Dang, Hai-Anh; Carletto, Calogero; Gourlay, Sydney; Abanokova, Kseniya

Working Paper

Addressing Soil Quality Data Gaps with Imputation: Evidence from Ethiopia and Uganda

GLO Discussion Paper, No. 1445

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Dang, Hai-Anh; Carletto, Calogero; Gourlay, Sydney; Abanokova, Kseniya (2024) : Addressing Soil Quality Data Gaps with Imputation: Evidence from Ethiopia and Uganda, GLO Discussion Paper, No. 1445, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/298382

Addressing Soil Quality Data Gaps with Imputation: Evidence from Ethiopia and Uganda

Hai-Anh Dang, Calogero Carletto, Sydney Gourlay, and Kseniya Abanokova*

June 2024

Abstract

Monitoring soil quality provides indispensable inputs for effective policy advice, but very few poorer countries can implement high-quality surveys on soil. We offer an alternative, low-cost imputation-based approach to generating various soil quality indicators. The estimation results validate well against objective measures based on benchmark surveys for Ethiopia and Uganda both for the mean values and the entire distributions of these indicators based on multiple imputation (MI) methods. Machine learning methods also perform well but mostly for the mean values. Furthermore, our imputation models can be combined with other publicly available, large-scale datasets on soil quality generated by model-based analysis with earth observations to provide improved estimates. Our results offer relevant inputs for future data collection efforts.

Key words: soil quality, multiple imputation, missing data, survey data, Ethiopia, Uganda

JEL: C8, O12, Q1, Q2

* Dang (hdang@worldbank.org; corresponding author) is a senior economist in the Living Standards Measurement Study Unit, Development Data Group, World Bank and is also affiliated with GLO, IZA, Indiana University, and London School of Economics and Political Science; Carletto (gcarletto@worldbank.org) is the senior manager of the Strategy and Collaboratives Unit, Development Data Group, World Bank; Gourlay (sgourlay@worldbank.org) is a senior economist in the Living Standards Measurement Study Unit, Development Data Group, World Bank; Abanokova (kabanokova@worldbank.org) is an economist in the Living Standards Measurement Study Unit, Development Data Group, World Bank. We would like to thank Songqing Jin, Dean Jolliffe, and participants at the Agricultural & Applied Economics Association meeting (Washington DC) and the World Bank Land Conference for useful feedback on an earlier version. We would also like to thank Siobhan Murray and Trong-Anh Trinh for helpful assistance with GIS data processing. We are grateful to the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through the Knowledge for Change (KCP) Research Program.

1. Introduction

The importance of healthy soils is well recognized for sustainable development and poverty reduction. Indeed, preserving soil quality is explicitly linked to several goals in the Sustainable Development Goals (SDGs), which range from improving agricultural production, food security and nutrition to end hunger (SDG number 2) to preserving clean water (SDG number 6) and reducing the harmful effects of climate change (SDG number 13). Monitoring soil quality thus provides indispensable inputs for effective policy advice. Yet, while richer countries such as the U.S. can maintain a century-old soil quality survey, very few, if any, low-income countries can afford such large-scale surveys on soil (Carletto *et al.*, 2021).

Unlike other household characteristics (such as gender or education attainment), collecting data on soil quality through farmers' self-reporting in a traditional household survey can be subject to severe measurement errors. Recent studies point to very weak or no correlation between farmers' subjective assessment of soil quality characteristics (inluding soil type, color and texture) and objective measurements using lab analysis or portable spectrometers in Ethiopia, Kenya and Tanzania (Carletto *et al.*, 2017; Gourlay *et al.*, 2017; Berazneva *et al.*, 2018; Kosmowski *et al.*, 2020). In this context, the arrival of large-scale, publicly available datasets on soil quality that are generated by model-based analysis with earth observations such as SoilGrids[1] and iSDASoil[2] are encouraging developments. But while these datasets provide useful data on soil that were not available before, it remains unclear to what extent their estimates compare with objective benchmark measures on soil quality.

[1] https://soilgrids.org/
[2] https://www.isda-africa.com/

We make several new contributions in this paper. First, we investigate whether we can explore alternative and less resources-intensive methods to produce high-quality data on soil quality. A useful approach is data imputation methods, which have seen increasingly more applications in various fields of social sciences such as health, psychology, and poverty measurement (van Buuren, 2018; Carpenter *et al.*, 2023; Dang and Lanjouw, 2023). But to our knowledge, hardly any study has employed imputation methods to produce estimates on soil quality. In our context, the central idea is to leverage a smaller benchmark survey with objective measures of soil quality—in combination with another larger (or more recent) dataset without any soil quality data (or with low-quality soil data)—to generate imputation-based estimates in the larger dataset. Since implementing objective measures of soil quality is expensive and requires much more logistical efforts than conducting a traditional self-report survey, this data generation setup can significantly reduce costs. In particular, we employ multiple imputation (MI) methods for analysis, and we also supplement our analysis with machine learning techniques for robustness checks.

Second, we examine how the estimates from large-scale, publicly available data sources such as SoilGrids and iSDASoil compare with the imputation-based estimates. Third, we further investigate whether we could employ imputation to improve the other data sources and provide better maps of soil quality. Finally, we also study the required sample sizes (for the base and target surveys) to implement imputation.

We find large differences between the benchmark measures of soil quality and those of iSDA and SoilGrids. The imputed estimates using multiple imputation methods are encouragingly similar to the benchmark estimates and reasonably approximate the whole distributions of several soil quality indicators, including pH, soil organic carbon (%), total nitrogen (%), soil texture, and

a soil quality index. Machine learning methods also perform well, but mostly for the mean values. We also find that imputation methods can be used in combination with iSDA and SoilGrids data to further improve on these data. Regarding sample sizes, the results vary by country but obtaining good imputation results appears to require a sample size for the benchmark survey between 752 plots (for Uganda) and 350 plots (for Ethiopia). Our results offer relevant inputs for both future survey design and improving the quality of the SoilGrids and iSDASoil databases.

This paper consists of six sections. We provide a brief overview of the country contexts and summary of the data in the next section before discussing the MI framework and the existing theory on selecting sample sizes for imputation in Section 3. We present in Section 4 the estimation results and further extension to the sample sizes. We further discuss the implications for survey implementation and finally conclude in Section 5.

2. Country background and data description

2.1. Country background

Ethiopia and Uganda are agriculture-centric economies, with agriculture, forestry, and fishing accounting for 37.6 and 24 percent of GDP respectively (World Bank, 2024). This highlights the importance of soil in these countries. Understanding soil health, particularly on land cultivated by smallholder farmers where consumption of own production is common, is critical for agricultural productivity and nutrient intake, especially in the face of climate change and growing food demand. Yet, soils in the African continent, 40% of which are estimated to be of low fertility, are being depleted further by unsustainable agricultural practices, low use of fertilizer inputs, and other degradation processes (FAO, 2022). In Ethiopia, soil erosion and degradation limit agricultural productivity, with an estimated 42 tons/ha of soil loss from cultivated lands and

additional constraints coming from strong soil acidity; this impacts over 28 percent of the country, as well as soil nutrient deficiencies (Kassahun, 2015). In Uganda, soil erosion also poses a significant challenge , where costs of environmental degradation, primarily through soil erosion, were estimated around 4-12% of GNP (NEMA, 2001). Furthermore, soil nutrient deficiencies have also been noted as key impediments to agricultural productivity and sustainable production in the country (Nkonya *et al.*, 2005a, 2005b; Wortmann and Kaizzi, 1998), with nutrient depletion rates in Uganda noted among the highest in sub-Saharan Africa at one time (Stoorvogel and Smaling, 1990).

2.2. Data description

We leverage ground-based, plot-level soil samples and survey data collected through methodological studies in Ethiopia and Uganda ('benchmark surveys'), nationally-representative longitudinal household surveys in Ethiopia and Uganda ('target surveys'), and modelled geospatial-based soil data from two publically available sources, namely SoilGrids 2.0 (Poggio *et al.*, 2021) and iSDAsoil (Hengl *et al.*, 2021).

