PRELIMINARY DRAFT – APRIL 2013

DO NOT CITE OR QUOTE WITHOUT AUTHOR'S WRITTEN PERMISSION

Missing(ness) in Action: Selectivity Bias in GPS-Based Land Area Measurements[1]

Talip Kilic*[†], Alberto Zezza[†], Calogero Carletto[†], Sara Savastano[‡]

* Corresponding Author: tkilic@worldbank.org
[†] Development Research Group, World Bank, Washington, DC, USA.
[‡] Department of Economics and Finance, University of Rome - Tor Vergata, Italy.

Land area is a fundamental component of agricultural statistics, and of analyses undertaken by agricultural economists. While household surveys in developing countries have traditionally relied on farmers' own, potentially error-prone, assessments to collect land area information, the availability of affordable and increasingly reliable Global Positioning System (GPS) units has recently made GPS-based area measurement a practical alternative. Nonetheless, survey implementing agencies typically require interviewers to record GPS-based area measurements *only* for plots within a given radius of dwelling locations, in order to reduce survey costs, keep interview durations within reasonable limits, and avoid the difficulty of asking respondents to accompany interviewers to distant plots. It is, therefore, common for as much as a third of the sample plots not to be measured, and available research has not shed light on the possible selection bias in analyses that rely on partial data due to gaps in GPS-based area measures. We explore the systematic patterns of missingness in GPS-based plot areas, and investigate their empirical implications in the context of the inverse scale-land productivity relationship. Using Multiple Imputation (MI) to predict missing GPS-based plot areas in nationally-representative survey data from Uganda and Tanzania, we highlight the potential of MI in reliably simulating the missing data, and document a stronger inverse scale-land productivity relationship with the complete data. Our study demonstrates the use of judiciously reconstructed GPS-based areas in alleviating concerns over potential measurement error in farmer–reported area assessments, and with regards to systematic bias in plot selection for GPS-based area measurement.

Keywords: Global Positioning System, Land Area Measurement, Land Productivity, Multiple Imputation

[1] The senior authorship is shared by Alberto Zezza and Talip Kilic. The results from all robustness checks referenced in the main text or the footnotes are available upon request. The authors wish to thank Dean Jolliffe for his comments on the earlier draft of the paper.

1. INTRODUCTION

In Ancient Egypt, dating back to the Old Kingdom (2686 BC – 2181 BC), the parcels along the Nile were allocated by the King to individuals, who were in turn recorded in the land administration system and were taxed for their land. However, the shape and the boundaries of the land on the banks of the Great River were often subject to change as a result of the annual flooding. This dynamic required the deployment of land surveyors, known as *rope stretchers*, in order to re-measure the land, replace the inscriptions that marked the boundaries, and resolve potential disputes between the neighbors so that the taxes could be determined accurately. The measuring devices used by these surveyors included an A-frame shaped level with a plumb bob, and a rope knotted at intervals (Lyons, 1927; Berger, 1934).

Land area measurement endures to be a fundamental component agricultural statistics[2], and of analyses undertaken by agricultural economists, including the large body of policy-relevant empirical work on testing the existence of an inverse scale–land productivity relationship (see Carletto et al., 2011 and the references cited therein). Although the traditional tools of land surveyors, such as rope and compass, that date back to hundreds of years has remained in the choice set of agricultural statisticians to measure land areas accurately, these conventional approaches are neither time- nor cost-effective in the context of large data collection efforts, which often only record farmers' own assessments of land areas. That said, the availability of affordable, portable and relatively reliable Global Positioning System (GPS) devices over the last two decades has made GPS-based land area measurement a practical alternative that is increasingly being applied in surveys worldwide.[3]

While the use of GPS technology in area measurement is the way of the future and there is increasing evidence for the benefits of utilizing GPS units as part of household survey operations in Africa (Dorward and Chirwa, 2010), in these relatively early stages of its application, it is important to assess the advantages it can provide over other methods as well as the possible issues associated with its use, particularly in the context of low income countries where the numeracy of respondents is limited, "non-standard" measurement units are commonplace, cadastral records are nonexistent, and the financial and human resources of national statistical offices are severely limited.

A key issue on which little analysis has been conducted is the potential selection bias associated with GPS-based plot area estimates. For a variety of reasons connected to (i) reducing survey costs, (ii) keeping household interview durations within reasonable limits, and (iii) the difficulty of asking respondents to accompany enumerators to agricultural plots that are situated far from dwelling locations, survey fieldwork protocols often require enumerators to record GPS-based area measurements only for plots that are within a given radius of the dwelling location. Plots that are located within an acceptable distance from the dwelling may still not be measured due to respondent refusal or lack of physical access at the time of the interview. Consequently, GPS-based plot area

[2] For instance, in the World Bank World Development Indicators database, 15 out of 32 indicators classified under the agricultural and rural development category rely on land area information.

[3] Schoning et al. (2005) document as part of the 2003 pilot of the Uganda Agricultural Census that the average time for measuring by rope and compass all parcels in a given agricultural holding was over 3 hours, which was more than three times as much as when GPS technology was employed to measure parcel areas.

measurements, which are still recorded alongside farmers' own assessments in agricultural and household surveys, present large gaps and a score of missing unit values that can be catalogued as "item non-response." It is not uncommon to have missing GPS-based area values for as much as a third of the plots owned and/or cultivated by sample households, thus limiting the analytical value of the GPS-based data collection effort, to the point that one could even question whether it is worth collecting such data. The potential selection bias in GPS-based plot area estimates is most recently recognized but not addressed by Carletto et al. (forthcoming), who explore the empirical implications associated with the choice of farm size measures based on farmer-reported vs. GPS-based plot areas, and document a stronger inverse farm size-land productivity relationship while using the GPS-based farm size measure.

The gaps in GPS-based plot area data are reminiscent of similar patterns of item non-response in other statistical series, such as income data in OECD countries. In the US, for instance, the problem is common in some of the Census Bureau series (Scheuren, 2005), in the National Interview Health Survey (NHIS; Schenker et al., 2006), but also in agricultural and forestry surveys such as the Agricultural Resource Management Survey (ARMS, Ahearn et al., 2011) and the National Survey on Recreation and the Environment (Zarnoch et al., 2010). Other examples include the Labor Force Survey of the Municipality of Florence in Italy (Giusti and Little, 2011), and the Labor Force Survey in South Africa (Vermaak, 2012). What several of these experiences have in common is the systematic approach to dealing with missing data as datasets are released in the public domain and analyzed for research and policy purposes. Increasingly, the technique of choice for dealing with missing data in public use datasets is "multiple imputation" (Rubin, 1987; 1996), which, provided that the assumptions regarding the missing data mechanism hold, allows for valid inference on a restored "complete" dataset, while taking into account the uncertainty associated with the imputation process itself.

As the drive towards open data advances in the developing world, methodical approaches to reliably address structural gaps in critical data, such as GPS-based land areas, are likely to be of increasing interest to national statistical agencies, their international partners, and data analysts. Our study (i) documents the way in which the process of selecting plots for GPS-based land area measurement in accordance with an arbitrary distance criterion results in a non-random sample of plots with GPS-based area measures, (ii) multiply imputes missing GPS-based plot areas in two recent, public use household survey datasets from Uganda and Tanzania, and (iii) presents evidence, in context of the inverse scale-land productivity relationship, of why accounting for missing data in a statistically valid fashion matters at the analysis stage. We find that the missing GPS-based plot areas could be reliably simulated by MI and that the inverse scale-land productivity relationship is strengthened with the complete data. Our study demonstrates the use of judiciously reconstructed GPS-based areas in alleviating concerns over potential measurement error in farmer–reported area assessments, and with regards to systematic bias in plot selection for GPS-based area measurement.

