Policy Research Working Paper 8138

# Mission Impossible?

## Exploring the Promise of Multiple Imputation for Predicting Missing GPS-Based Land Area Measures in Household Surveys

*Talip Kilic*
*Ismael Yacoubou Djima*
*Calogero Carletto*

WORLD BANK GROUP
Development Economics
Development Data Group
July 2017

Policy Research Working Paper 8138


## Abstract

Methodological research has showcased GPS technology as the new gold-standard in land area measurement in large-scale household surveys. Nonetheless, facing budget constraints, survey agencies continue to measure with GPS only plots within sampled enumeration areas or a given radius of dwelling locations. It is, subsequently, common for significant shares of plots not to be measured, and research has demonstrated that the incomplete datasets are subject to selection bias. This study relies on nationally-representative survey data from Malawi and Ethiopia that exhibit near-negligible missingness in GPS-based plot areas and uses these datasets to gauge the limits to the accuracy of a Multiple Imputation (MI) application for predicting GPS-based areas for plots that would typically be considered out-of-scope. The analysis (i) artificially creates missingness in area measures, ranging from 1 to 100 percent, among the plots that are beyond two operationally-relevant distance thresholds with respect to the dwellings; (ii) multiply-imputes "missing" values in each dataset created by a distance threshold-missingness combination; and (iii) compares the distributions of the imputed plot-level outcomes with the distributions of their true, observed counterparts. In Malawi, the multiply-imputed distribution of plot-level land productivity is statistically indistinguishable from the true distribution in each imputed dataset with up to 82 percent missingness in GPS-based plot areas that are more than 1 kilometer away from the associated dwellings. The comparable figure in Ethiopia is 56 percent. The study highlights the promise of MI for simulating missing area measures and provides recommendations for optimizing fieldwork to capture the minimum required data.


This paper is a product of the he Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at tkilic@worldbank.org.

# Mission Impossible?

## Exploring the Promise of MI for Predicting Missing GPS-Based Land Area Measures in Household Surveys

Talip Kilic[†], Ismael Yacoubou Djima[†], and Calogero Carletto[†1]

**JEL Codes:** C53, C83, Q12, Q15.

**Keywords:** Global Positioning System, Land Area Measurement, Productivity Analysis, Missing Data, Multiple Imputation, Household Surveys.

[1†] Development Data Group, World Bank. The senior authorship is shared between Talip Kilic and Ismael Yacoubou Djima, who is the corresponding author. The email addresses for the authors are tkilic@worldbank.org, iyacouboudjima@worldbank.org, and gcarletto@worldbank.org. The authors thank (in alphabetical order) Tomoki Fujii and Alberto Zezza for their comments on the earlier versions of this paper, and Heather Moylan for her insights on the field operations of the Malawi Integrated Household Survey. This work was supported by the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) initiative, funded by the Bill and Melinda Gates Foundation, and the Minding the (Agricultural) Data Gap Methodological Research program, funded by UKAid.

# 1 Introduction

Land is a fundamental component of household and personal wealth in rural areas and is the key factor of production in smallholder production systems. As such, the data on parcel and cultivated plot areas are the heart of economic research linked to agriculture and the design and implementation of land registration, titling and redistribution programs. Furthermore, the Sustainable Development Goal (SDG) Target 2.3 require doubling of agricultural productivity and incomes of small-scale food producers by 2030, and the monitoring of the progress towards this target rely on land area information sourced from household or farm surveys.

While large-scale household and farm surveys in low- and middle-income countries have traditionally relied on farmer reporting to elicit information on land areas, this can be problematic, particularly in the African context, which is characterized by the high incidence of smallholder farming and the fragmentation of farms into multiple parcels with irregular shapes and without formal titles. Several reasons may contribute to the inaccuracy in self-reported land areas. First, farmers may knowingly overstate or understate their landholdings for strategic reasons that may relate to access to development programs and/or taxation. Second, there is a natural tendency to round off numbers and provide approximations, which leads to heaping of the data around discrete values. Third, geography, particularly slope, can influence the way farmers assess distance and area. Fourth, the use of non-standard measurement units and within-country variation in the type and standard unit equivalence of these units complicate the compilation of conversion factors for land area measurement. In fact, methodological research has shown that self-reported land areas are subject to systematic measurement error with direct implications for the accurate measurement and analysis of land productivity (Carletto, et al., 2013, 2015).

These reasons, combined with (i) the validated accuracy of GPS-based land area measurement in household survey experiments in Ethiopia, Nigeria, and Tanzania (Zanzibar) (Carletto, et al., 2017), and (ii) the ever-

increasing affordability and accuracy of handheld GPS devices makes GPS-based land area measurement a desirable alternative for household and farm surveys in countries dominated by smallholder agricultural production. However, with the emergence of GPS-based area measurement as the new, scalable gold-standard for household and farm surveys, a key drawback is related to the operationalization of the technology.

To reduce transportation costs, keep household interview durations within reasonable limits, and avoid the difficulty of asking respondents to accompany enumerators to agricultural plots that are situated far from dwelling locations, survey implementing agencies often require enumerators to only obtain GPS-based area measures for plots within a given (arbitrary and non-cross-country-comparable) radius of dwelling locations. Consequently, non-ignorable shares of area measures are missing in public use datasets. For instance, among the selected national, multi-topic panel household surveys that are supported by the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) program, the rate of missingness in GPS-based plot areas range from 13 (Nigeria) to 44 percent (Uganda), as shown in Table 1. Given the potential selection biases that may be brought on by analyzing only non-missing portions of the datasets, there is a concern that the missing data may limit the operational relevance and the analytical value of GPS-based area measures. And this concern is pressing particularly in the context of a survey program such as the LSMS-ISA that has catalyzed a significant expansion in development research on Africa over the last decade.[2]

[2] As of November 12, 2018, the official World Bank Microdata Library download count for the publicly available, LSMS-ISA-supported household surveys stood at 37,750, and the lower-bound for the number of research outputs over the last decade based on the LSMS-ISA data, according to the continuing LSMS monitoring of online development research outlets, is estimated at 1,000.

*Table 1: Rates of Missingness in GPS-Based Plot Areas in Selected Datasets Generated by the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS ISA) & Survey Instructions on the Required Spatial Coverage of GPS-Based Plot Area Measurements*

| ***Survey*** | ***Rate of Missingness*** | ***Required Spatial Coverage of GPS-Based Plot Area Measurements*** |
|---|---|---|
| Niger Enquete Nationale sur les Conditions de Vie des Menages et l'Agriculture 2011 | 29% | Measure all plots in the same enumeration area as the household. |
| Nigeria General Household Survey - Panel 2012/2013 | 13% | Measure all plots in the same district of the household and within 3 hours of travel, regardless of mode of transportation. |
| Tanzania National Panel Survey 2010/2011 | 22% | Measure all plots within 1 hour of travel from the household, regardless of mode of transportation. |
| Uganda National Panel Survey 2011/2012 | 44% | Measure all plots in the same enumeration area as the household. |

Recognizing the need to address the problem of missing data for increasing the usability of household survey data, (Kilic, et al., 2017) use the LSMS-ISA data from Tanzania and Uganda to show that the missing GPS-based plot areas indeed constitute a non-random subset of the unit-record data and that the missing data can be simulated by Multiple Imputation (MI). In their analysis of plot-level land productivity, the authors document the non-trivial effects of using the datasets that are completed based on MI.

Underlined by the relatively recent adoption of GPS-based land area measurement in national household surveys, and the evolving appreciation of the recent methodological research on addressing (ultimately unavoidable) missingness in GPS-based land area measures, there is a continuing need to work on two fronts. The first is to offer operational guidance for survey practitioners and data users, including agricultural and development economists, regarding "acceptable" rates of item non-response in GPS-based areas for "distant" plots whose areas could instead be simulated. The second is to further elevate the importance of relying on easily-accessible, model-based simulation approaches, including MI, to judiciously address missingness in public use datasets that are at the core of development research.

To address these needs, we work with the unique, nationally-representative household survey data from Malawi and Ethiopia that exhibit near-negligible rates of missingness in GPS-based plot areas and use these datasets to gauge the validity and accuracy of an MI-based approach to predict missing GPS-based land areas among plots that would otherwise be deemed "distant" in a typical survey operation. In doing so, we test the typically-untestable assumptions of MI and identify the acceptable rates of missingness beyond which these assumptions are less likely to hold, specifically for the reliable estimations of cultivated area and agricultural productivity.