Benchmark surveys

Both methodological studies, the Land and Soil Experimental Research (LASER) study in Ethiopia and the Methodological Experiment on Measuring Maize Productivity, Soil Fertility and Variety (MAPS) study in Uganda, were conducted by the World Bank in collaboration with International Centre for Research in Agroforestry (ICRAF) and the respective national statistical agencies, and were purposefully designed to allow for validation of methods for measuring soil

health, among other domains, in the context of household surveys. The LASER and MAPS studies had similar designs in that they: (i) implemented a three-visit survey approach, with one visit at the post-planting stage, one visit for crop-cutting, and a final visit in the post-harvest period;[3] and (ii) collected soil samples, using the same collection protocols, from randomly selected plots cultivated by the selected households, with support from and analysis by ICRAF.

The Ethiopia LASER study was implemented in three administrative zones of the Oromia region (East Wellega, West Arsi, Borena), selected primarily based on their agroecological and topographic diversity, from September 2013 to February 2014 (Figure 1).[4] Using Ethiopia's Agricultural Sample Survey (AgSS) as a sampling frame, a total of 85 enumerations areas (EAs) were selected in accordance with a determined practical allocation of EA counts across agroecological and administrative zones to ensure variation in the resulting sample. Twelve households were randomly selected from each EA, drawn from the AgSS household listing conducted in September 2013. Up to two plots were selected for soil sampling from each household, with the first randomly selected among the purestand maize plots cultivated by the household, if any, and the second randomly selected from all remaining cultivated plots, irrespective of crop type. The plots with purestand maize were also subject to crop-cutting. Soil samples were collected following the protocol described in Gourlay *et al.* (2017), dried and processed in local soil research centers, and shipped to ICRAF Nairobi for analysis.

[3] In the first survey visit, each selected household was administered a questionnaire that collected information on household composition and demographics and an established roster of land parcels and plots with information collected on tenure type, cultivation status, management, agricultural inputs, subjective assessment of soil health, and farming practices, among others. Additionally, in this initial visit, a subsample of randomly selected plots was subject to soil sample collection, area measurement (via Garmin eTrex 30), and demarcation of crop-cutting subplots (as relevant).

[4] For more information on the LASER study, as well as access to the microdata, visit: https://microdata.worldbank.org/index.php/catalog/2671/study-description

Similarly, the Uganda MAPS study was conducted on a sub-national level, with coverage spanning three strata: Serere district, Sironko district, and a 400 km^2 remote sensing tasking area tht covered portions of Iganga and Mayuge districts, as illustrated in Figure 1. A total of 75 EAs were selected with probability proportional to size, based on the 2014 population and housing census housing counts (15 from Serere, 15 from Sironko, 45 from the remote sensing tasking area). A household listing exercise was undertaken in each selected EA, and from that list 12 maize-growing households were selected, with an effort to randomly select 6 purestand and 6 intercropped maize-crowing households in each.[5] The MAPS study is a longitudinal study, with Round 1 conducted in 2015 and Round 2 in 2016, though this analysis is limited to MAPS 1 as the second round did not include objective soil analysis. In the MAPS study, one maize plot was randomly selected for soil analysis per household, soil samples were collected and crop cutting was conducted on those plots.[6]

In this paper we focus on key chemical properties, namely pH, soil organic carbon (%), and total nitrogen (%), and soil texture (percent clay, silt, and sand). Using these properties, we construct an index guided by the nutrient storage capacity index put forth by Mukherjee and Lal (2014). Although this index also included electrical conductivity, which is available in the LASER and MAPS datasets, we exclude this component since it is not available in the geospatial soil

[5] Pure stand plots are those on which only maize is grown. Intercropped plots are those on which maize and at least one other crop is grown.

[6] The soil analysis conducted by ICRAF on the Ethiopia LASER and Uganda MAPS soil samples included mid-infrared (MIR) soil spectroscopy and laser diffraction particle size distribution (LDPSA) analysis on all collected samples, in addition to reference analyses which were conducted on a subset of soil samples and used to calibrate and validate the MIR-based prediction models (Shepherd & Walsh 2002, 2004, 2007) which included conventional wet chemistry, x-ray analysis (XRD) for mineralogy, and total element analysis (TXRF). The set of analyses conducted by ICRAF resulted in a dataset with plot-level estimates of multiple soil physical and chemical soil properties for both the top- and sub-soil samples collected in each study, with linkages to the survey data and plot coordinates.

products used for imputation.[7] The index is constructed by first assigning a score for each soil property on each plot, ranging from 0 to 3 (based on thresholds found in Annex X) and subsequently the scores are normalized over the study samples and then weighted according to the weights identified in Mukherjee and Lal (2014). We subsequently reweight the three soil elements such that the weights sum up to one. The final index, referred to henceforth as the *weighted soil index*, ranges from 0 to 1, where a higher value indicates greater nutrient storage capacity.[8] The index is created separately for top- and sub-samples, although we utilize the sub-soil-based index for robustness checks only.

Target surveys

The target surveys, upon which we apply our imputation models, include two nationally representative household surveys conducted with support from the World Bank's Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) project in Ethiopia and Uganda. For Ethiopia, we utilize the Ethiopia Socioeconomic Survey Wave 2 (ESS2), which is part of a nationally representative multi-topic longitudinal household survey. Wave 2 was selected as it is best aligned with the implementation period of the LASER study. The ESS2 sample consists of 5,262 agricultural and non-agricultural households. For Uganda, we utilize the 5th round of the Uganda National Panel Survey (UNPS5), implemented in 2015/16 with a sample of 3,305 agricultural and non-agricultural households. For the imputation analyses using the LSMS-ISA survey data, we create two separate analysis samples: one sample restricted to areas in which the

[7] Electrical conductivity was included in the 2015 version of the AfSoilGrids250m (Hengl et al., 2015) dataset, but was removed in later products. We replicate our analysis using electrical conductivity estimates derived from the 2015 AfSoilGrids250m product for robustness.

[8] Without reweighting, the soil index ranges from 0 to 0.7. Estimation results using this index yields qualitatively similarly results (not shown).

benchmark surveys were conducted (district-level restrictions in Uganda and zone-level restrictions in Ethiopia) and another restricted to areas in which the benchmark studies are were *not* implemented. We then employ each of the two samples separately as the target survey. [9]

Geospatial soil products

To complement the soil ground-based soil data collected through the LASER and MAPS studies, the key soil properties described above are extracted from two publicly available geospatial soil products, SoilGrids 2.0 and iSDAsoil, with the soil quality index subsequently computed from these extracted property values. The SoilGrids 2.0 product (Poggio *et al.*, 2021), henceforth refered to as *SoilGrids,* includes spatial predictions of soil properties at various depths globally, at 250m spatial resolution.[10] Similarly, building on the Africa Soil Information Service (AfSIS) project, the iSDASoil product provides soil property estimates for the African continent, at multiple depths, though at 30m spatial resolution (Hengl *et al.*, 2021). Data from SoilGrids and iSDASoil are extracted based on plot-level coordinates collected in both the benchmark surveys, the plot-level coordinates in ESS2, and the household coordinates for UNPS5.

Descriptive statistics

[9] For more details and access to the ESS2 and UNPS5 surveys, visit https://microdata.worldbank.org/index.php/catalog/2247 and https://microdata.worldbank.org/index.php/catalog/3460.

[10] SoilGrids 2.0 (version 2020) is a global digital soil mapping system produced by ISRIC - World Soil Information, which builds on previous iterations of the SoilGrids products including AfSoilGrids250m (Hengl et al., 2015) and SoilGrids250m (Hengl el at, 2017). Ground-based samples that factor into the SoilGrids 2.0 product were collected from 1924 to 2020, with "5 % of the profiles were sampled before 1960, 14 % between 1961–1980, 32 % between 1981–2000 and 16 % between 2001–2020; the date of sampling is unknown for 34 % of the shared profiles" (Poggio et al., 2021).

Table 1 compares topsoil quality in the benchmark surveys and the geospatial data sources for the same regions where the benchmark surveys were implemented. We find that both geospatial data sources provide biased estimates at the plot level for most of the soil properties. The difference in means across data sources for the weighted soil index, a key variable used in this analysis, is significant at the 1 percent level and overestimated in geospatial data for both countries. The iSDA data, on average, significantly underestimate Ethiopia's soil chemical properties, such as pH, organic carbon and total nitrogen. It also underestimates pH and overestimates total nitrogen in Uganda. The SoilGrids data significantly underestimate pH and overestimate organic carbon and total nitrogen in both countries. The approximation of soil physical properties is far from ideal, with geospatial data sources underestimating clay and overestimating silt and sand composition. Overall, Table 1 suggests that relying solely on iSDA or SoilGrids-derived values can result in an overestimation of the soil quality index and sand composition, as well as an underestimation of pH, at the plot level in both countries. Significant differences between the benchmark surveys and the geospatial data are also found for subsoil properties (Appendix B, Table B.1).