The paper is organized as follows. Section 2 provides a literature review on (i) the use of GPS-based land area measurement in agricultural productivity analysis, and (ii) multiple imputation. Section 3 describes the data and provides summary statistics. Section 4 details the multiple imputation approach, followed by an empirical application focused on the inverse scale-land productivity relationship presented in Section 5. Section 6 concludes.

2. LITERATURE REVIEW

This paper falls at the intersection of two different strands of literature. The first strand relates to the improvements in methodological tools available to agricultural statisticians in the area of land measurement. The second concerns the advances in statistical procedures to deal with missing data.

Land area measurement is one of the fundamental components of agricultural statistics, which have traditionally been fraught with large measurement errors. It is therefore not surprising that the interest of agricultural statisticians in applying technological innovations such as satellite imagery and GPS technology to land area measurement is growing rapidly with the increasing affordability, precision and applications of these tools. Kelly et al. (1995) have long identified the use of GPS technology as having the potential to render land area measurement a significantly less costly and time-consuming exercise than traditional methods such as rope and compass. An early study mentioning a comparison of GPS-based and farmer-reported plot area measures is Goldstein and Udry (1999). The authors touch upon the issue as part of a study on agrarian innovation in the Eastern Region of Ghana, reporting a correlation coefficient of only 0.15 between the two variables. They attempt to reason the finding in the context of traditional regional land area measures being based only on length (i.e. ropes), and farmers struggling to think in terms of a two dimensional measure (i.e. hectares) with increasing land scarcity.

Keita and Carfagna (2009) provide a discussion on the performance of different GPS devices in comparison to the rope and compass alternative, which they assume to be the ‘gold standard’ for land area measurement. The authors document that while GPS technology allows for the measurement of 80 percent of the sample plot areas with negligible error, it tends to, on average, underestimate plot areas marginally with respect to the gold standard.[4] Possibly the most important problem with relying on GPS-based area measures, however, relates to the inability of obtaining these measures for all agricultural plots owned and/or cultivated by sample households. The inability is typically underlined by field logistics- and cost-related considerations. Some plots may be distant from the place of the interview (usually the sample household’s dwelling), and respondents may not have time or willingness to accompany the enumerators to these locations. Even if this constraint does not hold, the operational costs, in terms of time and financial resources, required to record GPS-based area measures for all sample plots are often perceived as prohibitive in most survey operations. This raises analytical concerns, in the form of biased estimates, if the plots that are not measured are systematically different from the ones whose areas are captured with GPS technology.

Carletto et al. (forthcoming) use the data from the Uganda National Household Survey (UNHS) 2005/06 to document (i) the determinants of the discrepancy in land areas when using GPS-based versus farmer-reported assessments, and (ii) how the discrepancy varies systematically with plot and farm size.[5] Although the primary focus of their study is to

[4] Keita and Carfagna (2009) posit that the plot slope, density of the surrounding tree canopy cover, and the weather conditions at the time of the measurement could be among the factors driving the observed discrepancies. However, their validation was also carried out in less-than-ideal conditions, resulting in a very small sample size.

[5] Carletto et al. (forthcoming) document that the magnitude of discrepancy between GPS-based and farmer-reported farm size measures increases monotonically while moving from the first

document whether the use of farm size measures based on GPS-based versus farmer-reported plot areas weakens, reverses or strengthens the inverse farm size-productivity relationship, they are not able to address the possible selection bias stemming from restricting their sample to households/farms that do not have any plots with missing GPS-based area measures (i.e. discarding approximately 50 percent of the agricultural household sample). Our study explicitly deals with this problem by employing approaches developed over the years by statisticians to overcome issues that are conceptually and practically similar in nature.

Since most datasets feature missing values for at least some variables, it is not surprising that finding new ways of dealing with missing data has attracted the attention of statisticians and applied economists for a long time, and in a number of different contexts. Unless the appropriate assumption regarding why the data are missing in the first place holds, any imputation method runs the risk of understating the true variance in the data, and leading to biased hypothesis testing and parameter estimates. The appropriateness of a imputation approach to fill in missing data, therefore, depends on (i) the nature of the missing data mechanism, (ii) whether the missing data mechanism could be ignored, and (iii) under what conditions.

The data are missing completely at random (MCAR) if the missingness depends on neither observed nor unobserved variables. The missing data on a given variable thus constitute a simple random sample of that variable. By ignoring missing data, as in casewise deletion, researchers implicitly assume that the data are MCAR. If the assumption holds and the researcher opts not to address missingness, the resulting parameter estimates are not biased but have larger standard errors stemming from the smaller sample size. If the MCAR assumption is not tenable, however, the available data are not representative of the population of interest and the resulting parameters are biased.

If the missingness depends on observed but not on unobserved factors, the data are missing at random (MAR), and could be predicted based on observed data underlying the missingness. If the MAR assumption holds and researchers undertake casewise deletion, the resulting parameter estimates are associated with larger standard errors and bias. If the MAR assumption is tenable but analysts conduct random or conditional mean imputation within an arbitrary imputation class that does not appropriately represent the observable attributes underlying the missingness, the subsequent parameter estimates are biased with understated standard errors (Robbins and White, 2011).

Finally, if the missingness depends on both observed and unobserved data, the data are missing not at random (MNAR). Since unobserved data that predict missingness are *nonignorable* under MNAR, relying only on observable covariates to simulate missing data yields biased inferences on parameters of interest. The resulting bias is not tractable unless additional information from outside the survey can be used to take into account unobserved heterogeneity that predicts missingness (Scheuren, 2005; Giusti and Little, 2011).

decile of the farm size distribution (-97 percent) to the tenth decile (19 percent). The bias is, on average, negative throughout the 60 percent of the farm size distribution, and becomes positive in the top three deciles, leading the authors to deduce that smaller farmers tend to over-report their land relatively more than larger farmers and that larger farmers tend to under-report land size.

Graham et al. (1997) argue that sound ways of dealing partially with missing data are better than doing nothing, and that the impact of the non-random component of the data on statistical inference is often smaller than it is commonly thought. In practice, it is close to impossible to determine whether missing data belong to the MNAR or MAR type, precisely because the missing data are, by definition, not observed. Scheuren (2005) provides an interesting discussion on the prevalence of the different types of missing data in practice, and on the advantages and disadvantages of different approaches to dealing (or not dealing) with them. In empirical analyses, imputations procedures are often applied based on the assumption that the data are MAR, sometimes accompanied by sensitivity analysis aimed at gauging the impact that departure from the MAR hypothesis may have on the analysis (see Giusti and Little (2011) for an example).

The conditions under which valid inferences could be obtained from missing data is laid out in Rubin's (1987) seminal work on multiple imputation (MI), which is a Monte Carlo technique that replaces missing values for a given variable by m>1 simulated alternatives. In repeated imputation inference, each of the m imputed datasets are analyzed separately, and the results are combined so that the uncertainty associated with missing data is incorporated into the computation of estimates and confidence intervals. While MI had originally been deemed as a viable strategy to disseminate *complete* datasets from sample surveys and censuses, it has evolved to be part of the toolkit of a diverse array of researchers from biomedical, social and behavioral sciences, whose analyses may otherwise be hindered by missing data.

MI assumes the missing data to be MAR, and 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) an *analytical model* is estimated 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, in accordance with Rubin (1987). Multiply imputed values are chosen to represent “both uncertainty about which values to impute assuming the reasons for nonresponse are known and uncertainty about the reasons for nonresponse.” (Rubin 1988: 79) MI also incorporates in the analysis the uncertainty associated with the imputation process itself, by incorporating in the statistical inference the variability across imputations.

