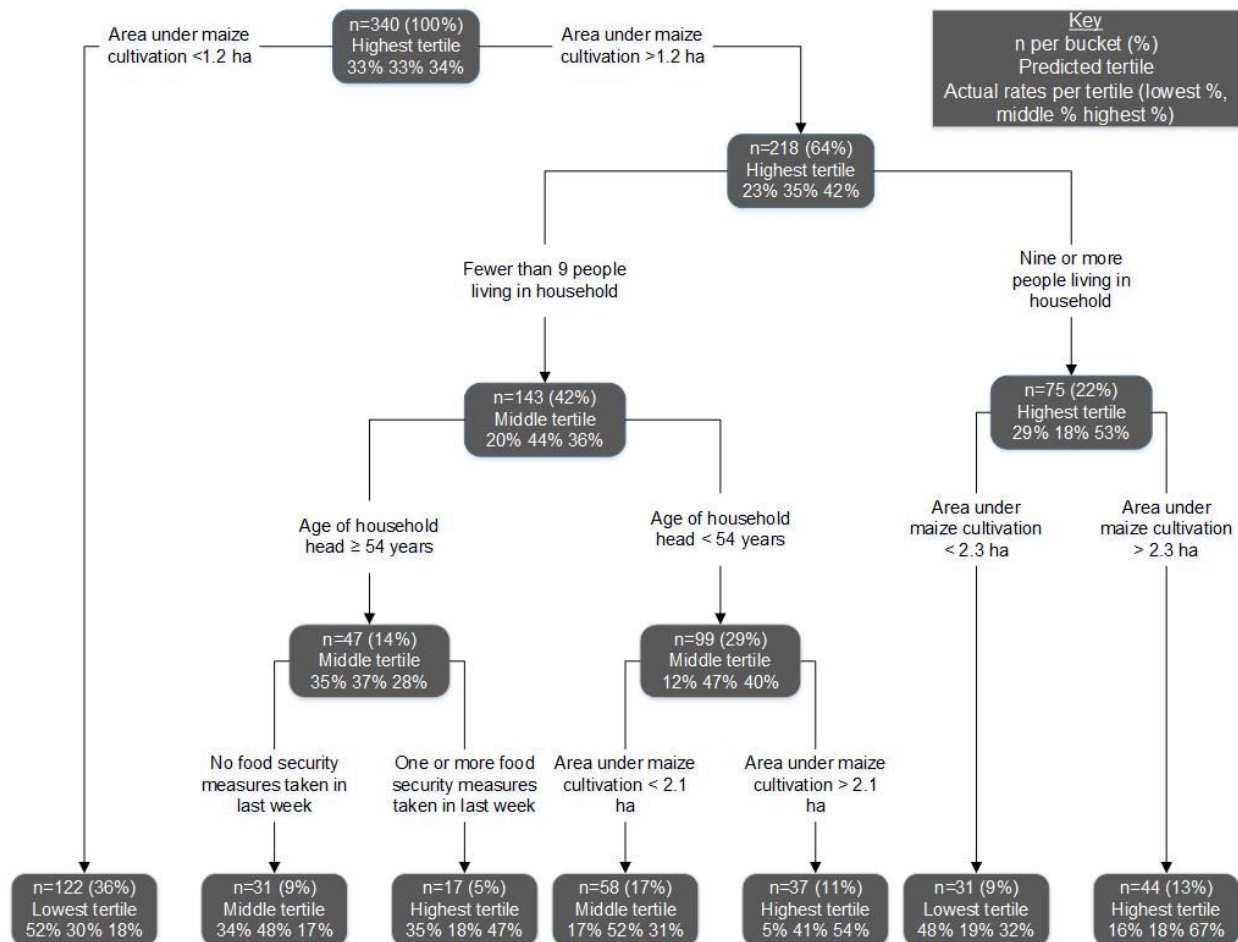


Figure 13. Classification tree for data from the *Storage* question. This tree was selected by pruning to a minimum bucket size of 15 farmers per node.