Tables A.1 and A.2 in Appendix A provide summary statistics and numbers of observations of key variables for the benchmark and target surveys in Ethiopia and Uganda for the sample with plot-level soil analysis (Panel A), as well as for the combined sample of plot-level soil analysis from LASER and MAPS1 with the iSDA and SoilGrids geospatial data (Panel B). Kolmogorov-Smirnov (KS) distribution tests suggest almost no difference across benchmark and target surveys regarding household head characteristics (such as gender and secondary and higher education in Ethiopia and all education levels in Uganda). While the distributions of some plot characteristics are statistically significantly different between the two surveys for Ethiopia, other characteristics are similar (such as the use of inorganic fertilizer and pesticides, and the type of crop stand). The

use of pesticides on the plot is not statistically significantly different between the surveys for Uganda.

Yet, soil properties from the geospatial data have different distributions across the benchmark and target survey locations in both countries, raising further concerns about data quality. Summary statistics for soil variables, including organic carbon, total nitrogen, pH, clay, silt, and sand contents, and the weighted soil index are reported in Panel B. Only silt content in Uganda collected in the benchmark districts of the UNPS5 survey is not statistically significantly different from those collected in MAPS1 survey.

3. Analytical framework

An established statistics literature exists on multiple imputation (MI) methods, which address missing data. Official agencies such as the U.S. Census Bureau routinely use imputation to fill in important missing data on various statistics for income (Census Bureau, 2016a) and labor (Census Bureau, 2016b). MI methods were also employed in economics to study various topics, ranging from children's family experiences and psychosocial adjustment (Davey *et al.*, 2001) to income inequality (Jenkins *et al.*, 2011) and poverty (Douidich *et al.*, 2016; Dang *et al.*, forthcoming). Yet, MI methods still remain little used in economics.[11] For this reason, we provide below a brief discussion of MI methods based on Rubin (1988) and Little and Rubin (2019).

Let x_j be a vector of characteristics that are commonly observed between the two surveys, which include the base survey (j= 1) and the target survey (j= 2). To make notation less cluttered,

[11] This stands in contrast with a growing literature on survey-to-survey poverty imputation in economics, which build on earlier efforts with "poverty mapping" (which imputes from a survey into a population census) (Elbers *et al.*, 2003). Further discussion on the differences between poverty imputation methods in economics and MI methods is provided in Dang and Lanjouw (2023).

we suppress the subscript for each household in the following equations. Subject to data availability, these characteristics can include individual-level and household-level characteristics. Individual characteristics include variables such as age, sex, education, ethnicity, religion, language, and occupation. Household characteristics include variables such as household size, the number of rooms in the house, the physical quality of the house (e.g., whether its roof or wall is of good quality), and the distance from the house to the nearest facilities, such as sources of water. These variables can capture the household's income levels.[12]

High-quality data on soil quality exist in the base survey (which is typically the benchmark survey but we also examine other base surveys later on) but are not available in the target survey. Let y_{1k} represent the outcomes of interest in survey *1*, the base survey, where k represents the different soil quality indicators. Our objective is to impute the missing (or low-quality) soil quality indicators in survey *2*, given that the survey characteristics x_j are available in both surveys.

We assume that the linear projection of soil quality indicators on household and other characteristics (x) is given by the following linear model

$$y_{jk} = \beta_{jk}' x_j + \varepsilon_{jk} \quad (1)$$

where β_{jk} are the vector of coefficients. x_j includes both household and plot-level characteristics which are considered to be correlated with soil quality. These include the incidence of fertilizer and pesticide use, incidence of hired labor, cropping patterns (pure stand or intercropped plots), cultivation history, and plot size. Geospatial climate and topology variables, such as elevation, precipitation, and temperature, were not included since these variables are already captured by geospatial-based soil data. Household characteristics include variables such as household head

[12] Household assets or income can also be included if such data are available.

age, household head gender and education, and household size, all of which are believed to be correlated with household income levels.

Conditional on the x_j characteristics, the error term is assumed to follow a normal distribution $\varepsilon_{jk}|x_j \sim N(0, \sigma^2_{\varepsilon_{jk}})$. Equation (1) thus provides a standard linear model that can be estimated using most available statistical packages.

We make the following assumption to further operationalize our estimation framework.

Assumption 1: Let x_j denote the values of the variables observed in survey j, for j= 1, 2, and let X_j denote the corresponding measurements in the population. Then x_j are consistent measures of X_j for all j (i.e., $x_j=X_j$ for all j).

Assumption 1 is crucial for imputation and ensures that the sampled data in each survey are representative of the target population. While somewhat different versions of this assumption are commonly employed in previous imputation studies (Elbers *et al.*, 2003; Tarozzi, 2007; Dang and Lanjouw, 2023), this assumption essentially implies that, for the surveys under consideration, measurements of the same characteristics *x* are identical, as they are consistent measures of the population values. While surveys of the same design (and sample frame) are more likely to be comparable and can thus satisfy Assumption 1, these surveys may not necessarily provide comparable estimates. Examples where Assumption 1 may be violated include cases where national statistical agencies change the questionnaire for the same survey over time.[13] Violation of Assumption 1 rules out the straightforward application of survey-to-survey imputation technique and would require further investigation of estimation results.

[13] The inconsistency between different rounds of the same survey or different surveys is well documented in studies using data from both poorer and richer countries. Survey design issues that compromise the comparability of poverty estimates are found in various countries such as China (Gibson, Huang, and Rozelle, 2003), Tanzania (Beegle *et al.*, 2012), and Vietnam (World Bank, 2012). See also Angrist and Krueger (1999) for a related review of comparability and other data issues with a focus on labor force surveys in the U.S.

Assumption 1 can be tested when the surveys under study are implemented in the same period. The discussion above (with Tables A.1 and A.2) suggest that the benchmark and target surveys generally satisfy this assumption; that is, the household and plot characteristics using the target surveys are not statistically significantly different from those based on the benchmark survey for both countries.

Given Assumption 1 and clearly writing out Equation (1), we can replace x_1 with x_2 as follows

$$y_{2k}^{1} = \beta_{1k}{}' x_2 + \varepsilon_{1k} \tag{2}$$

Equation (2) thus applies the model parameters β_{1k} and ε_1 based on the base survey to the x characteristics in the target survey to obtain estimates of the soil quality indicators y_{2k}^{1} in this survey.

Since the estimated parameters are obtained using a different survey from the target surveys, we can use simulation to estimate Equation (2) as follows

$$\hat{y}_{2k}^{1} = \frac{1}{S}\sum_{s=1}^{S}(\tilde{\tilde{\beta}}'_{1k,s} x_2 + \tilde{\tilde{\varepsilon}}_{1k,s}) \tag{3}$$

where $\tilde{\tilde{\beta}}'_{1k,s}$ and $\tilde{\tilde{\varepsilon}}_{1k,s}$ represent the s^{th} random draw (simulation) from their estimated distributions, for s= 1,…, S. The variance of $\hat{y}_{2k}^{1}$ can be estimated as

$$V(\hat{y}_{2k}^{1}) = \frac{1}{S}\sum_{s=1}^{S} V(\hat{y}_{2k,s}^{1}|x_2) + V(\frac{1}{S}\sum_{s=1}^{S}\hat{y}_{2k,s}^{1}|x_2) + \frac{1}{S}V(\frac{1}{S}\sum_{s=1}^{S}\hat{y}_{2k,s}^{1}|x_2) \tag{4}$$

As an alternative to the linear regression method offered in Equation (3), we can employ a predictive mean matching (PMM) algorithm to draw $\hat{y}_2^1$ instead from the nearest matching observation in the base survey. More formally, applying the estimated parameters from Equation (2) to the base survey itself for each simulation s, we have

$$\hat{y}_{1k,s}^{1} = \tilde{\tilde{\beta}}'_{1k,s} x_1 + \tilde{\tilde{\varepsilon}}_{1k,s} \tag{5}$$

We subsequently replace $\hat{y}^1_{2k,s}$ with $\hat{y}^1_{1k,s}$ such that the absolute difference $\left|\hat{y}^1_{2k,s} - \hat{y}^1_{1k,s}\right|$ for each individual is minimized, drawing from five nearest neighboring observations.