A typical example in the literature that is relevant for our discussion of missing GPS-based land areas is missing income data in household surveys. Respondents often refuse to report personal/household income, and the likelihood of refusal tends to increase with wealth. The analogy with GPS-based land areas is clear: the larger the variable of interest (i.e. income, land area), the higher the probability that the data may be missing. Recent empirical work relying on MI to deal with missingness in income data clearly show that the mean estimate and the associated standard error for the imputed variable are affected by the choice of the imputation method, particularly when the proportion of missing values is large (Schenker et al., 2006; Zarnoch et al., 2010, Giusti and Little, 2011; Ahearn et al., 2011; Vermaak, 2012). All four studies also face a nonresponse rate similar in magnitude to ours, at around 30 percent. In what follows, we take a similar approach, and multiply impute missing GPS-based plot area measures in recent household survey data from Uganda and Tanzania. A further desirable feature of the survey data from both settings is that we have the farmer-reported plot area for all plots, which is an alternative measure of the variable we will be imputing and a luxury that no other study featuring MI to simulate missing data has enjoyed. Following MI, we utilize the multiply imputed dataset in the study of the relationship between agricultural productivity and cultivated plot area, and present the resulting parameter estimates alongside those that are obtained under casewise deletion.

The use of the multiply imputed dataset in this empirical application is particularly pertinent since it has been claimed that the long-standing debate on the inverse relationship between farm size and land productivity may rest on a statistical artefact tied to measurement error in land area estimates (Lamb, 2003). Barrett et al. (2010) provide a similar explanation after failing to explain the inverse relationship otherwise. Carletto et al. (forthcoming), on the other hand, demonstrate that the inverse relationship is strengthened by using the aggregated farm size measure underlined by GPS-based *vis à vis* farmer-reported plot areas. Their GPS-based farm size measure, however, suffers heavily from missingness as they exclude all farming households that had a missing GPS-based area for any of the parcels reported to be cultivated. Their inferences could therefore be inefficient and biased. Our study is the latest contribution to this debate by testing the robustness of the observed relationship in the context of complete case analysis featuring multiply imputed GPS-based plot areas.

3. DATA

The Tanzania National Panel Survey (TZNPS) 2010/11 and the Uganda National Panel Survey (UNPS) 2009/10, which were implemented by the Tanzania National Bureau of Statistics and the Uganda Bureau of Statistics, respectively, inform our analyses.[6,7] The TZNPS 2010/11 sample size was 3,924 households, and the domains of inference include the major agro-ecological zones, Dar es Salaam, other mainland urban areas, mainland rural areas, and Zanzibar. The sample households were visited once during the 12-month survey round. The UNPS, on the other hand, has been designed to re-interview 3,123 households that had been interviewed by the Uganda National Household Survey (UNHS) 2005/06. The UNPS 2009/10 sample size was 2,975 households, and the domains of inference include Kampala City, Other Urban Areas, Central Rural, Eastern Rural, Western Rural, and Northern Rural. The sample households were visited twice during a given survey round.

In terms of questionnaire design, the TZNPS and the UNPS share a number of similar features. In each setting, all sample households were administered a multi-topic household questionnaire. The households that were involved in agricultural activities (through ownership and/or cultivation of land and/or ownership of livestock) were also given an agriculture questionnaire. On annual/temporary crop production, the same set of modules were administered separately for the long and short rainy seasons in a single visit in Tanzania, and for the two main agricultural seasons in two different visits in the case of Uganda. At the plot-level, the questionnaires solicited detailed information on

[6] The TZNPS and the UNPS are conducted with technical and financial support from the Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) initiative, whose primary purpose is to support governments in sub-Saharan Africa in the design and implementation of nationally-representative, multi-topic panel household surveys with a strong focus on agriculture (www.worldbank.org/lsms-isa). The data and documentation for the TZNPS, the UNPS and other survey efforts supported under the LSMS-ISA initiative are publically available (www.worldbank.org/lsms).

[7] The first round of the TZNPS was implemented in 2008/09 when 25 percent of agricultural plot areas were measured by handheld GPS units on an experimental basis. The GPS-based plot area measurement was fully adopted in the 2010/11 round.

land area, physical characteristics, labor and non-labor input use, cultivation, and production.

Both surveys relied on mobile survey teams, each of which was headed by a team leader and was composed of 3-4 enumerators and a data entry operator. In Tanzania, as long as respondent refusal or physical inaccessibility (due to inclement weather or unfavorable road conditions) were non-issues, the field protocol required enumerators to record GPS-based area measurements for agricultural plots that were owned and/or cultivated by the farming household sample that were situated within a 1 hour of travel radius with respect to the dwelling unit, regardless of mode of transportation. Similarly, the field protocol in Uganda required enumerators to take GPS-based plot area measurements as long as the plots were confirmed to be located within the sample enumeration area.

In both settings, the GPS-based plot area measurements were taken following the interview, with the help of respondents that accompanied enumerators to plot locations and identified the plot boundaries. The enumerators were instructed to report non-measured plots and the associated reasons to their team supervisors. Each enumerator was assigned a handheld GPS unit (Garmin eTrex HC in Tanzania, and Garmin 12 in Uganda) and was trained extensively on GPS-based area measurement during the field staff training prior to the start of the field work. For a given plot, the enumerator was supposed to walk clockwise the perimeter with the GPS unit active and at a reasonably slow speed, stopping for 10 seconds at every point that the plot outline changed direction. The area of each plot was calculated directly by the GPS unit in acres.

Our field experience and interactions with the survey teams in both settings, as well as the quantitative evidence on the reasons for not obtaining GPS-based plot areas, suggest that the missing data are overwhelmingly due to the aforementioned distance criteria for the selection of plots for GPS-based area measurement. Refusal (and physical accessibility) constitute a near-negligible share of reasons for the lack of GPS-based area measures in each setting, in line with the high response rates observed in rural East African settings.

Tables 1 and 2 present summary statistics based on the UNPS and TZNPS, respectively, and report the results from the tests of average differences by GPS-based plot area measurement status. Given the focus of the empirical application on agricultural productivity, we focus on cultivated plots in both samples. A number of noteworthy findings emerge. First, the number of plots lacking a GPS-based area measure is relatively large, with 1,519 and 759 missing data points accounting for 35 and 18 percent of the plot samples in Uganda and Tanzania, respectively. Second, the GPS-based and farmer-reported plot area estimates of average plot area are similar, and just over 2 acres in both countries. Average farmer-reported area is slightly larger for plots that were not measured, but the differences are not statistically significant. Third, several important plot- and household-level characteristics, which are expected to be associated with productivity-related outcomes, display statistically significant differences by GPS-based plot area measurement status. Taken together, 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.

Plots without a GPS-based area measure are clearly non-random picks and they differ from their GPS-measured counterparts in similar ways in both Uganda and Tanzania. In line with the expectations informed by the field work protocols, non-measured plots tend to be further away from the household location. They are additionally more likely to be rented and to lack desirable features in terms of soil quality and slope. Plots with a GPS-based area measure originate, on average, from households that are headed by older individuals who are more likely to report agriculture as their primary occupation. In Uganda, heads of households associated with plots with a GPS-based area measure are more likely to be female, while their counterparts in Tanzania have, on average, fewer years of education. Lastly, non-measured plots in both settings are more likely to stem from (intact) mover original and split-off households, as opposed to non-mover households re-interviewed in the TZNPS 2010/11 and the UNPS 2009/10. Given the additional workload associated with tracking these households in their new locations, and the possibility that households moving into distant areas might still keep their plots in their original locations, which in turn may not be GPS-measured in accordance with the survey protocol, this finding is also in line with our expectations.

The descriptive findings[8] reported thus far confirm our initial suspicion that the GPS-based plot area measures cannot be considered MCAR, and indicate that there are distinct observable attributes associated with missingness, and that the results of statistical analyses based on their use may not only be inefficient (due to the smaller sample size), but also biased (due to the non-random missing pattern). Together with the field insights regarding the missing data being overwhelmingly driven by the survey protocols that excluded plots from GPS-based area measurement based on distance criteria, we hypothesize that the data are to a large extent MAR, and that MI could be applied in a statistically valid fashion to predict missing GPS-based plot areas based on observables. The support to this hypothesis is also rooted in (i) the high degree of correlation between farmer-reported and GPS-based plot areas that will be central to the specification of the imputation model discussed in the subsequent section, and (ii) the fact that even if the data are partly MNAR, MI will still reduce the bias in comparison to casewise deletion or conditional mean imputation within an arbitrary imputation class (Graham, et al., 1997).