The use of actual data collected as part of large-scale household surveys that have adopted GPS-based area measurement is key to the operational relevance of our research. As such, we provide operational recommendations that can enable survey practitioners, including agricultural and development economists involved in primary data collection, to collect the minimum-required data for model-based imputation applications. In addition, we make available the constructed datasets and syntax files to replicate our analyses; enable future MI applications to address similar missing information problems; and catalyze further research for deriving alternative acceptable rates of missingness that can drive the design and implementation of future survey efforts that may have different analytical objectives than those that we work with.

Our headline finding is that in Malawi, the multiply-imputed distribution of plot-level agricultural productivity is statistically indistinguishable from the *true* distribution in each of the 50 imputed datasets with up to 82 percent missingness in GPS-based plot areas that are more than 1 kilometer away from the associated dwelling. The comparable figure in Ethiopia is 56 percent. If one sets the distance threshold at 500 meters, the tolerate rates of missing GPS-based areas among distant plots stand at 45 percent in Malawi, and 36 percent in Ethiopia. If one focuses on plot area as the outcome variable of interest, as opposed to productivity, the estimated tolerable rates of missingness present an even more optimistic outlook regarding the promise of MI, irrespective of the country and/or the distance threshold in question.

The paper is organized as follows. Section 2 describes the data. Section 3 presents the empirical approach. Section 4 discussed the results. Section 5 concludes, expanding on the relevance of our findings for household and farm surveys that visit sampled households at least twice and in sync with a given agricultural season.

## 2 Data

The Malawi Third Integrated Household Survey 2010/2011 (IHS3), and the Ethiopia Socioeconomic Survey Wave II 2013/2014 (ESS2), which were conducted respectively by the Malawi National Statistical Office (NSO) and the Central Statistics Agency (CSA) of Ethiopia inform our analysis. Both surveys were implemented under the Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) program.

The IHS3 data were collected within a two-stage cluster sampling design, and are representative at the national, urban/rural, regional, and district levels, covering 12,271 households in 768 enumeration areas (EAs). ESS2 is part of a long-term project to collect panel data. It covered all regional states including the capital, Addis Ababa. Much of the sample is comprised of rural areas as it was carried over from ESS1. The survey is representative at the national, urban/rural and, 6 strata (4 regions plus Addis Ababa and the other regions) covering 5,262 households in 433 EAs.

In terms of questionnaire instruments, the IHS3 and the ESS2 both had Household, Agriculture, and Community Questionnaires. In each setting, the sample households were administered a multi-topic Household Questionnaire that collected individual-disaggregated information on demographics, education, health, wage employment, nonfarm enterprises, anthropometrics, and control of income from non-farm income sources, as well as data on housing, food consumption, food and non-food expenditures, food security, and durable and agricultural asset ownership, among other topics. In addition, agricultural

households received the Agriculture Questionnaire, which solicited plot-level information on land areas, manager/holder identification, labor and non-labor input use, and crop cultivation and production.[3] Further, agricultural production data were collected for the two main agricultural seasons in each survey. Handheld global positioning system (GPS)-based locations and land areas of the plots were recorded, permitting us to link household- and plot-level data to outside geographic information system (GIS) databases.

The IHS3 required GPS-based area measurement of all plots that are owned and/or cultivated by the sampled households, within 2 hours of travel with respect to the household location, regardless of mode of transportation. For the distant plots, the field teams were advised to cluster them in accordance with their location, and to visit them in a coordinated fashion by using the team vehicle. For the sub-sample of IHS3 households that were visited twice, the first visit data were also reviewed, and the missing GPS-based plot areas were fed forward to the second visit interviews for potential capture by the field teams. While the first visit constraints leading to missing data still applied to most of these households during the second visit, the continuing emphasis on increasing the volume of GPS-based plot area measures did result in additional data capture. On the other hand, the ESS2 instructed the enumerators to take GPS-based area measures of all plots that are owned and/or cultivated by the sampled households, irrespective of distance. For plots less than 40 square meters, the enumerators measured areas by traversing, instead of GPS units. The overall rates of missingness in GPS-based plot areas were considerably low in both settings: 3.8 percent in Malawi and 6.2 percent in Ethiopia. These are in fact the lowest levels observed among the surveys supported by the LSMS-ISA program.

[3] Both the IHS3 and the ESS2 make a clear distinction between a *parcel* and a *plot*. A parcel is conceptualized as a continuous piece of land under a common tenure system, while a plot is defined as a continuous piece of land on which a unique crop or a mixture of crops is grown, under a uniform, consistent crop management system, not split by a path of more than one meter in width, and with boundaries defined in accordance with the crops grown and the operator. Therefore, a parcel can be made up of one or more plots. This distinction is key since for the purposes of within-farm analysis of agricultural productivity, the ideal is to capture within-parcel, plot area measurements linked with plot-level measurement of agricultural production Parcel-level GPS-based area measurement, on the other hand, could serve other objectives, such as surveying of land for land registration or titling programs or for land ownership measurement. An open empirical question is whether the extent to which parcel-area measurement could be reliably backed from aggregation of within-parcel, plot area measures – an exercise that will be mediated by the precision with which parcel and plot boundaries are established in the field prior to GPS-based area measurement.

Our analysis assumes both data sets to be complete and representative of the true distributions of interest and is subsequently conducted using plots for which GPS based-land area measurements are available.[4] Table 2 shows the distribution of plots according to their distance from the dwelling for both datasets. Table 3 presents the summary statistics based on the IHS3 and the ESS2, including the plot-level means for the entire sample; for the sample within 1 kilometer of the dwelling; and for the sample that lie outside of the 1-kilometer radius of the dwelling. Table 4 accomplishes the same objective but for the samples split by the alternative distance threshold of 500 meters. We provide the differences between the sample means and note when a given mean difference is statistically significant.

Several noteworthy findings emerge from Tables 2, 3 and 4. First, the distribution of plots per distance threshold is quite similar across the two countries. Between 54 and 60 percent of the plots are within 500 meters and between 72 to 77 percent are within 1 kilometer. Second, the plots within the distance threshold tend to be of significantly smaller areas than the plots beyond that threshold. Third, several important plot and household level characteristics which are expected to be associated with productivity-related outcomes, display statistically significant differences by distance threshold status. These observations highlight the importance of systematically addressing missingness in GPS-based plot areas, if such GPS data are to be used in a robust fashion.

[4]We cannot work with approximately 50 percent of the ESS2 plots in the public use data since the CSA ancillary dataset with the conversion factors for the non-standard land area measurement units (to express farmer-reported plot areas in hectares) does not include conversion factors for all non-standard measurement units. This limitation further underscores the importance using GPS-based land area measurements. Going forward, the ESS2 can be used to update the referenced ancillary dataset of conversion factors. Prior to the validation exercise, we elected not to update the ancillary dataset using the ESS2 since the imputation model performance would have improved dramatically in a mechanical manner. Further, most of the predictors that we use in the validation exercise based on the ESS2 data do present statistically significant differences across the plots depending on whether land area conversion factor is available. These predictors are included in the imputation model, and to the extent that they are correlated with observed and unobserved attributes that predict the likelihood of a farmer-reported plot area with a missing conversion factors, the ESS2 sample that we end up focusing on should be deemed satisfactory for validation purposes.