Since the PMM algorithm is non-parametric, it does not rely on the assumption of normality of the error term ε_{jk} and offers better estimation results where such assumption does not hold (Little, 1988). This advantage may be even more relevant in our study since the various non-benchmark surveys can potentially offer biased estimates due to their small sample sizes (even where the normality assumption holds). Consequently, the PMM imputation method is our preferred estimation method and will be employed for most of the analysis. However, we also show some estimates based on Equation (3) for comparison.

For alternative estimation options, we also employ several common machine learning (ML) techniques for robustness checks. These include LASSO, Elastic Net, and Random Forest. The standard ML procedures split a data sample into a training sample and an estimation sample. The training sample is used to estimate the imputation model, which is subsequently applied to the estimation sample to obtain out-of-sample predictions on the estimation sample. In our context, the training sample and the estimation sample respectively correspond to the benchmark survey and the target survey. Compared with MI methods which are based on statistical theory, ML methods are more data-oriented.[14]

4. Estimation results

4.1. Imputation-based estimates versus objective measures

We start by predicting plot-level soil properties from the benchmark surveys into larger LSMS surveys, with the latter being restricted to the same districts/zones included in the

[14] See Mullainathan and Spiess (2017) and Athey and Imbens (2019) for recent reviews of ML methods for economics.

benchmark surveys for better comparison. Our primary imputation method is predictive mean matching, although imputation results using the linear regression method are qualitatively similar (Appendix A, Tables A.3 to A.5).[15]

Mean values

Table 2 shows the "true rate" calculated based on the benchmark survey (column 1 for Ethiopia and column 3 for Uganda) and the imputation-based estimates (column 2 for Ethiopia and column 4 for Uganda). The (impution-based) estimates of the weighted soil index using the PMM method are 0.52 in Ethiopia and 0.56 in Uganda, which all fall within the 95 percent confidence intervals (CIs) around the true figures of 0.53 and 0.57 using the benchmark surveys, respectively.[16] In fact, the estimates in Ethiopia even fall within one standard error around the true weighted soil index of 0.53. The estimates of soil chemical properties, such as pH and organic carbon, fall within the 95 percent CIs in Ethiopia and Uganda, with organic carbon being within the one standard-error bandwidth around the true rate in Ethiopia. The estimates of total nitrogen, whose value is very close to the true total nitrogen of of 0.27 percent in Ethiopia and 0.11 percent

[15] The estimation results for the underlying linear regression model for both imputation approaches (based on Equation (1)) are shown in Appendix A, Table A.6 for Ethiopia, and Table A.7 for Uganda. Among the household characteristics, the household head`s gender and household size are significantly correlated with the weighted soil index and its components, such as organic carbon and total nitrogen, of which a higher value is preferred. The direction of correlation differs between countries, with male heads negatively associated with soil properties in Ethiopia but positively associated in Uganda and with bigger households positively associated in Ethiopia but negatively in Uganda. Pure stand and crop rotation are negatively correlated with soil quality index, organic carbon and total nitrogen in Ethiopia but are not significant in Uganda. The use of inorganic fertilizer positively correlated with organic carbon and total nitrogen content in both countries, as expected. In contrast, organic fertilizers and pesticides are not significant predictors of these soil properties. Cultivating the land in previous years is negatively correlated with soil quality in Uganda as it can deplete the soil over time (data not available for Ethiopia). Problems with erosion on the plot are negatively correlated with the weighted soil index, organic carbon and total nitrogen in Uganda, while the precautionary measures against erosion in Ethiopia are positively related to these soil properties. GPS-measured size of plot is a significant predictor across almost all soil properties in both countries, with bigger plot associated with lower concentrations of organic carbon and total nitrogen.

[16] The imputation-based estimates of the weighted soil index that includes electrical conductivity fall within the 95 percent confidence intervals (CIs) around the true rate with electrical conductivity is within the one standard-error bandwidth around the true rate (Table C.2, Appendix C).

in Uganda, does not. The estimates of soil physical properties all fall within the 95 percent CIs around the true rate, with silt and clay composition being within the one standard error of the true rate in both countries[17].

The linear regression method performs slightly worse than the PMM method for pH in Uganda with its estimate falling outside the 95 percent CIs around the true figures (Table A.3, Appendix A), but improved accuracy for the sand composition, keeping it within the one standard-error bandwidth. The linear regression method also improved accuracy for total nitrogen in Ethiopia, keeping the imputed estimates within the one standard-error bandwidth around the true rate.

Figure 2 further shows that while the imputed estimates almost mimic those based on the benchmark surveys in Ethiopia and Uganda (with ratios of approximately 1 for mean values), the divergence between experimental soil data and iSDA and SoilGrids can be high. The mean value of sand content produced by SoilGrids is 2.5 times higher than that collected in the benchmark survey in Ethiopia. However, the gap is even higher when compared with iSDA data, where the sand content value is three times higher than the benchmark survey. The sand content produced by iSDA for Uganda is twice higher than those collected in the MAPS1 survey, with a similar gap compared to data produced by SoilGrids. For Ethiopia, the organic carbon and total nitrogen are not well approximated by iSDA data, and both iSDA and SoilGrids do not identify clay

[17]As a robustness check we also restricted the UNPS5 sample to maize cultivated plots only to mirror the crop composition of the MAPS study. About 24% of all plots cultivated in the target period are used for growing maize, significantly reducing the target sample size. Even though the sample size is much smaller, the imputation method works with the accuracy of the imputation estimates slightly improved for chemical composition.

composition well. All these soil properties are not well approximated by SoilGrids in Uganda, where organic carbon is twice higher than the benchmark survey.

In summary, the imputed estimates for the mean values are largely not statistically different from the benchmark values (except for some perhaps negligible differences in nitrogen) in both countries indicates that MI can offer encouraging results. More importantly, while some plot characteristics of the target survey have different distributions from those of the benchmark survey, MI appears robust to these differences.

Entire distribution

We turn next to the question on how well imputation approximates the whole distribution of the variables of interest. To ensure that the imputation technique works for the whole distribution of soil properties, we show the imputed estimates of quintiles using the benchmark survey in Ethiopia (Figure 3) and Uganda (Figure 4). To do this, we split the benchmark surveys into two random samples with equal size (i.e., two random halves) and used one half as the base survey and the other half as the target sample.[18] Figures 3 and 4 clearly illustrate that the imputed quintiles (the black line) largely mimic "true" quintiles in the target survey (the gray zone) for most soil properties, excluding silt content in Ethiopia. For the latter, the imputed estimates of the 50th percentile are statistically different from the true estimates in this percentile. The linear regression method of imputation is less successful in approximating the distribution by mainly underestimating the extreme values of silt and sand in the tails of the distribution in Ethiopia

[18] For this exercise, we do not impute into the target surveys but work with the benchmark surveys alone because they offer a larger sample size. For example, a random half for Uganda (Figure 4) provides more than 400 plots, which is more than twice larger than the 163 plots in the benchmark regions for the target survey for this country.

(Appendix A, Figure A.1). It also underestimates the extreme values of total nitrogen and acidified carbon in Uganda and overestimates the middle portion of the distribution of weighted soil index in both countries (Appendix A, Figure A.2).

Machine learning as alternative approach

For robustness checks, we re-estimate the results in Table 2, using the ML techniques LASSO, Elastic Net, and Random Forest and showed the estimation results in Appendix D, Tables D.1 and D.2. The results are encouragingly strong with most of the estimates falling inside the 95 percent CIs, or even within one standard error, of the true rate, except for Random Forest estimates for Uganda. LASSO estimates perform even somewhat better than MI results, yielding estimates for Ethiopia that all fall inside the 95 percent CIs. However, when we plot the distribution graphs (Appendix D, Figures D.1 and D.2), ML methods appear to work only for the middle portion of the distributions for most soil quality indicators.

4.2. Imputation results with geospatial data

The estimation results in the previous discussion are obtained based on the assumption that the characteristics of the target surveys have the same distributions as those of the base surveys (which are the benchmark surveys). We now relax this assumption and examine whether estimation results still hold in different scenarios.