5.3 Event-based data

One limitation with our SMS-administered data collection is the impact that nonresponse has on event-based data. A typical household may buy or sell maize 3 or 4 times/year. For a question like “Did you sell any maize this week?” a missing response for a week when a household sold maize may mean that the SMS-based data reporting misses a major selling event. However, aggregation of data across a population of households can still be used to portray the seasonality of maize buying or selling through the year in a similar way as that described with the precipitation reporting: missing data from some respondents does not greatly impact the overall trend in the data provided there is a sufficient number of respondents in a particular week. But if the objective is to use SMS based data collection to document what proportion of harvest a household buys or sells through the year, a missed observation can lead to significant misrepresentation.

Households’ behavior in terms of maize selling and buying can be affected by many factors. Households have more maize available to sell after harvest, but they may retain a portion of harvest in hopes that prices increase as available maize stocks become scarcer over time. A household may also sell a large portion of harvest at one time in order to acquire cash for unexpected life events like health related expenses, school fees, or purchases of farm equipment. However, this one-time cash infusion may come at the expense of having to buy maize later in the year to satisfy household food demand. We cannot infer these household-level dynamics from the broad scale SMS data we gather on maize selling and buying, but any buying or selling event with this kind of temporal distribution (i.e., tied to maize farming cycles) will demonstrate some patterns. We can better contextualize such patterns given other macro-level data.

For example, our SMS data (Figure 14) show the temporal frequency of maize selling over the course of 24 weeks during and after harvest, from the week of May 22, 2017 (maize marketing season begins in May) through October 30, 2017. Looking at the trends, we might have expected to see a larger or more defined bump in selling around July, when farmers have finished harvesting maize. Instead, we see that the percentage of households selling maize ranges between about 12% and 26% across these months, with no very clear jump or fall in selling activity. When we consider other data to help understand this pattern, we see that prices for maize began to fall sharply and dramatically after May 2017, back to the five-year average for 2011-2016 (FAO 2018; FEWS NET 2017). A confluence of events across the 2015-2016 and 2016-2017 growing seasons led to this price drop (Chapoto et al. 2017). In the 2015-2016

season, there was both an export ban on maize and a favorable growing season that led to an excess of maize in storage in the country for that year. Then, in the 2016-2017 growing season, farmers experienced another good growing season and produced a bumper harvest. Prices did not pick back up through at least October for most of country (FEWS NET 2017). Given this context, it is less surprising to see the percent of households selling maize varying within this range, with few distinct peaks or valleys in the trendline, because the sudden price drop may have induced more farmers to hold on to their maize to wait and see if prices improve. If we were to examine other years of SMS data and compare patterns (data not shown), we could extend our understanding of the patterns of temporality of maize selling for farmers in Zambia.