4. MULTIPLE IMPUTATION OF GPS-BASED PLOT AREAS[9]

In building the imputation model, the literature (Rubin, 1996; 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.

Our imputation model follows the aforementioned guidelines, and includes the variables that relate to missingness, and all variables featured in the empirical application presented in Section 5, including sampling weights, and enumerator and district fixed effects. A key covariate that is included in the imputation model and that is both a powerful predictor and an alternative measure of the GPS-based plot area is the farmer-reported plot area.

[8] A multivariate Probit model with a dependent variable identifying whether or not a GPS-based plot area measure is missing yields associations that are qualitatively similar to those that have been highlighted thus far.

[9] The *mi impute* command in Stata 12 is used to multiply impute the missing data in our study.

The availability of this variable distinguishes our study from other studies that have employed MI to tackle item non-response. The availability of the farmer-reported plot area allows us to overcome what could otherwise be a limitation of this application, namely our inability to rely on a panel of plots. In the literature, the multiple imputation procedure is most often applied to longitudinal data, where variables from different survey rounds can be strong predictors of missing observations, and it is thought to be at "its weakest in cross-section surveys. In cross-section surveys, seldom are there strong predictors present" (Scheuren, 2005: 318). We posit that the presence of the self-reported land area makes our dataset an exception to this rule.[10]

Other independent variables in the imputation model include the distance of the plot from the household in minutes (Uganda) or kilometers (Tanzania). As noted earlier, these covariates are major predictors of missingness in accordance with the survey protocols. The variables on plot tenure status and farmer-assessed plot slope and soil quality are incorporated since, besides being control variables in the analytical model, they are possible predictors of missingness as well. The imputation model also includes variables on household characteristics related to (i) demographics (household size, dependency ratio, and the age, sex and education of the head of household), and (ii) whether the household is a (intact) mover original or a split-off household. The inclusion of the latter set of variables attempts to accommodate the possibility that the additional workload associated with tracking shifted households and individuals in their new locations may increase the likelihood of missing data. The demographics are deemed to influence the accuracy of the self-reporting, as well as being likely control variables for the analysis. Furthermore, the inclusion of enumerator fixed effects in the imputation model accounts for unobserved heterogeneity across enumerators, such as differences in effort, which might be correlated with the missingness in GPS-based plot area measures. Lastly, we use information from earlier survey rounds on household wealth to accommodate the possibility of higher likelihood of missing plot area measurements among richer households with higher opportunity costs of time.

The first step in our multiple imputation effort is to fit a plot-level Ordinary Least Squares (OLS) regression model with the GPS-based plot area as the dependent variable, to then obtain linear predictions for all plots in the dataset. Under the partially parametric method of predictive mean matching (PMM) (Rubin, 1987; Little, 1988; Schenker and Taylor, 1996), 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.[11] The imputation is carried out 50 times to reduce the potential sampling error due to imputation, and 50 complete datasets

[10] We cannot apply a panel data model because of issues with the survey design that do not allow matching plots across survey years. It is likely that future surveys in the LSMS-ISA program will allow us to overcome this limitation. Another aforementioned limitation in the case of the TZNPS is that only 25% of the plots were measured in the most recent survey round prior to the 2010/11 wave that informs our analysis.

[11] 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.

are generated.[12] The posterior estimates of the model parameters are obtained using sampling with replacement, which is standard practice when the asymptotic normality of parameter estimates is suspect.[13] 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 linear regression approach.

For the analysis of the multiply imputed data, we follow Rubin (1987), and suppose that there are m imputations, thereby m completed datasets, and let Q denote a scalar population parameter of interest. Estimation of the analysis model with the complete data from a multiply imputed data set i yields the point estimate $\widetilde{q}_{\mathrm{i}}$ and its estimated variance v_i, where i = 1, 2,…, m.

The overall multiple imputation estimate is defined as:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\widetilde{q}_{\mathrm{i}}$$

The variance of $\bar{Q}$ has two components. The average within-imputation variance is:

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m}v_i$$

The between-imputation variance is:

$$B == \frac{1}{m-1}\sum_{i=1}^{m}(\widetilde{q}_{\mathrm{i}} - \bar{Q})^2$$

The total variance is then specified as:

$$T = \bar{V} + \left(1 + \frac{1}{m}\right)B$$

The Student t approximation is used for constructing confidence intervals and significance tests, where:

$$t_{df} = \frac{\bar{Q}}{\sqrt{T}}$$

with degrees of freedom:

$$df = (m-1)(1 + \frac{m\bar{V}}{(m+1)B})^2$$

[12] The results are robust to carrying out 100, 150, 200, or 250 imputations.

[13] The results are robust to sampling estimates from the posterior distribution of model parameters, as opposed to bootstrapping.

Finally, the fraction of information about Q that is missing due to nonresponse is:

$$\gamma = [\frac{(m+1)}{m}]\frac{B}{T}$$

The results from an illustrative OLS regression that would have underlined MI in the absence of the bootstrapping option for the imputation model are reported in Tables 3 and 4. The models perform well in explaining the variance in the GPS-based plot area, with an R-square of 0.658 in Uganda and 0.688 in Tanzania. The coefficients on the farmer-reported plot area are highly significant and large, with a similar order of magnitude (0.945 in Uganda and 0.866 in Tanzania). 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 (Schafer and Graham, 2002). Instead, it attempts to be as comprehensive as possible in order 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: 479)

Table 5 provides separately for the UNPS 2009/10 and the TZNPS 2010/11 the summary statistics on (i) GPS-based and farmer-reported plot area for plots with an observed GPS-based area measure, and (ii) multiply-imputed GPS-based and farmer-reported plot area for the entire plot sample. The mean and the associated standard error for the multiply-imputed GPS-based plot area are computed in accordance with Rubin's rules. Looking at the *overall* multiply-imputed mean GPS-based area and the associated standard error for the entire plot sample in Uganda and Tanzania, we observe that they are in line with the mean GPS-based area and the associated standard error for measured plots. While MI, in comparison to casewise deletion, leads to marginally lower mean and standard error estimates for GPS-based plot area, a simpler method of handling missing data, such as conditional mean imputation, would have understated the true variance in the variable of interest (Schafer and Graham, 2002).[14] Moreover, the mean value and associated standard error for our land productivity measure, namely plot-level gross value of output per acre in local currency, exhibit more drastic changes under complete case analysis following MI in comparison to the incomplete case analysis based on plots with an observed GPS-based area measure. The next section delves deeper into the empirical implications of incomplete vs. multiply imputed complete case analysis of GPS-based plot areas in the context of the inverse scale-land productivity relationship.

[14] We conducted conditional mean imputation for missing GPS-based plot areas in each setting, replacing missing values by the EA mean, provided that there were at least ten GPS-based plot areas in a given EA. In case, the criterion on the EA-level observation count was not satisfied, a coarser grouping with at least ten GPS-based plot areas is used, in order of district, region, and nation. While the resulting mean for GPS-based plot areas were similar to those obtained under casewise deletion as well as MI, the associated standard errors were consistently lower, in accordance with the expectations. These results, while available upon request, have not been reported given the lack of a statistical basis for choosing any particular sample size (Graham et al., 1997). After all, "one imputed value cannot itself represent any uncertainty about which value to impute: if one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reasons for nonresponse are known." (Rubin, 1988:79)

5. EMPIRICAL APPLICATION: INVERSE SCALE-LAND PRODUCTIVITY RELATIONSHIP

Starting with the seminal work of Sen (1962, 1966), who observed an inverse relationship (IR) between farm size and output per hectare in Indian agriculture, a large number of empirical studies have presented evidence that appears to corroborate that hypothesis (Lau and Yotopoulos, 1971; Yotopoulos and Lau, 1973; Berry and Cline, 1979; Carter, 1984, Eswaran and Kotwal, 1985, 1986; Barrett, 1996; Benjamin and Brandt, 2002; Larson et al. 2012). A smaller set of empirical studies, however, does not find evidence of such a relationship (Hill, 1972; Kevane, 1996; Zaibet and Dunn, 1998). Binswanger et al. (1995) and Eastwood et al. (2010) provide careful discussions of both the theory and the empirics of the IR debate, a full review of which is beyond the scope of this paper.