First, we use as the base survey the four combined samples of the benchmark surveys (LASER or MAPS1) and geospatial soil data (iSDA or SoilGrids), with each benchmark survey combined with each geospatial database. We similarly use these combined samples of data as the

target survey. Pretending that the geospatial data are missing in the target survey, we first impute for these missing values before imputing for the soil quality indicators using these imputed geospatial values. We implement the procedures using a sequence of independent univariate conditional imputation methods, assuming that the modelling structure is monotone distinct (Rubin, 1988).[19]

Table 3 compares the imputation-based estimates against the benchmark survey estimates using the PMM method, where the target survey is restricted to the same areas as the benchmark surveys. The estimation results are encouraging, with many imputed estimates of soil quality indicators being within the 95 percent CIs of the benchmark rate. Using the iSDA-imputed samples, the estimates of the weighted soil index are 0.52 in Ethiopia and 0.56 in Uganda, which all fall within the 95 percent confidence intervals (CIs) around the benchmark rate of 0.53 and 0.57 respectively. The imputed results for the weighted soil index using the SoilGrids- and iSDA-imputed samples even fall within the one standard error around the benchmark estimates for Ethiopia.

The imputation method improves iSDA estimates for organic carbon and pH in Ethiopia, with the latter falling within the one standard error around the benchmark estimates. Both iSDA- and SoilGrids-imputed estimates of silt and sand in Ethiopia are improved, with the mean values falling within the 95 percent CIs around the benchmark values. The imputation method improves both iSDA and SoilGrids estimates of soil physical properties in Uganda which all fall within the

[19] The underlying linear regression results for the first step are shown in Appendix A, Tables A.8 and A.10 for Ethiopia, and Tables A.12 and A.14 for Uganda, and the corresponding results for the second step are shown in Appendix A, Tables A.9 and A.11 for Ethiopia and Tables A.13 and A.15 for Uganda. Given the correlated nature of soil properties, geospatial variables are highly significant in all plot-level soil properties, as expected.

95 percent CIs around the benchmark estimates. Furthermore, the imputation-based estimates for silt and clay even fall within one standard error of the true rate in Uganda.

The estimation results, shown in Appendix A, Table A.4 for linear regression methods, added more accuracy to imputed estimates of weighted soil index, organic carbon and total nitrogen in Ethiopia, keeping it within one standard error of the benchmark rate. The linear regression method also improved average values of the iSDA and SoilGrids-imputed estimates of organic carbon in Uganda.

In summary, obtaining the imputed values in the benchmark areas of the target survey that are not statistically different from the benchmark values indicates that the imputation model results in improved soil measures derived from geospatial data, although the geospatial plot characteristics of the target survey have different distributions from those of the benchmark survey.

Given the similarity between the imputed and benchmark estimates in Ethiopia and Uganda for benchmark areas in Table 3, we expect that the adjustment of geospatial data through imputation can improve estimates of soil quality indicators in non-benchmark areas (i.e., the remaining areas of the target surveys not covered in the benchmark surveys in Ethiopia and Uganda). The imputation results using the PMM method are shown in Table 4, which compares the imputed estimates that incorporate the geospatial data with those based on the existing geospatial data. Assuming the imputation results provide better estimates of the true values, Table 4 shows that all the geospatial soil quality indicators fall outside the 95 percent CIs around the imputed estimates. The alternative linear regression estimates offer similar results, except for the total nitrogen in Ethiopia and organic carbon in Uganda (Appendix A, Table A.5). Specifically, the SoilGrids estimate of total nitrogen is 0.27 in Ethiopia, which falls within the 95 percent CIs

around the estimated total nitrogen of 0.27 by adding SoilGrids data to the imputation model. The actual iSDA estimate of organic carbon is 1.52 in Uganda, which falls within the 95 percent CIs around the estimated organic of 1.56 by adding iSDA data to the imputation model.

We plot the imputed estimates of the soil quality index (from Table 4) to construct the improved soil map for each country and compare it with the maps derived from geospatial data sources. Figure 5 shows in the first row the variation across Ugandan districts for the iSDA-imputed soil quality index and the imputed soil elements pH, organic carbon and total nitrogen. Figure 5 also shows in the second row the corresponding estimates using the existing iSDA data. Figure 6 shows the similar results for Uganda. The weighted soil indexes derived from the existing iSDA data (second row) are generally biased upward, displaying distinct patterns towards areas with much higher soil quality in the south for both countries. On the other hand, the iSDA soil elements (second row) generally have lower levels than the imputed estimates (first row).

The imputed estimates using the SoilGrids data are shown in Appendix A, Figures A.5 and A.6. While there are considerable differences in the spatial variation of the soil quality indexes across the data sources, the differences between the indicators based on the existing SoilGrids data and the imputated estimates are reassuringly similar for both countries. These results are consistent with the results discussed in Section 4.1 above. This suggests that the proposed imputation method can be used to improve iSDA and SoilGrids data to better reflect plot-level soil composition.

4.3. Further extension

We turn next to examining the question of how large the appropriate sample sizes for the imputation model should be. We answer this question by splitting the benchmark data into two

random samples: 50% in-sample and 50% out-of-sample, and plotting the estimates of weighted soil index and soil quality indicators imputed from the benchmark sample (defined as a percentage varying from 10% to 100% of the "in-sample" data) into the target sample (defined as 100% of the "out-of-sample").

Since we have only 1672 observations for the LASER survey, we consider a range of 84 to 836 observations for the benchmark sample size in Ethiopia. Figure 7 shows that all estimates (the black line) fall within the 95% CIs of the true rate (the gray range) and fluctuate less at a sample size of 752 or larger. Figure 8 suggests that the estimates in Uganda fluctuate less at a sample size of around 307 observations, including falling somewhat outside the 95% CIs of the true rate, and not stabilizing until a minimum sample size of 350 observations. The range of 44 to 438 observations was selected for the benchmark sample size in Uganda (as we have 877 available observations in the MAPS1 survey).

In summary, obtaining good imputation results appears to require a larger sample size in Ethiopia than in Uganda (i.e., 752 versus 350).[20] The linear regression method of imputation gives a lower sample size for Ethiopia with estimates falling outside the 95% CIs of the true rate up to the total sample size of 502 plots (Appendix A, Figure A.5). As with the PMM method, the sample size requirements for Uganda are lower than for Ethiopia but higher than for the linear regression method – 384 plots (Appendix A, Figure A.6).

Similar analysis with sample sizes for the target survey suggests that obtaining good imputation results appears to require a sample size of 486 or larger in Ethiopia (Appendix A, Figure A.7) and a sample size of 163 or larger for most of the soil quality indicators in Uganda (Appendix

[20] Qualitively similar results were obtained using subsoils as opposed to topsoils (Appendix B, Figures B.1 and B.2).

A, Figure A.8). However, the results regarding choosing sample sizes for the target survey with total nitrogen in Ethiopia and Uganda are somewhat inconsistent.

5. Conclusion

We propose imputation as an alternative approach to address the shortage of high-quality survey data on soil quality in poorer countries. The results using multiple imputation (MI) methods, and to some extent machine learning methods, work reasonably well against objective measures based on benchmark surveys for Ethiopia and Uganda. We also show that imputation methods can be combined with publicly avaiable, large-scale datasets to further improve the quality of these datasets.

Our results provide relevant inputs for future survey design. If replicated for countries in other contexts (e.g., in a different region or at a different income level), these results could open up promising avenues for generating soil quality data in a cost-effective manner. Compared to implementing a full scale survey, imputation is far less expensive. Furthermore, while employing imputation is more demanding for analytical capacity, it is less demanding for survey implementation capacity. Consequently, in the absence of high-quality soil data, imputation offers a useful second-best option for data generation particularly where there are improved levels of local analytical capacity.

References

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.

Beegle, K., De Weerdt, J., Friedman, J., & Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of development Economics*, *98*(1), 3-18.

Berazneva, J., McBride, L., Sheahan, M., & Güereña, D. (2018). Empirical assessment of subjective and objective soil fertility metrics in east Africa: Implications for researchers and policy makers. *World Development*, *105*, 367-382.

Carletto, C., Dillon, A., & Zezza, A. (2021). Agricultural data collection to minimize measurement error and maximize coverage. In *Handbook of Agricultural Economics* (Vol. 5, pp. 4407-4480). Elsevier.