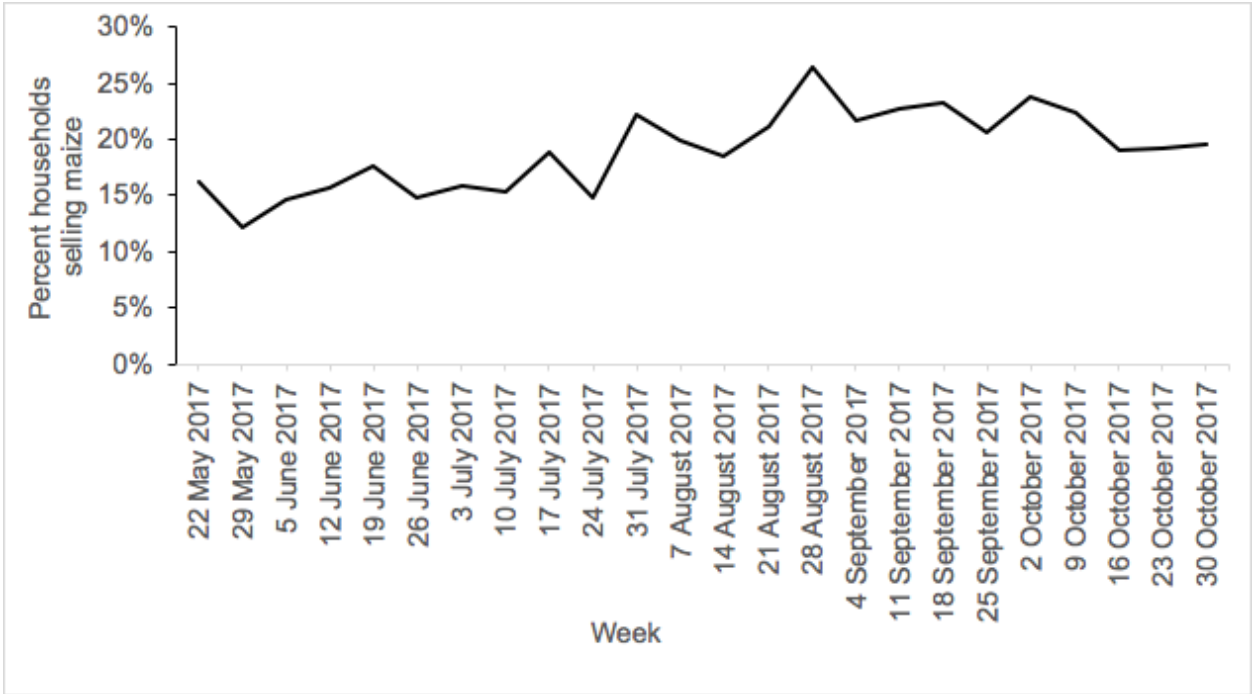


Figure 14. Percent of households reporting that they sold maize each week.

The potential for missing event-based data also limits our ability to compare households, but can be partly remedied by turning to other data gathered via SMS in combination with our survey data. For example, we asked about *Maize buying* because it is a coping mechanism households might employ when their levels of maize in storage are depleted or low. Because we also have data from the *Storage* question, we can identify the period during which certain households find themselves on the margins of food security and thus might be engaging in buying maize to get through their leanest period. We can also identify which households fall into the lowest tertile of food security, and could link them with their responses to the *Maize buying* question to see whether these households are employing this strategy. Further, because we

have survey data for a subset of the SMS participating households, we are able to understand, albeit with less temporal precision, what other food-related coping mechanisms households used during the year. The data from *Maize buying* give us an overall sense of the periodicity and distribution of this event, while the survey data can help us understand why a particular household might be buying maize.

Such frequent, event-based data are some of the most often missed data in traditional household surveys, but as is evident, they remain some of the most difficult to capture well. As our SMS survey program continues, cellular infrastructure becomes more reliable, and we improve our understanding of activities like maize buying and selling, this high-frequency, event-based data collection work will become more and more valuable.

6. Summary and conclusions

Noël and Cai (2017) noted that for models of coupled human and natural systems, environmental models have become increasingly complex as data improves and methods advance, while accurately modeling human behavior within environmental systems remains elusive. Bell et al. (2016) have recognized the need for intra-annual or intra-seasonal social data to capture the complexity of change over time and improve assessment and modeling. When attempting to track and measure human activity and needs related to issues such as farming practices and food security, more traditional methods of data collection such as household surveys fail to capture short-term variability in environmental conditions and behavior (Bell et al. 2016). Further, the temporal mismatch between annually collected household survey data and higher frequency environmental data (e.g. weather) limit the potential of both types of data. Harmonizing those temporal scales allows us to answer questions that we otherwise are not able to address. For example, if we were to only use an annual survey to ask about current maize in storage, depending on when that survey is administered, there could be little variability between households, which might lead us to believe that households in the study area are similar in terms of food security. However, when we examine food storage over the course of the season as we have done here, we see much greater differences in food security for households, which in turn is more meaningful if one were to factor in other survey data, such as HFIAS scores.