Following Barrett et al. (2010), an inverse relation between farm size and land productivity may have three main explanations: (i) imperfect factor markets, (ii) omitted variables and, in particular, omitted controls for land quality, and (iii) statistical issues related to the measurement of plot size. Imperfect factor markets (i.e. labor, land, insurance) are linked to differences in the shadow price of production factors that in turn lead to differences in the application of inputs per unit of land, in ways that are correlated with farm size.

Much of the earlier contributions to the IR debate focused on testing this type of explanation. Assunçao and Ghatak (2003) demonstrate how unobserved heterogeneity in farmer ability may theoretically explain the observed differences in productivity. On one hand, Bhalla and Roy (1988) and Benjamin (1995) have challenged the existence of the IR based on the observation that when land quality controls are introduced in the analysis, the strength of the IR often diminishes substantially or vanishes altogether. On the other hand, Barrett et al. (2010) utilize a dataset that includes laboratory measures of soil testing and conclude that only a very limited proportion of the IR can be explained by differences in land quality. Similarly, Nkonya et al. (2004) find a strong negative effect of farm size on plot output value after controlling for plot size, labor input, equipment and proxies for land quality, suggesting that not only land productivity but also total factor productivity is inversely related to scale. Carter (1984) and Heltberg (1998) control for village and household fixed effects, respectively, and still do not find that the strength of the IR diminishes. Taking these studies together, Eastwood et al. (2010: 3351) consequently posit that “the IR is immune to the land-quality objection.”

Lastly, it has been suggested that the existence of the IR may be a statistical artefact stemming from measurement error in land data (Lamb, 2003). A similar explanation is provided by Barrett et al. (2010), after failing to explain the observed IR otherwise. However, the data from the UNHS 2005/06, which collected both GPS-based and farmer-reported measures of plot areas, have been used by Carletto et al. (forthcoming) to show that the estimates of the IR are robust to potential measurement error introduced by farmer’s self-reporting and that the IR is strengthened when more accurate, GPS-based measure of farm size is used in the analysis.

Our study builds on the work of Carletto et al. (forthcoming), who exclude all farming households for whom GPS-based farm size cannot be computed due to one or more missing GPS-based plot areas. As a result, approximately 50 percent of the agricultural household sample is discarded in their analysis, even if the share of non-measured plots stands at 35 percent. In what follows, we investigate the robustness of the inverse scale-land productivity relationship to complete case analysis made possible by MI. We estimate a plot-level production function to test for the existence of the IR, based on the one originally proposed by Binswanger, Deininger and Feder (1995), and not dissimilar from the approach taken by Barrett et al. (2010) and Carletto et al. (forthcoming):

$$\ln \frac{Y_{ih}}{A_{ih}} = \alpha + \beta ln\, A_{ih} + \gamma\, P_{ih} + \ \delta\, H_{ih} + \partial D_{ih} + \varphi E_{ih} + \ \varepsilon_{ih} \quad (1)$$

where *i* and *h* denote plot and household, respectively; *Y* represents value of agricultural output; A denotes the plot area in acres; P is a vector of plot-level variables spanning logarithm of value of non-labor input use per acre, distance from the household, tenure status, and farmer-reported soil quality and slope; H is a vector of household-level attributes that might influence agricultural production, including household size, dependency ratio, and basic characteristics of the head of household; *D* and *E* are district and enumerator fixed effects, respectively; and ε is the error term.[15]

The coefficient β on A is the parameter of interest while testing for the existence of the IR. A negative β indicates the presence of the IR: the more negative the coefficient, the stronger the IR. We estimate Equation 1 separately with (i) the incomplete dataset using the observed GPS-based plot areas, and (ii) the complete dataset using the multiply imputed GPS-based plot areas. In the latter case, the regression is estimated m (i.e. 50) times with each of the m complete datasets, and the coefficient estimates and associated standard errors are combined in accordance with the framework presented in Section 5.

The regression results are presented in Tables 6 and 7 for the UNPS 2009/10 and the TZNPS 2010/11, respectively.[16, 17] The results from the incomplete case analysis and the multiply-imputed complete case analysis are presented in Columns 1 and 2, respectively. The coefficient on the plot area is negative and statistically significant at the 1 percent level across all columns reported in Tables 6 and 7. However, with respect to the β obtained under the incomplete case analysis, the β associated with multiply-imputed GPS-based plot area is 33 percent and 9 percent higher in Uganda, and Tanzania, respectively, revealing how the selection bias introduced in the GPS-based plot area measurement process carries all the way to the analysis stage. It should also be noted that not only does working with the full sample have implications for the IR, the coefficients and standard errors of several other explanatory variables of interest are observed to change when the complete dataset is restored. Notable in this respect are the changes associated with the distance variables in Tanzania, the rented/other plot tenure category in Uganda, and the household size and the number of plots in the holding in both settings.

[15] Our results are robust to (i) using profit (i.e. gross value of output net of gross value of non-labor input) per acre as the dependent variable in the analysis model, and/or (ii) including enumeration area, as opposed to district, fixed effects in the imputation and analysis models.

[16] Using the sample of plots with observed GPS-based areas, we also estimated Equation 1 using the farmer-reported plot areas in both settings. Just as in Carletto et al. (forthcoming), the resulting β is negative and its absolute value is lower than the β obtained by using the observed GPS-based plot areas. These results are available from the authors upon request, but are not reported as they are not the focus of this paper.

[17] The combined MI results reported in Tables 6 and 7 are underlined by complex survey regressions that correctly take into account clustering and stratification in weighted regressions. The results are robust to not weighting our regressions.

6. CONCLUSION

This paper attempts to address three interrelated questions concerning the use of the GPS technology for land area measurement in household surveys in developing countries: (i) Are the land area statistics informed by partial GPS-based land areas subject to selection bias?, (ii) Is it possible to fill the gaps in GPS-based land areas by using statistically valid techniques and other information available in the survey?, and (iii) Does having *complete* GPS-based land areas matter for the analysis of policy relevant issues, such as the inverse scale-land productivity relationship? Our analysis confirms that the missingness in GPS-based plot area measures is a serious concern, not only because it is pervasive, but also because plots that do not get measured are not random picks among plots that households own or cultivate. This is particularly concerning in the context of large, nationally-representative household surveys, which suggests that the prevailing survey protocols for GPS-based plot area measurement and its supervision should be revisited to minimize the extent to which GPS-based plot area measurements are not taken.

However, it is important to recognize that no matter how effective the protocols and the field supervision could be, a non-negligible degree of missingness in GPS-based plot areas data is bound to remain, as completely eliminating it is unfeasible without incurring prohibitive costs. This paper has, therefore, sought not only to advance our understanding of the limitations of current GPS-based land area data solicited as part of household surveys, but also to explore ways to remedy some of these shortcomings through sound statistical techniques. The good news is that the missingness is largely driven by observable variables. The better news is that we have the complete farmer-reported plot areas that are powerful predictors of the observed GPS-based plot areas and that could be used in generating reliable simulations of missing GPS-based plot areas using multiple imputation (MI). Advances in computing power and the availability of multiple imputation routines in an increasing number of statistical software packages make the routine adoption and application of these techniques feasible in developing country settings.