*Table 2: Plot Distribution Based on the Euclidean Distance from Household*

| | Malawi (IHS3) | | | Ethiopia (ESS2) | | |
|---|---|---|---|---|---|---|
| **Distance Interval** | **Frequency** | **Percentage** | **Cumulative Percentage** | **Frequency** | **Percentage** | **Cumulative Percentage** |
| [0.0, 0.5 Km) | 9,798 | 53.67 | 53.67 | 12,282 | 61.51 | 61.51 |
| [0.5, 1.0 Km) | 3,363 | 18.42 | 72.09 | 3,070 | 15.38 | 76.89 |
| [1.0, 2.0 Km) | 2,888 | 15.82 | 87.91 | 2,455 | 12.30 | 89.19 |
| [2.0, 3.0 Km) | 755 | 4.14 | 92.05 | 862 | 4.32 | 93.50 |
| [3.0, 5.0 Km) | 404 | 2.21 | 94.26 | 537 | 2.69 | 96.19 |
| [5.0, 10.0 Km) | 306 | 1.68 | 95.94 | 342 | 1.71 | 97.91 |
| [10.0, ~ Km) | 742 | 4.06 | 100.00 | 418 | 2.09 | 100.00 |
| Total | 18,256 | 100.00 | | 19,966 | 100.00 | |

*Table 3: Selected Plot-Level Means by Plot Distance to Household (Above versus Below 1 Kilometer)*

| | Malawi (IHS3) | | | | Ethiopia (ESS2) | | | |
|---|---|---|---|---|---|---|---|---|
| | **Entire sample** | **Sample [d < 1km]** | **Sample [d > = 1km]** | **x[d <1km] - x[d >= 1km]** | **Entire sample** | **Sample [d < 1km]** | **Sample [d > = 1km]** | **x[d <1km] - x[d >= 1km]** |
| Observations (Plots) | 18,256 | 13,161 | 5,095 | | 19,966 | 15,352 | 4,614 | |
| ***Plot Areas*** | | | | | | | | |
| GPS-based plot area (Ha) | 0.394 | 0.383 | 0.420 | -0.037*** | 0.197 | 0.177 | 0.261 | -0.084*** |
| Farmer-reported plot area (Ha) | 0.414 | 0.403 | 0.440 | -0.036*** | 0.193 | 0.175 | 0.251 | -0.075*** |
| ***Yields*** | | | | | | | | |
| Maize yield (Kg/Ha) | 1,693 | 1,694 | 1,692 | 2 | | | | |
| Value of output/Ha | | | | | 29,303 | 31,575 | 22,447 | 9,128 |
| ***Plot Manager Characteristics*** | | | | | | | | |
| Female † | 0.261 | 0.267 | 0.246 | 0.021 | 0.153 | 0.158 | 0.137 | 0.021 |
| Age (Years) | 43.147 | 43.511 | 42.273 | 1.238*** | 46.817 | 47.220 | 45.472 | 1.748* |
| Education (Years) | 5.028 | 4.934 | 5.252 | -0.318*** | 1.874 | 1.952 | 1.614 | 0.338 |
| ***Household Characteristics*** | | | | | | | | |
| Household size | 4.934 | 4.871 | 5.086 | -0.215*** | 6.476 | 6.491 | 6.427 | 0.064 |
| # of HH members - [0,5] | 0.981 | 0.974 | 0.998 | -0.024 | 0.916 | 0.920 | 0.904 | 0.016 |
| # of HH members - [6,14] | 1.396 | 1.369 | 1.461 | -0.093** | 1.932 | 1.930 | 1.940 | -0.010 |
| # of female HH members - [15,39] | 0.901 | 0.879 | 0.953 | -0.074*** | 1.111 | 1.102 | 1.143 | -0.041 |
| # of male HH members - [15,39] | 0.837 | 0.819 | 0.881 | -0.062** | 1.212 | 1.210 | 1.217 | -0.007 |
| # of female HH members - [40,59] | 0.270 | 0.270 | 0.268 | 0.002 | 0.386 | 0.394 | 0.363 | 0.031 |
| # of male HH members - [40,59] | 0.269 | 0.262 | 0.287 | -0.024 | 0.391 | 0.387 | 0.404 | -0.017 |
| # of HH members – 60 & above | 0.280 | 0.297 | 0.238 | 0.059*** | 0.527 | 0.548 | 0.456 | 0.092* |
| Household consumption expenditures per capita | 50,431 | 48,494 | 55,087 | -6,593*** | 5,723 | 5,804 | 5,453 | 350 |
| Number of plots in the holding | 2.374 | 2.359 | 2.410 | -0.051 | 15.857 | 16.086 | 15.091 | 0.995 |
| ***Plot Characteristics*** | | | | | | | | |
| Owned by household † | 0.904 | 0.917 | 0.872 | 0.045*** | 0.866 | 0.888 | 0.794 | 0.094*** |
| Use of hired labor † | 0.223 | 0.195 | 0.290 | -0.095*** | 0.057 | 0.050 | 0.082 | -0.032*** |
| Use of organic fertilizer † | 0.116 | 0.122 | 0.101 | 0.021*** | 0.183 | 0.213 | 0.085 | 0.128*** |
| Use of inorganic fertilizer † | 0.618 | 0.623 | 0.607 | 0.016 | 0.404 | 0.415 | 0.369 | 0.047 |
| Irrigated † | 0.005 | 0.005 | 0.006 | -0.001 | 0.016 | 0.017 | 0.011 | 0.006 |
| Soil quality good † | 0.467 | 0.453 | 0.503 | -0.050*** | 0.327 | 0.329 | 0.319 | 0.010 |
| Soil quality poor † | 0.113 | 0.112 | 0.116 | -0.004 | 0.173 | 0.168 | 0.188 | -0.019 |

**Note:** † denotes a dummy variable. *** p<0.01, ** p<0.05, * p<0.1. Sample of plots within a 1-kilometer radius is the comparison group for the tests of mean differences.

*Table 4: Selected Plot-Level Means by Plot Distance to Household (Above versus Below 500 meters)*

| | Malawi (IHS3) | | | | Ethiopia (ESS2) | | | |
|---|---|---|---|---|---|---|---|---|
| | **Entire sample** | **Sample [d < 500m]** | **Sample [d >= 500 m]** | **x[d <500m] - x[d >= 500m]** | **Entire sample** | **Sample [d < 500m]** | **Sample [d >= 500 m]** | **x[d <500m] - x[d >= 500m]** |
| Observations (Plots) | 18,256 | 9,798 | 8,458 | | 19,966 | 12,282 | 7,684 | |
| ***Plot Areas*** | | | | | | | | |
| GPS-based plot area (Ha) | 0.394 | 0.377 | 0.412 | -0.035*** | 0.197 | 0.163 | 0.249 | -0.086*** |
| Farmer-reported plot area (Ha) | 0.414 | 0.397 | 0.432 | -0.034*** | 0.193 | 0.162 | 0.239 | -0.076** |
| ***Yields*** | | | | | | | | |
| Maize yield (Kg/Ha) | 1,693 | 1,734 | 1,648 | 87 | | | | |
| Value of output/Ha | | | | | 29,303 | 25,373 | 34,563 | -9,190 |
| ***Plot Manager Characteristics*** | | | | | | | | |
| Female † | 0.261 | 0.270 | 0.251 | 0.019 | 0.153 | 0.171 | 0.125 | 0.046*** |
| Age (Years) | 43.147 | 44.017 | 42.235 | 1.782*** | 46.817 | 47.508 | 45.759 | 1.749** |
| Education (Years) | 5.028 | 4.961 | 5.098 | -0.137 | 1.874 | 1.965 | 1.735 | 0.230 |
| ***Household Characteristics*** | | | | | | | | |
| Household size | 4.934 | 4.843 | 5.031 | -0.188*** | 6.476 | 6.490 | 6.455 | 0.035 |
| # of HH members - [0,5] | 0.981 | 0.964 | 0.999 | -0.035 | 0.916 | 0.929 | 0.897 | 0.032 |
| # of HH members - [6,14] | 1.396 | 1.369 | 1.424 | -0.055 | 1.932 | 1.938 | 1.923 | 0.015 |
| # of female HH members - [15,39] | 0.901 | 0.868 | 0.936 | -0.068*** | 1.111 | 1.092 | 1.142 | -0.050 |
| # of male HH members - [15,39] | 0.837 | 0.797 | 0.879 | -0.081*** | 1.212 | 1.207 | 1.219 | -0.011 |
| # of female HH members - [40,59] | 0.270 | 0.269 | 0.270 | -0.001 | 0.386 | 0.388 | 0.384 | 0.004 |
| # of male HH members - [40,59] | 0.269 | 0.264 | 0.275 | -0.011 | 0.391 | 0.365 | 0.431 | -0.066** |
| # of HH members – 60 & above | 0.280 | 0.311 | 0.247 | 0.063*** | 0.527 | 0.571 | 0.458 | 0.113*** |
| Household consumption expenditures per capita | 50,431 | 48,099 | 52,876 | -4,777*** | 5,723 | 5,787 | 5,625 | 161 |
| Number of plots in the holding | 2.374 | 2.337 | 2.414 | -0.077** | 15.857 | 16.160 | 15.392 | 0.768 |
| ***Plot Characteristics*** | | | | | | | | |
| Owned by household † | 0.904 | 0.927 | 0.879 | 0.048*** | 0.866 | 0.902 | 0.811 | 0.091*** |
| Use of hired labor † | 0.223 | 0.188 | 0.260 | -0.071*** | 0.057 | 0.045 | 0.075 | -0.030** |
| Use of organic fertilizer † | 0.116 | 0.123 | 0.108 | 0.016** | 0.183 | 0.246 | 0.088 | 0.157*** |
| Use of inorganic fertilizer † | 0.618 | 0.630 | 0.606 | 0.024** | 0.404 | 0.419 | 0.383 | 0.036 |
| Irrigated † | 0.005 | 0.003 | 0.007 | -0.004** | 0.016 | 0.016 | 0.015 | 0.001 |
| Soil quality good † | 0.467 | 0.449 | 0.487 | -0.038*** | 0.327 | 0.338 | 0.309 | 0.030 |
| Soil quality poor † | 0.113 | 0.112 | 0.114 | -0.002 | 0.173 | 0.162 | 0.190 | -0.028 |

**Note:** † denotes a dummy variable. *** p<0.01, ** p<0.05, * p<0.1. Sample of plots within a 1-kilometer radius is the comparison group for the tests of mean differences.