We locate our research among work done by Bell et al. (2016), who provided smartphones to rural Bangladeshis to collect real-time survey data via a custom interface, and efforts by applied researchers, such as those being undertaken by the WFP in their mVAM program, which utilizes multiple mobile data collection modes to monitor individuals' current food

security status in many places in the world. Like these, we are collecting social data in real-time when we ask farmers about activities like buying and selling maize. Unlike Bell et al. (2016), we utilize existing data infrastructure in that farmers we survey are using their own phones. We also integrate high-frequency environmental data by asking farmers about rainfall and via mobile meteorological stations we have set up in the study areas. The environmental data we collect addresses gaps found in existing social-environmental data in the area: our weekly question about rainfall, linked to household location and then aggregated across provinces, is a level of detail missing from precipitation data that are currently available for SSA. As a result of all this, we see our work as integrative across scientific disciplines and pertinent to sectors outside of academia.

There are two separate but related issues that our methods to collect, manage, visualize, and synthesize the high-frequency data address. The first is the need for improved agricultural statistics, especially about farmers living in areas where funding, personnel, and infrastructure present barriers to collecting and synthesizing agricultural data. Carletto et al. (2015) note that improvements to methods for collecting data about smallholders have been particularly slow. They also highlight the need to harness technological advances to collect such data more efficiently, and to better integrate agricultural data with other types of data. We are working toward all three of these goals. The second is the issue of scale mismatch. Scale mismatch occurs when the scale of management and scale of ecosystem processes do not match, leading to problems in the system in the institutions charged with managing the system, or in the ecological systems themselves (Cumming et al. 2006). Matched scales, however, are essential to characterizing the resilience of a system and to the ability of those within the system to successfully manage change. Both ecological and social change contribute to scale mismatch, and improving ways to measure and ameliorate scale mismatch is urgent, as poor socio-ecological data can lead to poor policies, institutions, and management. There is no clear path for fixing scale mismatches (Cumming et al. 2013), but more closely matching the data in terms of temporal resolution will help. Our methods are specifically designed to address the observational scale mismatch between annual surveys and farmer decision-making processes. The high frequency with which we are collecting these data can provide more detailed statistics, and temporal data can be put in the context of larger forces. Our example of examining the SMS-gathered maize selling data in light of movement in the market and growing conditions allows us to see how these outside forces are affecting households over time.

Despite the potential of our approach to reduce scale mismatch in the characterization of coupled socio-ecological systems by collecting high-frequency data, bringing this data collection effort to scale is not without its own hurdles. While it is far less expensive to capture these data via SMS compared to in-person surveys, in terms of database infrastructure and cleaning steps, it is costly to prepare it for widespread use. In addition, understanding the types of data gaps that are inherent in high-frequency data and having ways to manage them are necessary to use the data to their fullest potential. We have noted some of these data gaps and presented ideas for addressing or rectifying them. As cellular networks expand, more people acquire cell phones, and smartphone use becomes more widespread, the opportunities for collecting data with greater frequency will continue to improve, both in terms of lowering the rates of technology-related nonresponse by those living in less well-connected places, and in the ways we can utilize these tools to expand and improve data collection. Building tools and knowledge about the process of collecting high-frequency data now will be invaluable as technology presents these new opportunities.

This work was supported by the National Science Foundation [award numbers SES-1360463, SES-1534544, and BCS-1115009] and the NASA New Investigator Program [award number NNX15AC64G].

References

- Aker, J. C., & Ksoll, C. (2016). Can mobile phones improve agricultural outcomes? Evidence from a randomized experiment in Niger. *Food Policy*, *60*, 44-51.
- Bauer, J. M., Akakpo, K., Enlund, M., & Passeri, S. (2013). Tracking vulnerability in real time: Mobile text for food security surveys in Eastern Democratic Republic of Congo. *Africa Policy Journal*, *9*, 36.
- Beegle, K., Carletto, C., & Himelein, K. (2012). Reliability of recall in agricultural data. *Journal of Development Economics*, *98*, 1, 34-41.
- Bell, A. R., Ward, P. S., Killilea, M. E., & Tamal, M. E. H. (2016). Real-time social data collection in rural Bangladesh via a 'microtasks for micropayments' platform on android smartphones. *PLoS ONE*, *11*, 11, e0165924.
- Bogen, K. (1996). The effect of questionnaire length on response rates: a review of the literature. In *Proceedings of the Section on Survey Research Methods* (pp. 1020-1025). American Statistical Association, Alexandria, VA.