This is very timely as policies enabling open access to data are spreading quickly across the world, and GPS technology is becoming a standard feature in survey work. Having data producers, as opposed to data analysts, deal with handling missing data has the advantage of (i) reducing the duplication of work, (ii) increasing the comparability between the final data used to perform analyses, and (iii) allowing more analysts (who do not necessarily have the time or ability to deal with the missing data problem correctly) to work with complete case data. Regardless, it is essential to document the imputation approach as part of the survey documentation.

Obtaining complete case GPS measures at the fieldwork stage is overly costly, as some plots will always be located at a prohibitive distance from the location of the interview. On the other hand, GPS-based plot area measures exhibiting up to 35 percent of missing values would be of limited value to many data analysts. By combining good fieldwork training, strict fieldwork quality control, sound imputation methods, and emphasis on always soliciting farmer-reported plot areas as possible predictors for missing GPS-based counterparts, it is possible to obtain reliable simulations of GPS-based plot areas while keeping survey costs in check. One limitation of our results is that they are based on two cases from East Africa. Further applications in different geographical, cultural, and socio-economic contexts are needed to establish the extent to which our findings are generalizable on a broader scale.

Last but not least, our findings show how the collection of GPS-based land areas matters for advancing the debate on key policy-related research questions, such as the longstanding debate on the existence of an inverse relationship (IR) between farm size and land productivity. When (complete) GPS-based plot area information is utilized, the presence of the IR is confirmed and strengthened, and MI allows us to overcome the limitations to the power of that analysis and the concerns over potential selection bias rooted in the available GPS-based measures.

References

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

Assunçao, J., & Ghatak, M. (2003). "Can unobserved heterogeneity in farmer ability explain the inverse relationship between farm size and productivity?" *Economics Letters*, 80(2), 189-194.

Barrett, C. (1996). "On price risk and the inverse farm size–productivity relationship." *Journal of Development Economics*, 51(2), 193–215.

Barrett, C., Bellemare, M., & Hou, J. (2010). "Reconsidering conventional explanations of the inverse productivity–size relationship." *World Development*, 38(1), 88–97.

Benjamin, D. (1995). "Can unobserved land quality explain the inverse productivity relationship?" *Journal of Development Economics*, 46, 51-84.

Benjamin, D., & Brandt, L. (2002). "Property rights, labor markets, and efficiency in a transition economy: the case of rural China." *Canadian Journal of Economics*, 35(4), 689–716.

Berger, S. (1934). "A note on some scenes of land-measurement." *The Journal of Egyptian Archaeology*, 20, 54-56.

Berry, R., & Cline, W. (1979). *Agrarian structure and productivity in developing countries*. Baltimore, MD: Johns Hopkins University Press.

Bhalla, S., & Roy, P. (1988). "Mis-specification in farm productivity analysis: the role of land quality." *Oxford Economic Papers*, 40, 55–73.

Binswanger, H., Deininger, K., & Feder, G. (1995). "Power, distortions, revolt and reform in agricultural land relations." In J. Behrman, & T.N. Srinivasan (Eds.), *Handbook of Development Economics*, Vol. 3, Amsterdam: Elsevier.

Carletto, C., Savastano, S., & Zezza, A. (Forthcoming). "Fact of artefact: the impact of measurement errors on the farm size – productivity relationship." *Journal of Development Economics*.

Carter, M. (1984). "Identification of the inverse relationship between farm size and productivity: an empirical analysis of peasant agricultural production." *Oxford Economic Papers*, 36, 131–145.

Dorward, A. & Chirwa, E. (2010). "A review of methods for estimating yield and production impacts" Unpublished paper. Retrieved on March 6, 2013 from
http://www.wadonda.com/Dorward_Chirwa_2010_FISP_ProdMethodologies.pdf

Eastwood, R., Lipton, M., & Newell, A. (2010). "Farm size," In R. Evenson, & P. Pingali (Eds.), *Handbook of Agricultural Economics*, Volume 4, Amsterdam: Elsevier.

Eswaran, M., & Kotwal, A. (1985). "A theory of contractual structure in agriculture." *American Economic Review*, 75, 352-367.

________________________________ (1986). "Access to capital and agrarian production organization." *Economic Journal*, 96, 482-498.

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

Goldstein, M., Udry C., 1999. Agricultural Innovation and Risk Management in Ghana. Unpublished Final Report to International Food Policy Research Institute (IFPRI).

Graham, J., Hofer, S., Donaldson, S., MacKinnon, D., & Schafer, J. (1997). "Analysis with missing data in prevention research." In K. Bryant, M. Windle, & S. West (Eds.), *The Science of Prevention: Methodological Advances from Alcohol and*

Substance Abuse Research. Washington, D.C.: American Psychological Association.

Heltberg. R. (1998). "Rural market imperfections and the farm size-productivity relationship: evidence from Pakistan." World Development, 26(10), 1807-1826.

Hill, P. (1972). *Rural Hausa: a village and a setting*. Cambridge, UK: Cambridge University Press.

Keita N., & Carfagna, E. (2009). "Use of modern geo-positioning devices in agricultural censuses and surveys." Bulletin of the International Statistical Institute, the 57th Session, Proceedings, Special Topics Contributed Paper Meetings (STCPM22), Durban, August 16-22.

Kelly, V., Diagana, B., Reardon, T., Gaye, M., & Crawford, E. (1995). "Cash crop and foodgrain productivity in Senegal: historical view, new survey evidence, and policy implications." Michigan State University Staff Paper No. 95-05.

Kevane, M. (1996). "Agrarian structure and agricultural practice: Typology and application to Western Sudan." *American Journal of Agricultural Economics*, 78(1), 236–245.

Lamb, R. (2003). "Inverse productivity: Land quality, labor markets, and measurement error." *Journal of Development Economics*, 71(1), 71–95.

Larson, D., Otsuka, K., Matsumoto, T., & Kilic, T. (2012) "Should African rural development strategies depend on smallholder farms? An exploration of the inverse productivity hypothesis."World Bank Policy Research Working Paper No. 6190.

Lau, L., & Yotopoulos, P. (1971). "A test for relative efficiency and application to Indian agriculture." *American Economic Review*, 61, 92-109.

Little, R. (1988). "Robust estimation of the mean and covariance matrix from data with missing values." *Applied Statistics*, 37, 23-38.

Lyons, H. (1927). "Ancient surveying instruments." *The Geographical Journal*, 69, 132-145.

Nkonya, E., Pender, J., Jagger, P., Sserunkuuma, D., Kaizzi, C., & Ssali, H. (2004). "Strategies for sustainable land management and poverty reduction in Uganda." Research Report 133. Washington, DC: IFPRI.

Robbins, M. & White, T. (2011). "Farm commodity payments and imputation in the Agricultural Resource Management Survey." *American Journal of Agricultural Economics*, 93, 606-612.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons.

__________ (1988). "An overview of multiple imputation." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-84.

__________ (1996). "Multiple imputation after 18+ years." *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J., & Graham, J. (2002). "Missing data: our view of the state of the art." *Psychological Methods*, 7, 147–177.

Sen, A. K. (1962). "An aspect of Indian agriculture." *Economic Weekly*, 14, 243–266.

__________ (1966). "Peasants and dualism with or without surplus labor." *Journal of Political Economy*, 74(5), 425–450.

Schenker, N., & Taylor, J. (1996). "Partially parametric techniques for multiple imputation." *Computational Statistics & Data Analysis*, 22, 425-446.

Schenker, N., Raghunathan, T., Chiu, P.-L., Makuc, D., Zhang, G., & Cohen, A. (2006). "Multiple imputation of missing income data in the National Health Interview Survey." *Journal of the American Statistical Association*, 101, 924-933.

Scheuren, F. (2005). “Multiple imputation: how it began and continues.” *The American Statistician*, 59(4), 315-319.