# 3 Empirical Approach

## 3.1 Artificial Missingness Creation

Missing GPS-based plot areas measurements are often tied to numerous field logistics and cost constraints. The variable that underlies the overwhelming majority of missing GPS-based plot areas in household survey operations is the plot distance from the dwelling or the plot location with respect to the EA boundaries.[5] As noted above, the IHS3 instructed the enumerators to measure all plots within 2 hours travel time from the dwelling locations, while the ESS2 required the measurement of all plots, with the exception of those less than 40 square meters, irrespective of distance/travel time. For a more time and/or budget constrained operation, a lower threshold for GPS based land areas measurements could be enforced.

The first step in our analysis is to generate missing GPS-based plot areas in a way that would be similar to real-life field experience, or in other words, identify plots that could be deemed as "distant" in a large-scale survey operation. Towards this end, we work with two arbitrary but operationally-relevant distance thresholds that map well to the existing practices. In the first scenario, the plots are identified as distant if the georeferenced plot corner is located greater than 500 meters from the georeferenced dwelling unit of the associated household. The second scenario relies on a threshold of 1 kilometer instead.[6] Since survey implementers may want to get sense of the time requirements associated with visiting plot locations that

[5] Kilic et al. (2017) report that missingness in GPS-based plot areas due to refusal or physical inaccessibility, as opposed to distance, is near-negligible in the LSMS-ISA-supported surveys in Tanzania and Uganda. In the case of Malawi IHS3 2010/11, the considerable missingness in the stated reasons for lack of GPS-based plot area measurement prevents us from reporting specific statistics on missingness due to refusal or physical inaccessibility. In the case of the more recent Malawi Integrated Household Panel Survey (IHPS) 2016, the GPS-based plot area measures that are missing due to refusal or physical inaccessibility constitute 9 percent of the overall sample of plots without GPS-based areas, and 1 percent of the overall sample of plots.

[6] The variable underlying each distance threshold is the Euclidean (crow-fly) distance between the geo-referenced plot and dwelling location. Other geospatial measures of the plot distance to the dwelling were considered, including the estimated minimum cost distance that considers topography; the walking time associated with the minimum cost distance; and the inclination-adjusted measures of these two variables. The weighted pairwise correlation between any of the alternatives and our Euclidean distance measure is above 99 percent, and our results are robust to the use of these alternative distance measures.

are below versus above the chosen distance thresholds, consider, for instance, the walking time associated with the inclination-adjusted minimum cost distance between dwelling and plot locations in Malawi. For plots that are within the 500-meters and within the 1-kilometer threshold, the average walking time is 4 minutes and 6 minutes, respectively. Conversely, for plots that are outside the 500-meters and outside the 1-kilometer threshold, the average walking time is 33 minutes and 47 minutes, respectively.

Once the non-random, distant plots are identified according to one of the distance thresholds, we artificially create missingness in GPS-based areas among these plots at random, at a rate of 1 to 100 percent and with an increment of 1 percentage point and save these datasets separately. The choice of creating artificial missingness at random above an arbitrary distance threshold that identifies a non-random portion of the data is anchored in the specific way in which we see our findings would be operationalized, as explained below, particularly as part of multi-visit household and farm surveys that are in sync with a given agricultural season and that field a first, post-planting visit for parcel and plot demarcation and area measurement.

### 3.2 Multiple Imputation

The second step in our analysis is to carry out Multiple Imputation (MI) to fill the gaps in GPS-based plot areas in each unique data set that is created by a given distance threshold-artificial missingness combination.

MI, first proposed by (Rubin, 1987), is a Monte Carlo technique that replaces missing values for a given variable with $m > 1$ simulated alternatives. MI typically consists of three steps: (i) $m$ imputations (i.e. $m$ complete datasets) are generated based on an *imputation model* that encompasses a vector of observable covariates that predict the missingness in a given variable, (ii) statistical analysis is performed separately with each of the $m$ complete datasets, and (iii) the results obtained from $m$ complete data analyses are combined into a single set of multiply-imputed parameter estimates and standard errors. The conditions

under which valid inferences could be obtained from missing data has been laid out by (Rubin, 1987). Our procedure assumes that data are missing at random (MAR), that is that missing data could be predicted based on observable attributes underlying missingness. While the MAR assumption is not empirically testable, the limits of its tenability could be assessed in our study since we have otherwise complete datasets that are used as validation samples.

In building the imputation model, the literature (Rubin, 1996) or (van Buuren, et al., 1999) advises to include as explanatory variables: (i) the variables appearing in the analysis model that features the multiply-imputed variable(s), (ii) the variables that are known to have influenced the occurrence of missing data, and other variables for which the distributions differ between the response and non-response groups, (iii) the variables that explain a considerable amount of variance of the multiply-imputed variable(s) and that help to reduce the uncertainty of the imputations, and (iv) the variables with information on the features of the complex survey design, including stratum and cluster identifiers, and sampling weights.

In their MI application to missingness in GPS-based land areas in Tanzania and Uganda, (Kilic, et al., 2017) attempt to provide support for the MAR assumption by (i) detailing the field work processes underlying the missing data, (ii) providing insights from their field experience and interactions with the survey teams, (iii) systematically documenting the established guidelines on imputation model specification, and (iv) including in the imputation model explanatory variables that influence the occurrence of missing data; that have different distributions between the response and non-response groups; that explain a considerable amount of variance of the multiply-imputed variable; and that include information on the survey design.

Our imputation model specification is anchored in these considerations, and includes farmer-reported plot area, which is both a powerful predictor and an alternative measure of the GPS-based plot area. The use of a self-reported variable in the imputation model to tackle item non-response in an objectively-measured

variable has been pursued also by (Schenker, et al., 2010), who feed self-reported health measures into a model to multiply impute clinical values in a different survey.[7]

For illustration, Table 5 and Table 6 show the details of the Ordinary Least Squares (OLS) imputation model for Malawi and Ethiopia, respectively. In addition to farmer-reported plot area, we include plot manager, household and other plots attributes as predictors. The model specification differs slightly between the IHS3 and the ESS2 depending of the availability of the variables or the specificity of the data set. For example, the raw data on farmer-reported plot areas could have been expressed in non-standard measurement units in the ESS2, as such we add dummy variables for these units in the imputation model for Ethiopia.

[7] The literature on the use of MI to address missing information problems is vast and cuts across several disciplines, including but not limited to economics, statistics, sociology, public health, medicine, and epidemiology. The empirical work relying on MI to deal with missingness in income data is noteworthy given the types of household surveys that inform our analyses, and considering the rates of item non-response that are dealt with in the literature on income and that present similarities to the patterns in Table 1 (Schenker, et al., 2006); (Zarnoch, et al., 2010); (Giusti & Little, 2011); (Ahearn, et al., 2011); (Vermaak, 2011).