Burney, J. A., Naylor, R. L., & Postel, S. L. (2013). The case for distributed irrigation as a development priority in Sub-Saharan Africa. *Proceedings of the National Academy of Sciences*, 110, 31, 12513–12517. doi: 10.1073/pnas.1203597110.

Burnham, M., & Ma, Z. (2016). Linking smallholder farmer climate change adaptation decisions to development. *Climate and Development*, 8(4), 289-311.

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage.

Carletto, C., Jolliffe, D., & Banerjee, R. (2013). The emperor has no data! Agricultural statistics in Sub-Saharan Africa. *World Bank Working Paper*.

Carletto, C., Jolliffe, D., & Banerjee, R. (2015). From tragedy to renaissance: improving agricultural data for better policies. *The Journal of Development Studies*, 51(2), 133-148.

Chapoto, A., Chisanga, B., & Kabisa, M. (2017). Zambia Agricultural Status Report 2017. Indaba Agricultural Policy Research Institute. Available at https://www.researchgate.net/publication/322676437_Zambia_Agriculture_Status_Report_2017. Accessed December 10, 2018.

Cinnamon, J., Jones, S. K., & Adger, W. N. (2016). Evidence and future potential of mobile phone data for disease disaster management. *Geoforum*, 75, 253-264.

Coates, J., Swindale, A. & Bilinsky, P. (2007). Household Food Insecurity Access Scale (HFIAS) for measurement of household food access: Indicator guide (v. 3). Washington, D.C.: FHI 360/FANTA.

Cumming, G. S., D. H. M. Cumming, and C. L. Redman. (2006). Scale mismatches in social-ecological systems: causes, consequences, and solutions. *Ecology and Society* 11, 1, 14. [online] URL: <http://www.ecologyandsociety.org/vol11/iss1/art14/>

Cumming, G. S., Olsson, P., Chapin, F. S., & Holling, C. S. (2013). Resilience, experimentation, and scale mismatches in social-ecological landscapes. *Landscape Ecology*, 28(6), 1139-1150.

Debats, S. R., Luo, D., Estes, L., Fuchs, T. J., & Caylor, K. K. (2016). A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes." *Remote Sensing of Environment*, 179, 210-21. <https://doi.org/10.1016/j.rse.2016.03.010>.

de Lepper, A. M., Eijkemans, M. J., Beijma, H., Loggers, J. W., Tuijn, C. J., & Oskam, L. (2013). Response patterns to interactive SMS health education quizzes at two sites in Uganda: a cohort study. *Tropical Medicine & International Health*, 18(4), 516-521.

de Tolly, K., Skinner, D., Nembaware, V., & Benjamin, P. (2012). Investigation into the use of short message services to expand uptake of human immunodeficiency virus testing, and whether content and dosage have impact. *Telemedicine and e-Health*, 18(1), 18-23.

Ericsson. (2014). Ericsson Mobility Report, Sub-Saharan Africa.

Estes, L. D., Chaney, N. W., Herrera-Estrada, J., Sheffield, J., Caylor, K. K., & Wood, E. F. (2014). Changing water availability during the African maize-growing season, 1979–2010. *Environmental Research Letters*, 9(7), 075005.

Estes, L. D., McRitchie, D., Choi, J., Debats, S., Evans, T., Guthe, W., ... & Caylor, K. K. (2016). A platform for crowdsourcing the creation of representative, accurate landcover maps. *Environmental Modelling & Software*, 80, 41-53.

FAO (2013). Smallholders and Family Farmers. Factsheet. <http://www.fao.org/3/a-ar588e.pdf>. Accessed January 17, 2018.

FAO (2018). Global information and early warning system on food and agriculture (GIEWS) country brief, Zambia. <http://www.fao.org/giews/countrybrief/country/ZMB/pdf/ZMB.pdf>. Accessed December 10, 2018.

FEWS NET, (2017). Zambia price bulletin, November 2017. Downloaded from <http://fewnet.net/southern-africa/zambia/price-bulletin/november-2017> December 10, 2018.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., ... & Obersteiner, M. (2012). Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31, 110-123.