Schoning, P., Apuuli, J., Menyha, E., & Zake-Muwanga, E. (2005). “Handheld GPS equipment for agricultural statistics surveys: experiments on area-measurement and geo-referencing of holdings done during fieldwork for the Uganda Pilot Census of Agriculture.” Statistics Norway Report 2005/29.

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

Vermaak, C. (2012) “Tracking poverty with coarse data: evidence from South Africa.” *Journal of Economic Inequality*, 10(2), 239-265.

Yotopoulos, P. & Lau, L. (1973). “A test for relative efficiency: some further results.” *American Economic Review*, 61, 92-109.

Zaibet, L., & Dunn E. (1998). “Land tenure, farm size, and rural market participation in developing countries: the case of the Tunisian olive sector.” *Economic Development and Cultural Change*, 46(4), 831-848.

Zarnoch, S., Cordell, H., Betz C., & Bergstrom, J. (2010) “Multiple imputation: an application to income nonresponse in the National Survey on Recreation and the Environment.” United States Department of Agriculture Forest Service Research, Paper SRS–49.

Table 1: Averages by Plot GPS Measurement Status & Results from Tests of Mean Differences
Data: Uganda National Panel Survey (UNPS) 2009/10

	Entire Sample	*W/ GPS Measurement*	*W/o GPS Measurement*	*Difference*	
Plot Area					
GPS-Based (Acres)	2.13	2.13	--	--	
Farmer-Reported (Acres)	2.05	2.00	2.12	-0.12	
Plot Output					
Value of Total Output	471,783	532,708	367,645	165,063	***
Plot Input Use					
Value of Total Inputs	89,824	89,870	89,746	124	
Plot Location (Farmer-Assessed)					
Less Than 15 Mins Away from Household †	0.62	0.80	0.31	0.49	***
15-30 Mins Away from Household †	0.17	0.14	0.21	-0.07	***
30+ Mins Away from Household †	0.22	0.06	0.48	-0.42	***
Plot Tenure					
Owned w/ Title †	0.10	0.12	0.06	0.06	***
Owned w/o Title †	0.64	0.74	0.48	0.26	***
Rented/Other †	0.26	0.14	0.46	-0.32	***
Plot Soil Quality (Farmer-Assessed)					
Good †	0.60	0.57	0.66	-0.09	***
Fair †	0.32	0.34	0.28	0.06	***
Poor †	0.08	0.09	0.06	0.04	***
Plot Slope (Farmer-Assessed)					
Flat †	0.44	0.43	0.46	-0.03	
Gentle †	0.36	0.40	0.30	0.11	***
Hilly, Steep or Valley †	0.20	0.17	0.25	-0.08	***
Household Characteristics					
Household Size	6.22	6.18	6.28	-0.10	
Dependency Ratio	1.39	1.43	1.32	0.11	**
Head of Household: Female †	0.26	0.28	0.22	0.06	***
Head of Household: Years of Age	45.06	46.07	43.33	2.74	***
Head of Household: Years of Education	5.37	5.29	5.51	-0.22	
Head of Household: Primary Occupation: Agriculture †	0.68	0.72	0.62	0.10	***
# of Plots in Holding	3.31	3.18	3.54	-0.36	***
Non-Mover Original Household †	0.82	0.92	0.65	0.27	***
Mover Original Household †	0.04	0.01	0.09	-0.08	***
Split-Off Household †	0.13	0.06	0.26	-0.19	***
Wealth Index 2005/06	-0.65	-0.76	-0.46	-0.29	***
Observations	4,333	2,814 (65%)	1,519 (35%)		

Note: *** p<0.01, ** p<0.05, * p<0.1. † indicates a dummy variable. The statistics are weighted through the use of household sampling weights, and are based on UNPS 2009/10 data, unless otherwise stated.

Table 2: Averages by Plot GPS Measurement Status & Results from Tests of Mean Differences
Data: Tanzania National Panel Survey (TZNPS) 2010/11

	Entire Sample	*W/ GPS Measurement*	*W/o GPS Measurement*	*Difference*	
Plot Area					
Observed GPS-Based (Acres)	2.59	2.59	--		
Farmer-Reported (Acres)	2.31	2.30	2.35	-0.05	
Plot Output					
Value of Total Output	90,012	94,624	63,722	30,902	**
Plot Input Use					
Value of Total Inputs	65,855	72,358	28,784	43,574	*
Plot Location (Farmer-Assessed)					
Distance to Home (KM)	3.74	1.95	13.92	-11.97	***
Distance to Road (KM)	2.18	1.62	5.39	-3.76	***
Plot Tenure					
Owned w/ Title †	0.10	0.11	0.07	0.03	*
Owned w/o Title †	0.78	0.80	0.67	0.13	***
Rented/Other †	0.12	0.09	0.25	-0.16	***
Plot Soil Quality (Farmer-Assessed)					
Good †	0.45	0.45	0.45	0.00	
Average †	0.47	0.47	0.49	-0.02	
Bad †	0.07	0.08	0.06	0.02	*
Plot Slope (Farmer-Assessed)					
Flat †	0.55	0.56	0.53	0.03	
Slight †	0.33	0.33	0.34	-0.02	
Steep †	0.12	0.12	0.12	-0.01	
Household Characteristics					
Household Size	5.88	5.85	6.08	-0.23	
Dependency Ratio	1.23	1.24	1.20	0.03	
Head of Household: Female †	0.22	0.22	0.21	0.01	
Head of Household: Years of Age	48.81	49.18	46.69	2.49	***
Head of Household: Years of Education	4.78	4.69	5.27	-0.58	***
Head of Household: Primary Occupation: Agriculture †	0.89	0.90	0.86	0.04	*
# of Plots in Holding	3.09	3.08	3.15	-0.07	
Non-Mover Original Household †	0.85	0.86	0.76	0.10	***
Mover Original Household †	0.06	0.05	0.09	-0.03	**
Split-Off Household †	0.09	0.08	0.15	-0.07	***
Wealth Index 2008/09	-1.06	-1.09	-0.88	-0.21	**
Observations	4,142	3,383 (82%)	759 (18%)		

Note: *** p<0.01, ** p<0.05, * p<0.1. † indicates a dummy variable. The statistics are weighted through the use of household sampling weights, and are based on TZNPS 2010/11 data, unless otherwise stated.

Table 3: OLS Regression Results
Data: UNPS 2009/10
Dependent Variable: Observed GPS-Based Plot Area (Acres)

Plot Area	
Farmer-Reported Plot Area (Acres)	0.945***
	(0.016)
Plot Characteristics	
Log [Value of Plot Output (UShs)]	0.023
	(0.023)
Log [Value of Total Plot Input (UShs)]	0.027**
	(0.012)
15-30 Mins Away from Household † *A*	-0.296*
	(0.152)
30+ Mins Away from Household † *A*	-0.080
	(0.220)
Owned w/o Title † *B*	0.172
	(0.198)
Rented/Other † B	0.115
	(0.232)
Soil Quality: Fair † *C*	0.024
	(0.115)
Soil Quality: Poor † *C*	0.346*
	(0.195)
Slope: Hilly, Steep or Valley †	-0.021
	(0.166)
Household Characteristics	
Household Size	0.026
	(0.020)
Dependency Ratio	-0.024
	(0.050)
Head of Household: Female †	0.000
	(0.133)
Head of Household: Years of Age	0.000
	(0.004)
Head of Household: Years of Education	-0.000
	(0.014)
Head of Household: Primary Occupation: Agriculture †	0.018
	(0.124)
# of Plots in Holding	-0.141***
	(0.032)
Mover Original Household † D	-0.015
	(0.422)
Split-Off Household † D	0.622
	(0.841)
Wealth Index 2005/06	0.047
	(0.052)
Observations	2,814
R2	0.658

Note: *** p<0.01, ** p<0.05, * p<0.1. Constant is estimated but not reported. District and enumerator fixed effects as well as sampling weights are included as covariates but not reported. † denotes a dummy variable. The comparison categories are (i) less than 15 minutes, (ii) owned with a title, (iii) good, and (iv) non-mover original household for A, B, C, and D, respectively. All variables based on the UNPS 2009/10 data, unless otherwise stated.