*Table 5: OLS Imputation Model Results for Malawi - Dependent Variable: GPS-Based Plot Area (HA)*

| | Full Sample | Plots [d <= 1km] | Plots [d <= 500m] |
|---|---|---|---|
| ***Plot Area*** | | | |
| Farmer-reported plot area (Ha) | 0.583*** | 0.583*** | 0.613*** |
| | (0.006) | (0.007) | (0.008) |
| ***Plot Manager Characteristics*** | | | |
| Female † | -0.060*** | -0.069*** | -0.069*** |
| | (0.013) | (0.015) | (0.016) |
| Age (Years) | 0.003*** | 0.003*** | 0.002*** |
| | (0.001) | (0.001) | (0.001) |
| Education (Years) | -0.003** | -0.004** | -0.005** |
| | (0.001) | (0.002) | (0.002) |
| Plot manager is respondent † | 0.032*** | 0.042*** | 0.039*** |
| | (0.011) | (0.013) | (0.014) |
| Has a chronic disease † | -0.039** | -0.043** | -0.054** |
| | (0.017) | (0.020) | (0.022) |
| Religion: Christian † | 0.054** | 0.062** | 0.037 |
| | (0.025) | (0.029) | (0.032) |
| Religion: Muslim † | -0.018 | -0.005 | -0.042 |
| | (0.030) | (0.036) | (0.039) |
| Religion: Traditional † | 0.012 | 0.045 | -0.014 |
| | (0.047) | (0.057) | (0.070) |
| ***Plot Characteristics*** | | | |
| Soil quality good † | -0.025 | -0.021 | -0.015 |
| | (0.016) | (0.019) | (0.021) |
| Use of organic fertilizer † | 0.036** | 0.048*** | 0.033* |
| | (0.016) | (0.018) | (0.020) |
| Use of inorganic fertilizer † | 0.086*** | 0.085*** | 0.064*** |
| | (0.011) | (0.012) | (0.013) |
| Use of hired labor † | 0.113*** | 0.104*** | 0.109*** |
| | (0.013) | (0.015) | (0.017) |
| Irrigated † | -0.142** | -0.088 | -0.035 |
| | (0.072) | (0.088) | (0.111) |
| ***Household Characteristics*** | | | |
| # of HH members - [0,5] | 0.006 | 0.005 | 0.008 |
| | (0.006) | (0.007) | (0.007) |
| # of HH members - [6,14] | 0.022*** | 0.022*** | 0.022*** |
| | (0.004) | (0.005) | (0.005) |
| # of female HH members - [15,39] | 0.011 | 0.001 | 0.001 |
| | (0.008) | (0.009) | (0.010) |
| # of female HH members - [40,59] | 0.057*** | 0.055*** | 0.068*** |
| | (0.013) | (0.015) | (0.017) |
| # of male HH members - [15,39] | 0.023*** | 0.022*** | 0.016* |
| | (0.006) | (0.008) | (0.008) |
| # of male HH members - [40,59] | 0.039*** | 0.040** | 0.044** |
| | (0.014) | (0.016) | (0.017) |
| # of HH members – 60 & above | 0.030** | 0.025 | 0.041** |
| | (0.015) | (0.017) | (0.018) |
| Wealth index | 0.010*** | 0.008** | 0.006 |
| | (0.003) | (0.003) | (0.004) |

*Table 5 (Cont'd)*

| | Full Sample | Plots [d <= 1km] | Plots [d <= 500m] |
|---|---|---|---|
| ***Household Characteristics (Cont'd)*** | | | |
| Agriculture implement index | 0.022*** | 0.028*** | 0.030*** |
| | (0.004) | (0.005) | (0.006) |
| Number of plots in the holding | -0.053*** | -0.052*** | -0.050*** |
| | (0.005) | (0.006) | (0.006) |
| Access to non-farm labor income † | -0.054*** | -0.052*** | -0.065*** |
| | (0.010) | (0.012) | (0.013) |
| Access to non-Farm non-labor income † | -0.018* | -0.024** | -0.018 |
| | (0.010) | (0.012) | (0.013) |
| **Observations** | 18,256 | 13,161 | 9,798 |
| **Adjusted R2** | 0.430 | 0.425 | 0.466 |

**Note:** † denotes a dummy variable. *** p<0.01, ** p<0.05, * p<0.1. Constant, district fixed effects (30 in total) included but not reported.

*Table 6: OLS Imputation Model Results for Ethiopia - Dependent Variable: GPS-Based Plot Area (HA)*

| | Full Sample | Plots [d <= 1km] | Plots [d <= 500m] |
|---|---|---|---|
| ***Plot Area*** | | | |
| Farmer-reported plot area (Ha) | 0.871*** | 0.827*** | 0.865*** |
| | (0.004) | (0.004) | (0.004) |
| Unit reported: Square Meters † | 0.278*** | 0.270*** | 0.380*** |
| | (0.021) | (0.023) | (0.027) |
| Unit reported: Timad† | 0.241*** | 0.253*** | 0.371*** |
| | (0.018) | (0.021) | (0.025) |
| Unit reported: Boy † | 0.160*** | 0.167*** | 0.285*** |
| | (0.020) | (0.022) | (0.026) |
| Unit reported: Senga † | 0.301*** | 0.247*** | 0.336*** |
| | (0.024) | (0.027) | (0.031) |
| Unit reported: Kert † | 0.188*** | 0.200*** | 0.295*** |
| | (0.028) | (0.030) | (0.034) |
| ***Plot Manager Characteristics*** | | | |
| Female † | -0.011* | -0.007 | -0.008 |
| | (0.007) | (0.007) | (0.007) |
| Age (Years) | -0.000 | -0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| Education (Years) | -0.002** | -0.002*** | -0.002* |
| | (0.001) | (0.001) | (0.001) |
| Religion: Orthodox † | 0.023 | 0.037 | 0.031 |
| | (0.024) | (0.025) | (0.028) |
| Religion: Protestant † | 0.030 | 0.046* | 0.043 |
| | (0.025) | (0.025) | (0.029) |
| Religion: Muslim † | 0.022 | 0.024 | 0.019 |
| | (0.024) | (0.025) | (0.028) |
| Religion: Traditional † | 0.028 | 0.045 | 0.039 |
| | (0.037) | (0.039) | (0.042) |
| | 0.030 | 0.046* | 0.043 |
| ***Plot Characteristics*** | | | |
| Cultivated † | 0.037*** | 0.030*** | 0.028*** |
| | (0.008) | (0.008) | (0.009) |
| Pasture † | 0.078*** | 0.073*** | 0.066*** |
| | (0.011) | (0.010) | (0.011) |
| Fallowed † | 0.047*** | 0.047*** | 0.040*** |
| | (0.013) | (0.013) | (0.015) |
| Soil quality good † | -0.013* | -0.007 | -0.012 |
| | (0.007) | (0.007) | (0.008) |
| Use of organic fertilizer † | -0.040*** | -0.044*** | -0.035*** |
| | (0.008) | (0.008) | (0.009) |
| Use of hired labor † | 0.070*** | 0.063*** | 0.051*** |
| | (0.010) | (0.010) | (0.012) |
| Irrigated † | -0.022 | -0.014 | -0.010 |
| | (0.015) | (0.015) | (0.016) |
| ***Household Characteristics*** | | | |
| # of HH members - [0,5] | 0.003 | 0.003 | 0.003 |
| | (0.003) | (0.003) | (0.003) |

*Table 6 (Cont'd)*

| | Full Sample | Plots [d <= 1km] | Plots [d <= 500m] |
|---|---|---|---|
| ***Household Characteristics (Cont'd)*** | | | |
| # of HH members - [6,14] | 0.001 | -0.000 | -0.002 |
| | (0.002) | (0.002) | (0.002) |
| # of female HH members - [15,39] | 0.007** | 0.005 | 0.004 |
| | (0.003) | (0.003) | (0.003) |
| # of female HH members - [40,59] | 0.012** | 0.009* | 0.009 |
| | (0.005) | (0.005) | (0.006) |
| # of male HH members - [15,39] | 0.010*** | 0.006** | 0.006** |
| | (0.003) | (0.003) | (0.003) |
| # of male HH members - [40,59] | 0.012** | 0.015*** | 0.011* |
| | (0.005) | (0.005) | (0.006) |
| # of HH members – 60 & above | 0.007** | 0.006* | 0.002 |
| | (0.003) | (0.003) | (0.003) |
| Household consumption expenditure per capita | 0.001* | 0.001*** | 0.001*** |
| | (0.000) | (0.000) | (0.000) |
| Number of plots in the holding | -0.001* | -0.001 | -0.001 |
| | (0.000) | (0.000) | (0.000) |
| **Observations** | 19,966 | 15352 | 12282 |
| **Adjusted R2** | 0.789 | 0.768 | 0.805 |

**Note:** † denotes a dummy variable. *** p<0.01, ** p<0.05, * p<0.1. Constant, woreda fixed effects (228 in total) included but not reported.