Garlick, R., Orkin, K., & Quinn, S. (2016). Call me maybe: experimental evidence on using mobile phones to survey African microenterprises. *Economic Research Initiatives at Duke (ERID) Working Paper*, (224).

Harmer, N., & Rahman, S. (2014). Climate change response at the farm level: a review of farmers' awareness and adaptation strategies in developing countries. *Geography Compass*, 8(11), 808-822.

Hoogeveen, J., Croke, K., Dabalén, A., Demombynes, G., & Giugale, M. (2014). Collecting high frequency panel data in Africa using mobile phone interviews. *Canadian Journal of Development Studies/Revue canadienne d'études du développement*, 35(1), 186-207.

Horsburgh, J. S., Leonardo, M. E., Abdallah, A. M., & Rosenberg, D. E. (2017). Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system. *Environmental Modelling & Software*, 96, 83-94.

Hotz, C, Palaniappan, U, Chileshe, J, Kafwembe, E, & Siamusantu, W. (2011). Nutrition Survey in Central and Eastern Provinces, Zambia, 2009: Focus on Vitamin A and Maize Intakes, and Vitamin A Status among Women and Children. Lusaka, Zambia, Washington, DC, and Ndola,

Zambia: National Food and Nutrition Commission, HarvestPlus, and Tropical Diseases Research Centre.

Kadlec, J., & Ames, D. P. (2017). Using crowdsourced and weather station data to fill cloud gaps in MODIS snow cover datasets. *Environmental Modelling & Software*, 95, 258-270.

Kotir, J. H. (2011). Climate change and variability in Sub-Saharan Africa: a review of current and future trends and impacts on agriculture and food security. *Environment, Development and Sustainability*, 13(3), 587-605.

Kumar, S. K. (1994). Adoption of hybrid maize in Zambia: Effects on gender roles, food consumption, and nutrition (Vol. 100). Washington, DC: International Food Policy Research Institute.

Leo, B., Morello, R., Mellon, J., Peixoto, T., & Davenport, S. T. (2015). Do mobile phone surveys work in poor countries? Center for Global Development Working Paper 398.

Lowder, S. K., Skoet, J., & Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, 87, 16-29.

Lynn, P. (2014). Longer interviews may not affect subsequent survey participation propensity. *Public Opinion Quarterly*, 78(2), 500-509.

Mendelsohn, R. (2008). The impact of climate change on agriculture in developing countries. *Journal of Natural Resources Policy Research*, 1(1), 5-19.

Mock, N., Singhal, G., Olander, W., Pasquier, J. B., & Morrow, N. (2016). mVAM: A new contribution to the information ecology of humanitarian work. *Procedia Engineering*, 159, 217-221.

Morrow, N., Mock, N., Bauer, J. M., & Browning, J. (2016). Knowing Just in Time: Use cases for mobile surveys in the humanitarian world. *Procedia Engineering*, 159, 210-216.

Mwingira, U. J., Downs, P., Uisso, C., Chikawe, M., Sauvage-Mar, M., Malecela, M. N. Crowley, K. & Ngondi, J. M. (2017). Applying a mobile survey tool for assessing lymphatic filariasis morbidity in Mtwara Municipal Council of Tanzania. *mHealth*, 3.

National Science Foundation. (2007). Cyberinfrastructure vision for the 21st century discovery. Retrieved from <https://www.nsf.gov/pubs/2007/nsf0728/index.jsp>. Accessed November 25, 2017.

Nayak, A., Poriya, A., & Poojary, D. (2013). Type of NOSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems* 5(4), 16-19.

Noël, P. H., & Cai, X. (2017). On the role of individuals in models of coupled human and natural systems: Lessons from a case study in the Republican River Basin. *Environmental Modelling & Software*, 92, 1-16.

Pettitt, A. N. "A Non-Parametric Approach to the Change-Point Problem." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, no. 2 (January 1, 1979): 126–35.
<https://doi.org/10.2307/2346729>.

Pew Research Center. (2015). Cell Phones in Africa: Communication Lifeline.

Pew Research Center. (2018). Internet Connectivity Seen as Having Positive Impact on Life in Sub-Saharan Africa.