Carletto, C., Aynekulu, E., Gourlay, S. and Shepherd, K., 2017. Collecting the dirt on soils: advancements in plot-level soil testing and implications for agricultural statistics. *World Bank Policy Research Working Paper 8057*.

Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., & Kenward, M. G. (2023). *Multiple imputation and its application*. John Wiley & Sons.

Dang, H. A., & Lanjouw, P. F. (2023). Regression-based imputation for poverty measurement in data-scarce settings. In *Research handbook on Measuring Poverty and Deprivation* (pp. 141-150). Edward Elgar Publishing.

Dang, H. A., Kilic, T., Abanokova, K., & Carletto, C. (forthcoming). Poverty imputation in contexts without consumption data: A revisit with further refinements. *Review of Income and Wealth*.

Davey, A., Shanahan, M. J., & Schafer, J. L. (2001). Correcting for selective nonresponse in the National Longitudinal Survey of Youth using multiple imputation. *Journal of Human Resources*, 500-519.

Douidich, M., Ezzrari, A., Van der Weide, R., & Verme, P. (2016). Estimating quarterly poverty rates using labor force surveys: a primer. *World Bank Economic Review*, *30*(3), 475-500.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

FAO. 2022. *Soils for Nutrition: State of the Art*. Rome. https://doi.org/10.4060/cc0900en

Gibson, J., Huang, J., & Rozelle, S. (2003). Improving estimates of inequality and poverty from urban China's household income and expenditure survey. *Review of Income and Wealth*, *49*(1), 53-68.

Gourlay, S., Aynekulu, E., Carletto, C., & Shepherd, K., 2017. *Spectral Soil Analysis & Household Surveys: A Guidebook for Integration*. Washington, DC: World Bank.

Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, *et al.* (2015) Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE 10(6).

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748.

Hengl, T., Miller, M.A.E., Križan, J. *et al.* African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci Rep 11, 6130 (2021).

Jenkins, S. P., Burkhauser, R. V., Feng, S., & Larrimore, J. (2011). Measuring inequality using censored data: a multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *174*(1), 63-81.

Kassahun, B. (2015). Soil fertility mapping and fertilizer blending. Ethiopian Agricultural Transformation Agency (Ethiopian ATA) report, Addis Ababa.

Kosmowski, F., Ambel, A., Tsegay, A. H., Negawo, A. T., Carling, J., Kilian, A., & Central Statistics Agency. (2021). A large-scale dataset of barley, maize and sorghum variety identification using DNA fingerprinting in Ethiopia. *Data*, *6*(6), 58.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data. 3rd edition.* John Wiley & Sons.

Lobell, D.B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T. and Murray, S., 2020. Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, 102(1), pp.202-219.

Mukherjee, A., & Lal, R. (2014). Comparison of soil quality index using three methods. *PloS One*, 9(8), e105981.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Nabiollahi, K., Golmohamadi, F., Taghizadeh-Mehrjardi, R., Kerry, R. and Davari, M., 2018. Assessing the effects of slope gradient and land use change on soil quality degradation through digital mapping of soil quality indices and soil loss rate. *Geoderma*, 318, pp.16-28.

National Environment Management Authority (NEMA). (2001). National State of the Environment Report for Uganda, 2000/2001. National Environment Management Authority (NEMA), Kampala. Retrieved from: http://nile.riverawarenesskit.org/english/nrak/Resources/Document_centre/Uganda_SoE_2000.pdf

Nkonya, E., Pender, J., Kaizzi, C., Edward, K., & Mugarura, S. (2005a). Policy Options for Increasing Crop Productivity and Reducing Soil Nutrient Depletion and Poverty in Uganda. Intl Food Policy Res Inst. EPT Discussion paper 134.

Nkonya, E., Kaizzi, C., & Pender, J. (2005b). Determinants of nutrient balances in a maize farming system in eastern Uganda. *Agricultural systems*, 85(2), 155-182.

Park, Colin N. and Arthur L. Dudycha. (1974). "A cross-validation approach to sample size determination for regression models." *Journal of the American Statistical Association,* 69(345): 214-218.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Shepherd, K. D., & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Social Science Society of American Journal*, 66(3), 988–998.

Shepherd, K. D., & Walsh, M. G. (2004). Diffuse reflectance spectroscopy for rapid soil analysis. In Lal, R. (Ed.), *Encyclopedia of Soil Science.* New York: Marcel Dekker.

Shepherd, K. D., & Walsh, M. G. (2007). Infrared spectroscopy: Enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. *Journal of Near Infrared Spectroscopy*, 15, 1–19.

Stoorvogel, J. J., & Smaling, E. M. A. (1990). Assessment of soil nutrient depletion in sub-Saharan Africa: 1983-2000. Report 28. Wageningen, The Netherlands: Winand Staring Centre for Integrated Land, Soil and Water Research.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Vasu, D., Singh, S.K., Ray, S.K., Duraisami, V.P., Tiwary, P., Chandran, P., Nimkar, A.M. and Anantwar, S.G., 2016. Soil quality index (SQI) as a tool to evaluate crop productivity in semi-arid Deccan plateau, India. *Geoderma*, 282, pp.70-79.

World Bank. (2012). "Well Begun, Not Yet Done: Vietnam's Remarkable Progress on Poverty Reduction and the Emerging Challenges". *Vietnam Poverty Assessment Report 2012*. Hanoi: World Bank.

World Bank. (2024). World Development Indicators. Agriculture, forestry, and fishing, value added (% of GDP), Retrieved from https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=ET-UG

Wortmann, C. S., & Kaizzi, C. K. (1998). Nutrient balances and expected effects of alternative practices in farming systems of Uganda. *Agriculture, ecosystems & environment*, 71(1-3), 115-129.

Table 1. Comparison of Top Soil Characteristics Across Sources: Plot-Level (LASER 2013/14 and MAPS1 2015/16) vs Geospatial (SoilGrids 2020 and iSDA 2021)

	Ethiopia (2013/14)			Uganda (2015/16)		
	LASER	**LASER-iSDAsoil**	**LASER-SoilGrids**	**MAPS1**	**MAPS1-iSDAsoil**	**MAPS1-SoilGrids**
Weighted Soil Index	0.53	-0.11***	-0.15***	0.57	-0.08***	-0.08***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
			Components of soil index			
pH	6.23	0.34***	0.16***	6.42	0.49***	0.61***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
Carbon (%)	3.17	1.31***	-0.23***	1.49	0.01	-1.62***
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)
Total Nitrogen (%)	0.27	0.08***	-0.01***	0.11	-0.02***	-0.10***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
			Soil particle composition			
Clay (%)	63.84	29.18***	27.77***	56.46	24.48***	18.75***
	(0.34)	(0.30)	(0.34)	(0.55)	(0.51)	(0.54)
Silt (%)	23.37	-0.63***	-7.60***	20.68	-2.09***	-2.45***
	(0.20)	(0.18)	(0.23)	(0.15)	(0.22)	(0.16)
Sand (%)	12.82	-26.64***	-20.16***	22.86	-22.51***	-16.42***
	(0.19)	(0.18)	(0.27)	(0.44)	(0.42)	(0.42)
Number of plots	1,529			875		

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. LASER and MAPS1 topsoil sample are at a soil depth of 0-20cm, SoilGrids topsoil is a weighted avg of 0-5/5-15cm, and iSDA topsoil is at soil depths of 0-20cm. The weighted soil index includes pH, carbon, and total nitrogen but excludes electrical conductivity. Table C.1 (Appendix C) compares the weighted soil index calculated with electrical conductivity across different sources. The sample is restricted to plots with non-missing soil properties in household surveys and geospatial data sources. The distributions of the soil characteristics are shown in Table A.1, Appendix A (Panel B). The comparison of subsoil across different sources is shown in Table B.1, Appendix B.