We estimate the imputation model using each unique dataset that is created by a given distance threshold-artificial missingness combination. While the results confirm that the predictions are essentially driven by the farmer-reported plot area, the more comprehensive model improves the accuracy and precision of our predictions. As pointed out by (Kilic, et al., 2017), it is worth emphasizing that the imputation model neither intends to provide a parsimonious description of the data nor attempts to portray structural relationships among variables. Instead, it attempts to be as comprehensive as possible to minimize any bias that could stem from omitting variables that might be relevant to the pattern of missingness or the subsequent analysis. "The possible lost precision when including unimportant predictors is usually viewed as a relatively small price to pay for the general validity of analyses of the resultant multiply-imputed database" (Rubin, 1996).

In multiply imputing missing values that have been artificially created in each scenario, we fit plot-level OLS regression models with the GPS-based plot area as the dependent variable and obtain linear predictions

for all plots in the dataset. Under the partially parametric method of predictive mean matching (PMM)[8], we use the linear prediction as a distance measure to form a set of 5 nearest neighbors chosen from the plot sample with GPS-based area measures, and randomly pick one of the neighbors whose observed GPS-based plot area value replaces the missing value for the incomplete case at hand.[9] The imputation is carried out 50 times to reduce the potential sampling error due to imputation;[10] 50 complete datasets are generated; and the posterior estimates of the model parameters are obtained from a bootstrapped sample[11]. By drawing from the observed data, PMM preserves the distribution of observed values in the missing part of the data, which makes it more robust than the fully parametric regression approach. In total, we generate 50 complete datasets of GPS-based land plot areas for each of rate missingness (from 1 to 100) above each distance threshold for each country. These data sets are used to assess the tolerable rates of missingness, as explained below.

### 3.3 Assessing the tolerable rates of missingness in GPS-based plot areas

To assess the performance of the imputation model, we compare, the distributions of the *true*, *observed* versions of key variables that rely on GPS-based plot areas with the distributions of their completed (observed plus imputed) counterparts. The key outcomes that our assessment focuses on is GPS-based plot area and plot-level agricultural productivity, which is measured as the quantity or value of crop harvested based on farmer-reporting (the numerator) over cultivated land (the denominator). As discussed earlier, plot-level agricultural productivity is of policy relevance.

[8] For more insights on the use of predictive mean matching, see (Little, 1988).
[9] The results are robust to using linear regression, as opposed to PMM. The number of nearest neighbors in the PMM framework is inversely related to the correlation among imputations. While high correlation may increase the variability in MI point estimates, low correlation may increase the bias in MI point estimates. The literature does not provide definitive guidance on the decision regarding the number of nearest neighbors, but the results are robust to the specification of ten nearest neighbors, with or without bootstrapping.
[10] The results are robust to performing 100, as opposed to 50, imputations.
[11] The results are robust to sampling estimates from the posterior distribution of model parameters.

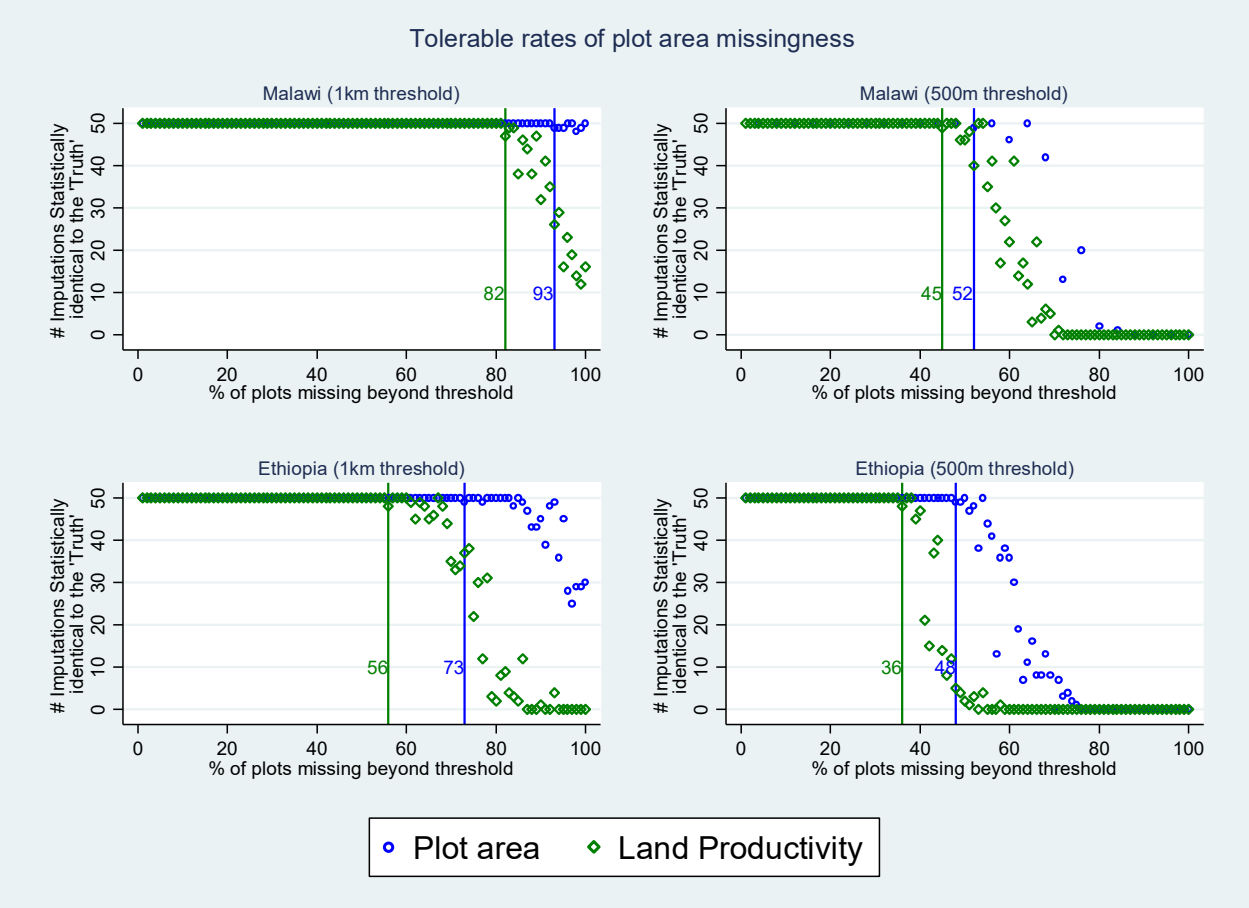
Given the nature of the problems to which MI is applied, it is often difficult for analysts to verify the appropriateness of their imputation procedures. Imputation values are guesses of unobserved, unknown values (Abayomi, et al., 2008). In this study, however, missingness is artificially created such that the true values are known. This allows direct comparison of the distributions of the observed versus the completed data. Numerically, the comparison of the empirical distributions is done using the Kolmogorov-Smirnov (KS) test separately for each outcome variable for each of the levels of missingness, raising the flag if there are statistically significant differences at the 5 percent level[12] for at least 1 of the 50 imputations generated. As noted by (Abayomi, et al., 2008), there is no reason to suppose that setting a 5 percent level of significance will be appropriate when producing a MI diagnostic through density comparisons. However, it is useful to start with this rule and further examine the results.

## 4 Results

The results of our simulations are illustrated in Figure 1. Each panel shows the results from the analyses conducted with a given distance threshold in a given country. In each panel, the y-axis shows the number of imputations out of 50 for which the KS test indicates that the distribution of the relevant outcome variable derived from the imputed dataset is statistically indistinguishable from its observed counterpart. We additionally highlight the tolerable rates of missingness with vertical lines. The x-axis, on the other hand, shows the percentage of simulated missing GPS-based plot areas measurements beyond a given distance threshold. Three general observations emerge from Figure 1.

[12] The *p*-values for the test are approximate. The imputations are generated from the observed data. Hence, the empirical distributions are not independent of the observed data.

*Figure 1: Tolerable Rates of Missingness in GPS-Based Plot Areas Above a Given Distance Threshold for Plot Area & Plot-Level Yield Analysis*




First, at low rates of missingness, the distribution of each outcome variable in each of the 50 imputed datasets are statistically indistinguishable from the observed counterpart. As the rate of missingness increases, this count starts to decrease until only a small number (between 0 and 10) of the imputations appear to have distributions that are not statistically different from the true distribution. Second, the tolerable rates of missingness are lower with the 500-meter distance threshold in comparison to the 1-kilometer counterpart. Third, plot-level agricultural productivity is more sensitive to missingness than plot area (i.e. the tolerable rate of missingness is reached earlier in the case of the latter).