Table 4: OLS Regression Results
Data: TZNPS 2010/11
Dependent Variable: Observed GPS-Based Plot Area (Acres)

Plot Area	
Farmer Reported Plot Area (Acres)	0.866***
	(0.014)
Plot Characteristics	
Log [Value of Plot Output (TShs)]	0.056***
	(0.019)
Log [Value of Total Plot Input (TShs)]	0.032***
	(0.011)
Distance to Home(KM)	-0.034*
	(0.017)
Distance to Road(KM)	0.002
Owned w/o Title † *A*	0.308
	(0.195)
Rented/Other † *A*	0.028
	(0.246)
Soil Quality: Average † *B*	-0.050
	(0.115)
Soil Quality: Bad † *B*	-0.171
	(0.212)
Slope: Slight † *C*	0.274**
	(0.133)
Slope: Steep † *C*	0.169
	(0.187)
Household Characteristics	
Household Size	0.089***
	(0.019)
Dependency Ratio	-0.053
	(0.057)
Head of Household: Female †	-0.217
	(0.138)
Head of Household: Years of Age	0.009**
	(0.004)
Head of Household: Years of Education	0.000
	(0.018)
Head of Household: Primary Occupation: Agriculture †	0.260
	(0.175)
# of Plots in Holding	-0.094**
	(0.041)
Mover Original Household † D	0.014
	(0.235)
Split-Off Household † D	0.117
	(0.196)
Wealth Index 2008/09	-0.058
	(0.040)
Observations	3,363
R2	0.688

Note: *** p<0.01, ** p<0.05, * p<0.1. Constant is estimated but not reported. District and enumerator fixed effects as well as sampling weights are included as covariates but not reported. † denotes a dummy variable. The comparison categories are (i) owned with a title, (ii) good, (iii) flat, and (iv) non-mover original household for A, B, C, and D, respectively. All variables based on the TZNPS 2010/11 data, unless otherwise stated.

Table 5: Key Descriptive Statistics Following Multiple Imputation

	UNPS 2009/10			*TZNPS 2010/11*		
Plots w/ GPS-Based Plot Area	*Obs*	*Mean*	*Std Err*	*Obs*	*Mean*	*Std Err*
Actual GPS-Based Plot Area (Acres)	2,814	2.130	0.132	3,383	2.588	0.148
Value of Output Per Acre		484,861	27,119		94,000	10,857
Entire Plot Sample - Following MI	*Obs*	*Mean*	*Std Err*	*Obs*	*Mean*	*Std Err*
GPS-Based Plot Area (Acres)	4,333	2.124	0.125	4,141	2.564	0.147
Value of Output Per Acre		553,951	55,010		92,142	11,053

Table 6: OLS Regression Results
Dependent Variable = Log Value of Plot Output/Acre
Sample: UNPS 2009/10

	[1] *Observed* *GPS-Based Plot Area*	*[2]* *Multiply Imputed* *GPS-Based Plot Area*
Log Plot Area [Acres]	-0.388***	-0.515***
	(0.047)	(0.054)
Plot Characteristics		
Log [Value of Total Plot Input (UShs)]	0.094***	0.111***
	(0.013)	(0.011)
15-30 Mins Away from Household † *A*	-0.180	-0.146
	(0.149)	(0.135)
30+ Mins Away from Household † *A*	-0.220	-0.200
	(0.250)	(0.126)
Owned w/o Title † *B*	0.032	-0.030
	(0.152)	(0.143)
Rented/Other † B	-0.447**	-0.580***
	(0.181)	(0.168)
Soil Quality: Fair † *C*	-0.188*	-0.133
	(0.102)	(0.091)
Soil Quality: Poor † *C*	-0.392**	-0.379**
	(0.162)	(0.170)
Slope: Hilly, Steep or Valley †	0.143	0.114
	(0.107)	(0.098)
Household Characteristics		
Household Size	0.046**	0.059***
	(0.021)	(0.020)
Dependency Ratio	-0.070	-0.106**
	(0.049)	(0.048)
Head of Household: Female †	0.064	0.053
	(0.115)	(0.107)
Head of Household: Years of Age	-0.001	0.000
	(0.003)	(0.003)
Head of Household: Years of Education	0.027**	0.026**
	(0.011)	(0.011)
Head of Household: Primary Occupation: Agriculture †	0.148	0.204**
	(0.101)	(0.092)
# of Parcels in Holding	-0.074**	-0.088***
	(0.031)	(0.034)
Mover Original Household † D	0.525	-0.095
	(0.336)	(0.264)
Split-Off Household † D	-0.967	-0.354
	(0.797)	(0.351)
Wealth Index 2005/06	0.015	0.006
	(0.032)	(0.034)
Number of Imputations	N/A	50
Observations	2,814	4,333

Note: *** p<0.01, ** p<0.05, * p<0.1. Complex survey regressions, which correctly take into account clustering and stratification in weighted regressions, underlie the combined MI estimates reported here. Constant, and district and enumerator fixed effects are estimated but not reported. † denotes a dummy variable. The comparison categories are [i] less than 15 minutes from household, [ii] owned with a title, [iii] good, and [iv] non-mover original household for A, B, C, and D, respectively. All variables are based on the UNPS 2009/10, unless otherwise stated.

Table 7: OLS Regression Results
Dependent Variable = Log Value of Plot Output/Acre
Sample: TZNPS 2010/11

	[1]	*[2]*
	Observed	*Multiply Imputed*
	GPS-Based Plot Area	*GPS-Based Plot Area*
Log Plot Area [Acres]	-0.448***	-0.487***
	(0.064)	(0.060)
Plot Characteristics		
Log Value ofTotal Input/Acres	0.053***	0.055***
	(0.014)	(0.013)
Distance to Home (KM)	-0.082***	-0.002
	(0.024)	(0.004)
Distance to Road (KM)	-0.031	-0.060***
	(0.020)	(0.014)
Owned w/o Title †	-0.069	-0.010
	0.217	(0.219)
Rented/Other †	-1.174***	-1.063***
	(0.283)	(0.261)
Average †	-0.045	-0.110
	(0.120)	(0.108)
Bad †	0.046	0.058
	(0.216)	(0.200)
Slight †	0.253*	0.256**
	(0.131)	(0.128)
Steep †	0.267	0.244
	(0.202)	(0.196)
Household Characteristics		
Household Size	0.047**	0.035*
	(0.020)	(0.019)
Dependency Ratio	0.002	0.012
	(0.058)	(0.053)
Head of Household: Female †	-0.083	-0.112
	(0.140)	(0.132)
Head of Household: Years of Age	0.008**	0.007**
	(0.004)	(0.003)
Head of Household: Years of Education	0.004	-0.007
	(0.019)	(0.018)
Head of Household: Primary Occupation: Agriculture †	0.386*	0.223
	(0.200)	(0.207)
# of Plots in Holding	0.176***	0.190***
	(0.043)	(0.040)
Mover Original Household † D	0.165	-0.067
	(0.252)	(0.221)
Split-Off Household † D	-0.403	-0.338
	(0.260)	(0.210)
Wealth Index 2008/09	0.056	0.037
	(0.047)	(0.043)
Number of Imputations	N/A	50
Observations	3,383	4,121

Note: *** p<0.01, ** p<0.05, * p<0.1. Complex survey regressions, which correctly take into account clustering and stratification in weighted regressions, underlie the combined MI estimates reported here. Constant, and district and enumerator fixed effects are estimated but not reported. † denotes a dummy variable. The comparison categories are [i] owned with a title, [ii] good, [iii] flat, and [iv] non-mover original household for A, B, C, and D, respectively. All variables are based on the TZNPS 2010/11 unless otherwise noted.