Table 2. Benchmark Survey vs. Imputation-Based Estimates (for the zones/districts of the benchmark surveys), Top Soil Analysis

Soil Properties	*Ethiopia (2013/14)*		*Uganda (2015/16)*	
	Benchmark (LASER)	**Imputed (ESS2)**	**Benchmark (MAPS1)**	**Imputed (UNPS5)**
	(1)	*(2)*	*(3)*	*(4)*
Weighted Soil Index	0.53	**0.52**[a]	0.57	**0.56**
	(0.00)	(0.00)	(0.00)	(0.01)
	Components of soil index			
pH	6.25	**6.22**	6.42	**6.45**
	(0.02)	(0.05)	(0.02)	(0.06)
Carbon (%)	3.16	**3.14**[a]	1.49	**1.44**
	(0.03)	(0.07)	(0.03)	(0.09)
Total Nitrogen (%)	0.27	0.27	0.11	0.11
	(0.00)	(0.01)	(0.00)	(0.01)
	Soil particle composition			
Clay (%)	63.91	**64.12**[a]	56.48	**56.04**[a]
	(0.32)	(0.84)	(0.55)	(1.94)
Silt (%)	23.29	**23.48**[a]	20.67	**20.55**[a]
	(0.19)	(0.47)	(0.15)	(0.50)
Sand (%)	12.82	**12.56**	22.85	**23.60**
	(0.18)	(0.50)	(0.44)	(1.62)
Number of plots	1,672	608	877	163

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of benchmark rate. Standard errors are in parentheses. Estimates are obtained using the PMM method, with 100 iterations. The target survey is restricted to the same zones (Ethiopia) or districts (Uganda) as the benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.1 and A.2, Appendix A (Panel A). The estimates obtained using the linear regression method are shown in Table A.3, Appendix A. Imputation models are shown in Tables A.6 and A.7, Appendix A.

Table 3. Benchmark Survey vs. Imputation-Based Estimates with Additional Geospatial Soil Quality Information (for the zones/districts of the benchmark surveys), Top Soil Analysis

Soil Properties	*Ethiopia (2013/14)*			*Uganda (2015/16)*		
	Benchmark (LASER)	**Imputed with iSDAsoil (ESS2)**	**Imputed with SoilGrids (ESS2)**	**Benchmark (MAPS1)**	**Imputed with iSDAsoil (UNPS5)**	**Imputed with SoilGrids (UNPS5)**
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Weighted Soil Index	0.53	**0.52**[a]	**0.52**[a]	0.57	**0.56**	**0.56**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
			Components of soil index			
pH	6.23	**6.23**[a]	**6.22**[a]	6.42	6.45	6.47
	(0.02)	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)
Carbon (%)	3.17	**3.13**	**3.13**	1.49	1.43	**1.45**
	(0.03)	(0.07)	(0.07)	(0.03)	(0.08)	(0.08)
Total Nitrogen (%)	0.27	0.26	**0.26**	0.11	0.11	0.11
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
			Soil particle composition			
Clay (%)	63.84	**64.26**	**64.39**	56.46	**56.45**[a]	**56.14**[a]
	(0.34)	(0.81)	(0.81)	(0.55)	(1.73)	(1.83)
Silt (%)	23.37	**23.27**[a]	**23.31**[a]	20.68	**20.55**[a]	**20.60**[a]
	(0.20)	(0.47)	(0.45)	(0.15)	(0.51)	(0.52)
Sand (%)	12.82	**12.69**[a]	**12.53**	22.86	**23.17**[a]	**23.55**
	(0.19)	(0.50)	(0.46)	(0.44)	(1.48)	(1.49)
Number of plots	1,529	608		875	157	

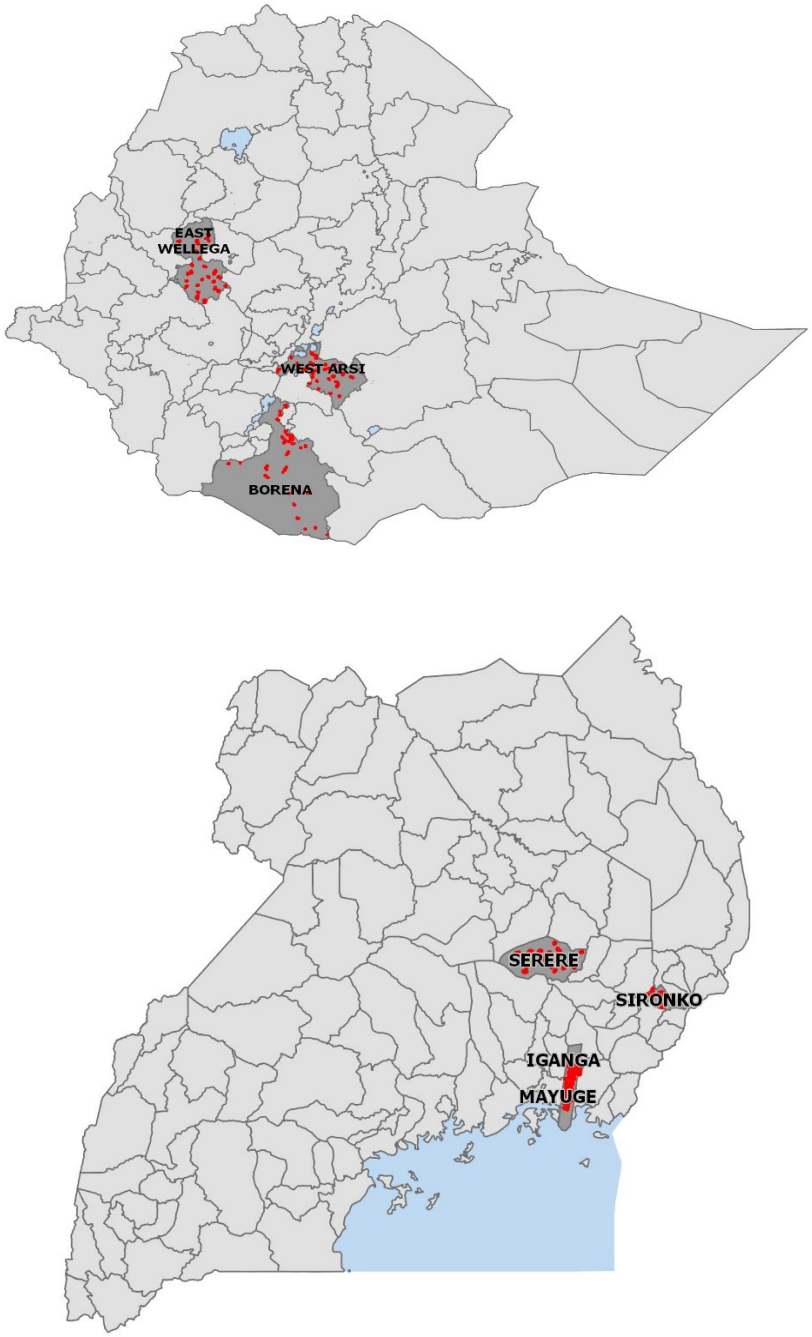
Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Estimates are obtained using the PMM method, with 100 iterations. The target survey is restricted to the same zones (Ethiopia) or districts (Uganda) as the benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.1 and A.2, Appendix A (Panel B). The estimates obtained using the linear regression method are shown in Table A.4, Appendix A. Imputation models are shown in Tables A.8-A.11 for Ethiopia and A.12-A.15 for Uganda, Appendix A.

Table 4. Imputation-Based Estimates with Additional Geospatial Soil Quality Information vs. Geospatial Estimates (for the remaining areas), Top Soil Analysis

Soil Properties	*Ethiopia (2013/14)*				*Uganda (2015/16)*			
	iSDA		**SoilGrids**		**iSDA**		**SoilGrids**	
	Imputed with iSDA (ESS 2)	**iSDA**	**Imputed with SoilGrids (ESS2)**	**SoilGrids**	**Imputed with iSDA (UNPS5)**	**iSDA**	**Imputed with SoilGrids (UNPS5)**	**SoilGrids**
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
Weighted Soil Index	0.52	0.62	0.51	0.64	0.57	0.64	0.56	0.63
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Components of soil index								
pH	6.43	6.12	6.43	6.32	6.43	5.89	6.52	5.89
	(0.02)	(0.00)	(0.02)	(0.01)	(0.02)	(0.00)	(0.02)	(0.01)
Carbon (%)	3.09	1.77	3.07	3.26	1.61	1.52	1.55	3.29
	(0.03)	(0.00)	(0.03)	(0.01)	(0.04)	(0.01)	(0.03)	(0.01)
Total Nitrogen (%)	0.26	0.18	0.27	0.27	0.14	0.15	0.14	0.25
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Soil particle composition								
Clay (%)	65.71	35.73	65.25	37.19	54.87	31.51	53.52	35.67
	(0.33)	(0.03)	(0.32)	(0.04)	(0.62)	(0.09)	(0.70)	(0.07)
Silt (%)	23.29	25.04	22.83	32.09	21.83	20.08	21.46	24.99
	(0.21)	(0.02)	(0.19)	(0.03)	(0.15)	(0.03)	(0.27)	(0.05)
Sand (%)	12.16	37.88	12.10	30.62	24.73	46.97	22.95	38.75
	(0.16)	(0.04)	(0.18)	(0.05)	(0.49)	(0.11)	(0.52)	(0.07)
Number of plots	20,575				4,065			

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Estimates are obtained using the PMM method, with 100 iterations. The target survey is restricted to the remaining zones (Ethiopia) or districts (Uganda), other than the zones/districts of the benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.1 and A.2, Appendix A (Panel B). The estimates obtained using the linear regression method are shown in Table A.5, Appendix A. Imputation models are shown in Tables A.8-A.11 for Ethiopia and A.12-A.15 for Uganda, Appendix A.