The first and second observations confirm the expectations anchored in the descriptive analyses discussed in Section 2. Plots that are further from the dwelling are inherently different from the ones that are closer. Thus, as missingness increases, the pool of plots with similar characteristics (and thus comparable areas) to choose from gets smaller, and it is understandable that the distribution differs substantially. The third observation is also foreseen: land area being the denominator of the formula for yield, a small deviation of the imputed values from the observed land values brings about a relatively more important deviation in the yield estimates obtained from them. Consequently, the yields calculated from the imputed land areas differ substantially from the true yields at lower rates of missingness.

We now compare the results obtained in the different panels depicted in Figure 1. For convenience, the tolerable rates of missing GPS-based plot areas are summarized in Table 7. Along with the tolerable rates in terms of the percentages of plot areas observations that could go missing beyond a given distance threshold, we report the corresponding overall rates of missingness in parentheses. In the discussion that follows, we focus on the discussion of the results pertaining to plot-level agricultural productivity, given the policy relevance of the outcome and its lower tolerance to missingness vis-à-vis plot area.

*Table 7: Tolerable Rates of Missingness in GPS-Based Plot Areas Above a Given Distance Threshold for Plot Area & Plot-Level Yield Analysis*

| | | **Plot Area** | **Yield** |
|---|---|---|---|
| | | *Tolerable rate (%)* | *Tolerable rate (%)* |
| **Malawi** | **1.0 km** | 93 (26) | 82 (23) |
| | **500 m** | 52 (24) | 45 (21) |
| **Ethiopia** | **1.0 km** | 73 (18) | 56 (13) |
| | **500 m** | 48 (20) | 36 (15) |

**Note:** The overall rates of missingness implied by the tolerable rates of missingness above a given distance threshold are noted in the parentheses.

The results obtained with the 1-kilometer threshold are very encouraging. In Malawi, the multiply-imputed distribution of the plot-level productivity measure is statistically indistinguishable from the true distribution in each of the 50 imputed datasets with up to 82 percent missingness in GPS-based plot areas that are more than 1 kilometer away from the associated dwelling. The comparable figure in Ethiopia is 56 percent. Put differently, the number of plots for which GPS-based area measurement can be forgone represent 23 percent and 13 percent of the overall plot sample in Malawi and Ethiopia, respectively. These findings indicate that in Ethiopia, the plot-level agricultural productivity estimation is more sensitive, compared to Malawi, to missingness in GPS-based distant plot areas.

Further, as noted above, we get lower tolerable rates of missingness among distant GPS-based plot areas when we lower the distance threshold from 1 kilometer to 500 meters. In Malawi, the multiply-imputed distribution of the plot-level productivity measure is statistically indistinguishable from the true distribution in each of the 50 imputed datasets with up to 45 percent missingness in GPS-based plot areas that are more than 500 meters from the associated dwelling. The comparable figure in Ethiopia is 36 percent. In this case, the number of plots for which GPS-based area measurement can be omitted represent 21 percent and 15 percent of the overall plot sample in Malawi and Ethiopia, respectively.

The cross-country differences in tolerable missingness rates are likely in part tied to the differences in farm organization, which mediate the differences in variability in the outcomes of interest.[13] On the one hand, the average plot size in hectares in Malawi (0.4) is twice as much as the comparable statistic in Ethiopia (0.2), as reported in Table 3. On the other hand, the household-level average number of plots per holding in Ethiopia (11.7) is more than six times the comparable figure in Malawi (1.9). While the spatial distribution of the plot samples across the distance intervals in Table 2 are comparable across the two settings, the average plot distance from the dwelling is 2.19 kilometers in Malawi, with a 95 percent

[13] Unless otherwise stated, the statistics in this paragraph are not reported in any of the tables but have been computed based on the same datasets used for analysis.

confidence interval of 1.91-2.47, versus 1.10 kilometers in Ethiopia with a 95 percent confidence interval of 0.76-1.43. The plot distance from the dwelling further exhibits cross-country distributional differences that are statistically significant at the 1 percent level.

Finally, Table 8 presents country-specific multiply-imputed mean versus true mean comparisons for plot-level area and agricultural productivity, following MI at identified tolerable rates of missingness above the distance thresholds as reported in Table 7. Irrespective of the distance threshold and country in question, the root mean square error for plot area is close to zero and the difference between the MI mean and the true mean as a percentage of the true mean does not exceed 1.5 percent. For plot-level agricultural productivity, we have more promising findings in Malawi compared to Ethiopia. In Malawi, for instance, at 82 percent missingness above the 1-kilometer threshold, the difference between the MI mean and the true mean as a percentage of the true mean stands at 7.5 percent. The comparable statistic for Ethiopia is 40.4 percent. These findings underscore the relative sensitivity to missingness of plot-level agricultural productivity measures vis-à-vis plot area, and the fact that this sensitivity is likely to vary by country and production system complexity, as in this study.

*Table 8: Country-Specific Multiply Imputed Overall Mean versus True Mean Comparisons Following Multiple Imputation At Identified Tolerable Rates of Missingness above the Distance Thresholds as Specified in Table 7*

| Country | Distance Threshold | Tolerable Rate of Missingness Above Distance Threshold | Variable | MI Mean | True Mean | Difference | Difference % of True Mean | RMSE | RMSE % of True Mean |
|---|---|---|---|---|---|---|---|---|---|
| Ethiopia | 1 Kilometer | 73 | Plot Area | 0.206 | 0.209 | -0.003 | -1.4% | 0.003 | 1.4% |
| | | 56 | Yield | 41,141 | 29,303 | 11,839 | 40.4% | 11,839 | 40.4% |
| | 500 Meters | 48 | Plot Area | 0.207 | 0.209 | -0.002 | -1.1% | 0.002 | 1.1% |
| | | 36 | Yield | 39,628 | 29,303 | 10,325 | 35.2% | 10,325 | 35.2% |
| Malawi | 1 Kilometer | 93 | Plot Area | 0.390 | 0.394 | -0.004 | -0.9% | 0.004 | 0.9% |
| | | 82 | Yield | 1,821 | 1,693 | 128 | 7.5% | 128 | 7.5% |
| | 500 Meters | 52 | Plot Area | 0.391 | 0.394 | -0.003 | -0.8% | 0.003 | 0.8% |
| | | 45 | Yield | 1,794 | 1,693 | 101 | 5.9% | 101 | 5.9% |

**Note:** RMSE stands for Root Mean Squared Error. Plot area is in hectares. Yield is maize production in kilograms per hectare in Malawi and value of output per hectare in Ethiopia.

## 5 Conclusion

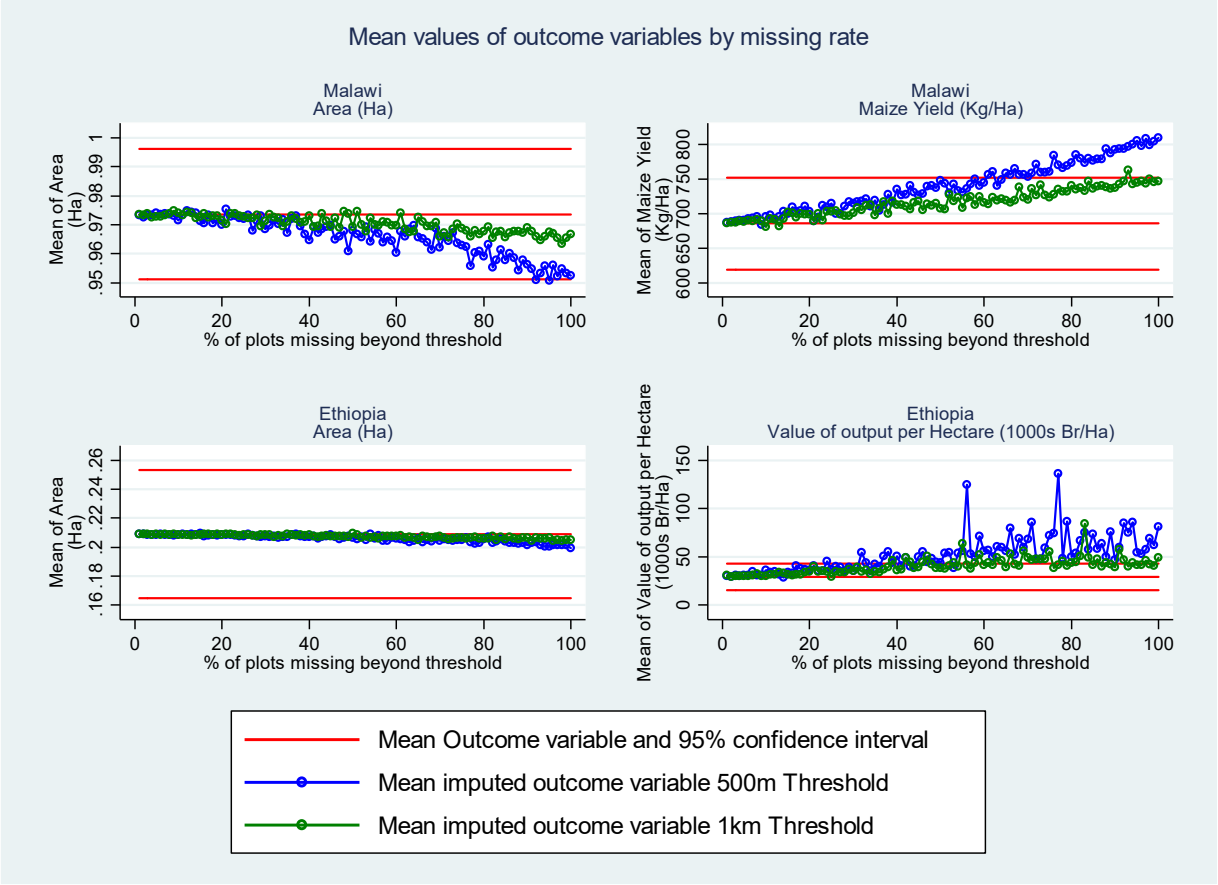
This paper provides further evidence that combining GPS-based plot areas measurements with farmer-reported plots areas in a Multiple Imputation (MI) application can result in reliable simulations of missing GPS-based plot areas. The analysis extends previous research by using survey data from Malawi and Ethiopia that feature negligible levels of missing GPS-based area measures. By artificially simulating missingness among distant plots in otherwise assumed-to-be-complete data, we compare the MI-based predictions to the true, observed values and gauge the levels of missingness in GPS-based land area measurements that can be handled with MI without compromising the robustness of key land area related statistics.