Plale, B., & Kouper, I. (2017). The centrality of data: Data lifecycle and data pipelines. In M. A. Chowdhury, A. Apon, & K. Dey (Eds.), *Data analytics for intelligent transportation systems* (pp. 91–111). Cambridge, MA: Elsevier Inc. doi: 10.1016/B978-0-12-809715-1.00004-3

Pop-Eleches, C., Thirumurthy, H., Habyarimana, J. P., Zivin, J. G., Goldstein, M. P., De Walque, D., ... & Ngare, D. (2011). Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. *AIDS (London, England)*, 25(6), 825.

Prasanna, B. M. (2016). Developing and deploying abiotic stress-tolerant maize varieties in the tropics: challenges and opportunities. In V. R. Rajpal, S. R. Rao, & S. N. Raina (Eds.), *Molecular Breeding for Sustainable Crop Improvement* (pp. 61-77). Springer.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2011. <http://www.R-project.org/>.

Raimi, K. T., & Carrico, A. R. (2016). Understanding and beliefs about smart energy technology. *Energy Research & Social Science*, 12, 68-74.

Shiferaw, B., Prasanna, B. M., Hellin, J., & Bänziger, M. (2011). Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security*, 3(3), 307-327.

TextIt Ingestor. (n.d.). In Data To Insight Center Github. Retrieved from <https://github.com/Data-to-Insight-Center/smallholder-ag/tree/master/textit>.

The World Bank. (2016). Mobile cellular subscriptions (per 100 people), Sub-Saharan Africa. <https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=ZG>. Accessed January 17, 2018.

Therneau, T., Atkinson, B., Ripley, B. (2012). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Waldman, K. B., Vergopolan, N., Estes, L. D., Attari, S. Z, Sheffield, J. Caylor, K. K., and Evans, T. P. (forthcoming) Cognitive biases about climate variability in smallholder farming systems in Zambia. *Climatic Change*.

Yu, Q., Shi, Y., Tang, H., Yang, P., Xie, A., Liu, B., & Wu, W. (2017). eFarm: A Tool for Better Observing Agricultural Land Systems. *Sensors*, 17(3), 453.

Appendix A.

These are questions asked of farmers in the SMS data collection program in Zambia. This is not an exhaustive list of all questions ever asked of these farmers, because questions have evolved or been dropped over time depending on their performance and the research foci of team members.

Question	Question nickname
Have you planted any maize in the last 7 days?	maize planting
How many seed varieties did you plant?	variety number
What is the first seed variety you planted?	variety 1 of 2
How many kg of [that variety] did you plant?	kg 1 of 2
What is the second seed variety you planted?	variety 2 of 2
How many kg of [that variety] did you plant?	kg 2 of 2
What seed variety did you plant?	variety 1
How many kg of [that variety] did you plant?	kg 1
How many 50kg bags of maize do you have in storage now?	storage
Are all of your maize fields planted now?	planting complete
Did it rain on your fields in the last 7 days?	rain
Did you weed your maize fields in the last two weeks?	weeding
Did you apply fertilizer to your maize fields in the last two weeks?	fertilizer
In the last two weeks, how many days did you work outside your	piecework

farm for pay?

At this point, how many 50kg bags do you expect to harvest at the end of the season from all of your maize fields? expected harvest

How many hours did your household spend collecting water in the last 7 days? water collection

How many hours did your household spend collecting firewood in the last 7 days? firewood

Did your household use charcoal for cooking in the last 7 days? charcoal

Have you harvested any of your maize fields in the last 7 days? maize harvest

How many 50kg bags of maize have you harvested from your maize fields in the last 7 days? 50kg harvest

Have you harvested all of your maize fields? harvest complete

Did you buy any maize for your household usage this week? maize buying

How much maize did you purchase? maize purchased

Did you sell any maize this week? maize selling

How much maize did you sell? maize sold

In the past 14 days, did anyone give you maize so your household would have enough food? receive maize

In the past 14 days did you give maize or mealie meal to any neighbors to they would have enough food? give maize

In the past 14 days how many days did your household consume meat? meat