Since the microdata on land areas inform a wide range of research applications on smallholder agriculture, agricultural and development economists that rely heavily on public use datasets are therefore encouraged to think more critically about the use of model-based approaches to address missingness not only in GPS-based land areas but also in other variables with known missing information mechanisms that can be captured with confidence as part of the imputation model specification.

Among the outcome variables of interest, plot-level agricultural productivity, as measured by maize yield in Malawi and total harvest value per land area in Ethiopia, is found to be more sensitive to missingness. Still, we show that in Malawi, by obtaining GPS-based area measures for only 18 percent of the distant plots in Malawi that are further than 1 kilometer with respect to the dwelling location and that would be selected at random, and by multiply-imputing the remaining missing GPS-based plot areas, we can derive comparable means and distributions with respect to the true data. In the case of Ethiopia, we need a randomly selected sample of at least 44 percent of the distant plots based on the same distance threshold to achieve the same objective.

These findings are further buttressed by Figure 2. At these rates of missingness, the mean values of the multiply-imputed outcome variables of interest are within the 95 percent confidence intervals of the corresponding true mean values.

*Figure 2: Mean Values of Plot Area & Plot-Level Yield by Rate of Missingness Above a Given Distance Threshold*




Overall, further investment into the operationalization of this evidence could result in significant savings in terms of time and resources. Our findings are particularly relevant for household and farm surveys that visit sampled households at least twice and in sync with a given agricultural season, such as the LSMS-ISA supported surveys in Ethiopia, Malawi, Niger, Nigeria, and Mali. In these surveys, the sampled households are visited for the first time during the post-planting period for parcel and plot demarcation and area

measurement. As explained above, GPS-based area measures are obtained within a specific radius that is defined ex-ante by the implementing agency, traditionally in terms of subjective assessments of distance, travel time, and plot location as it relates to enumeration area/village boundaries. Hence, to the extent that a smallholder production system presents a sufficient degree of similarity to the Malawian or Ethiopian contexts and a given implementing agency is fielding a survey that mirrors the fieldwork set up in these countries, following the post-planting visit, the survey management team could consider reviewing the set of unmeasured, distant plots, and select a random subset of those in an attempt to achieve one of eight tolerable rates of missingness, as reported in Table 7, and as a function of the distance threshold-outcome variable combination. This random subset of distant plots could then be prioritized for GPS-based area measurement during the subsequent visit(s) to the sampled households, and the resulting, "more complete" dataset could be subject to MI to predict the remaining missing GPS-based plot areas.

However, since the tolerable missingness rates may vary by country, distance threshold and outcome variable, similar analyses could be replicated (i) based on the IHS3 and ESS datasets that inform our analysis but by using alternative distance thresholds and outcome variables, and (ii) using other survey data that exhibit low rates of missingness in GPS-based plot areas. The findings would in turn catalyze the convergence onto comprehensive operational guidelines for survey practitioners.[14] Finally, although dealing with missingness empirically in the post-fieldwork period is always an option, there is no substitute for good fieldwork to prevent unwarranted missing measurements as much as possible. Thus, survey practitioners, including agricultural and development economists involved in primary data collection, should follow a combination of (i) well-supervised field practices to reduce missingness, as exemplified in Section 2, and (ii) sound MI applications to fill the data gaps that will still be unavoidable to a degree.

[14] Along with the paper, we make available the Stata formatted datasets and syntax, along with a readme file, that could be used not only to replicate our analyses, but to select different distance thresholds and outcome variables based on which tolerable rates of missingness can be derived.

## 6 References


Abayomi, K., Gelman, A. & Levy, M., 2008. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 57(3), pp. 273-291.

Ahearn, M., Banker, D., Clay, D. M. & Milkove, D., 2011. Comparative survey imputation methods for farm household income. *American Journal of Agricultural Economics,* 93(2), pp. 613-618.

Carletto, C., Gourlay, S., Murray, S. & Zezza, A., 2017. Cheaper, faster, and more than good enough. Is GPS the new gold standard in land area measurement? *Survey Research Methods*, 11(3), pp. 235-265.

Carletto, C., Gourlay, S. & Winters, P., 2015. From guesstimates to GPStimates: land area measurement and implications for agricultural analysis. *Journal of African Economies,* 24(5), pp. 593-628.

Carletto, C., Savastanao, S. & Zezza, A., 2013. Fact or artifact: The impact of measurement errors on the farm size–productivity relationship. *Journal of Development Economics,* 103, pp. 254-261.

Dorward, A. & Chirwa, E., 2010. *A review of methods for estimating yield and production impacts,* s.l.: Centre for Development, Environment and Policy, SOAS, University of London, and Wadonda Consult.

Giusti, C. & Little, R. J., 2011. An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics,* 27(2), pp. 211-229.

Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2016. [Online] Available at: https://goo.gl/A7nSq2 [Accessed 24 6 2016].

Kilic, T., Palacios-López, A. & Goldstein, M., 2015. Caught in a Productivity Trap: A Distributional Perspective on Gender Differences in Malawian Agriculture. *World Development,* 70, pp. 416-463.

Kilic, T., Zezza, A., Carletto, C. & Savastano, S., 2017. Missing(ness) in action : selectivity bias in GPS-based land area measurements. *World Development*, 92, pp. 143-157.

Little, R. J., 1988. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics,* 6(3), pp. 287-296.

Marchenko, Y. V. & Eddings, W., 2011. *A note on how to perform multiple-imputation diagnostics in Stata,* College Station, TX: StataCorp.

Rubin, D. B., 1987. *Multiple Imputation for Nonresponse in Surveys.* Hoboken, NJ: Jon Wiley & Sons.

Rubin, D. B., 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association,* 91(434), pp. 473-489.

Schenker, N., Raghunathan, T. & Bondarenko, I., 2010. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine,* 29(5), pp. 533-545.

Schenker, N. et al., 2006. Multiple Imputation of Missing Income Data in the National Health Interview Survey. *Journal of the American Statistical Association,* 101, pp. 924-933.

van Buuren, S., Boshuizen, H. C. & Knook, D. L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine,* 18(6), pp. 681-694.

Vermaak, C., 2011. Tracking poverty with coarse data: evidence from South Africa. *The Journal of Economic Inequality,* 10(2), pp. 239-265.

Zarnoch, S. J., Cordell, H. K., Betz, C. J. & Bergstorm, J. C., 2010. *Multiple Imputation: An Application to Income Nonresponse in the National Survey on Recreation and the Environment,* s.l.: United States Department of Agriculture Forest Service.