Figure 1. Soil data collection areas for Ethiopia's LASER **study (top) and Uganda's MAPS1 study (bottom)**



Note: darker gray color indicate areas where soil samples were collected: Borena, East Wellega and West Arsi zones of Oromia region in Ethiopia and Serere, Sironko, Iganga, and Mayuge districts of Uganda's Eastern region (Serere district was located in Soroti, Uganda). The red markers indicate plots where soil samples are collected from interviewed housheolds.

Figure 2. Ratio of Estimates Across Different Sources to the Benchmark Survey, Top Soil Analysis, Ethiopia (2013/14) and Uganda (2015/16)

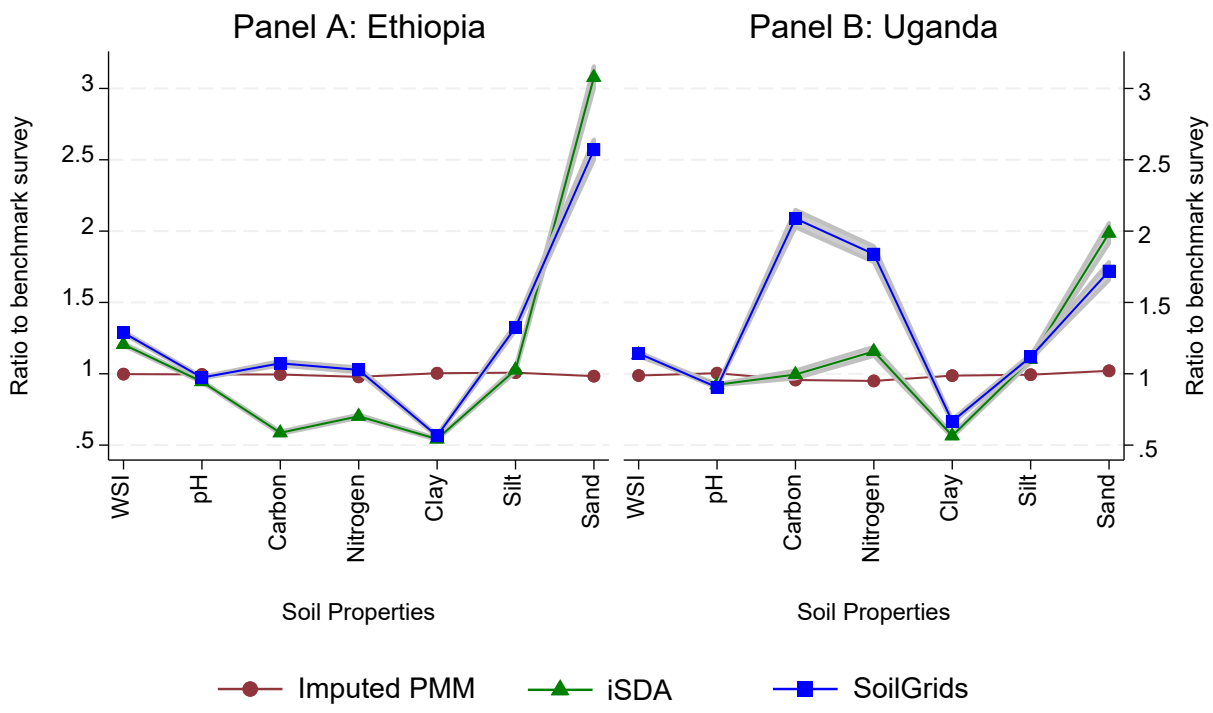
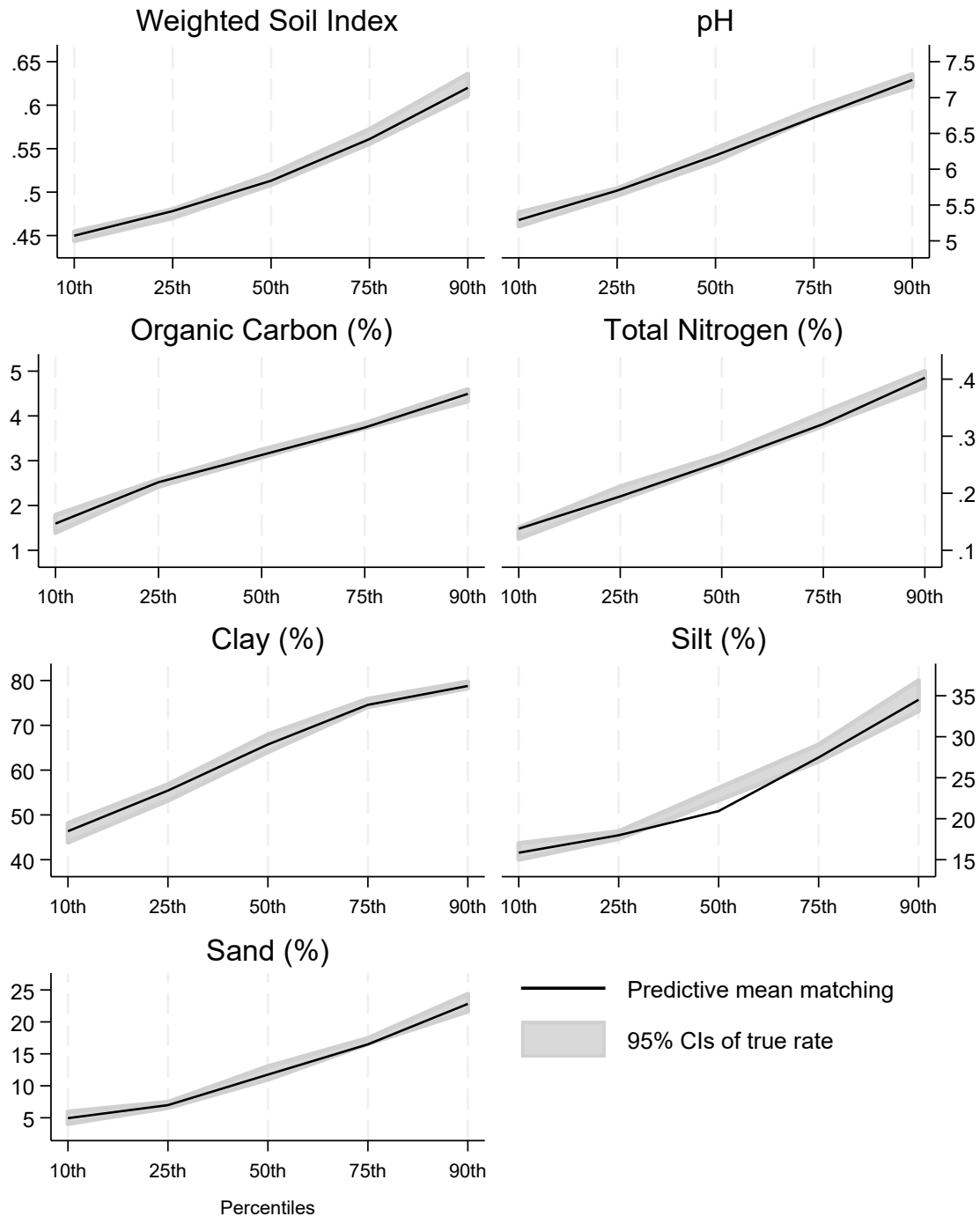
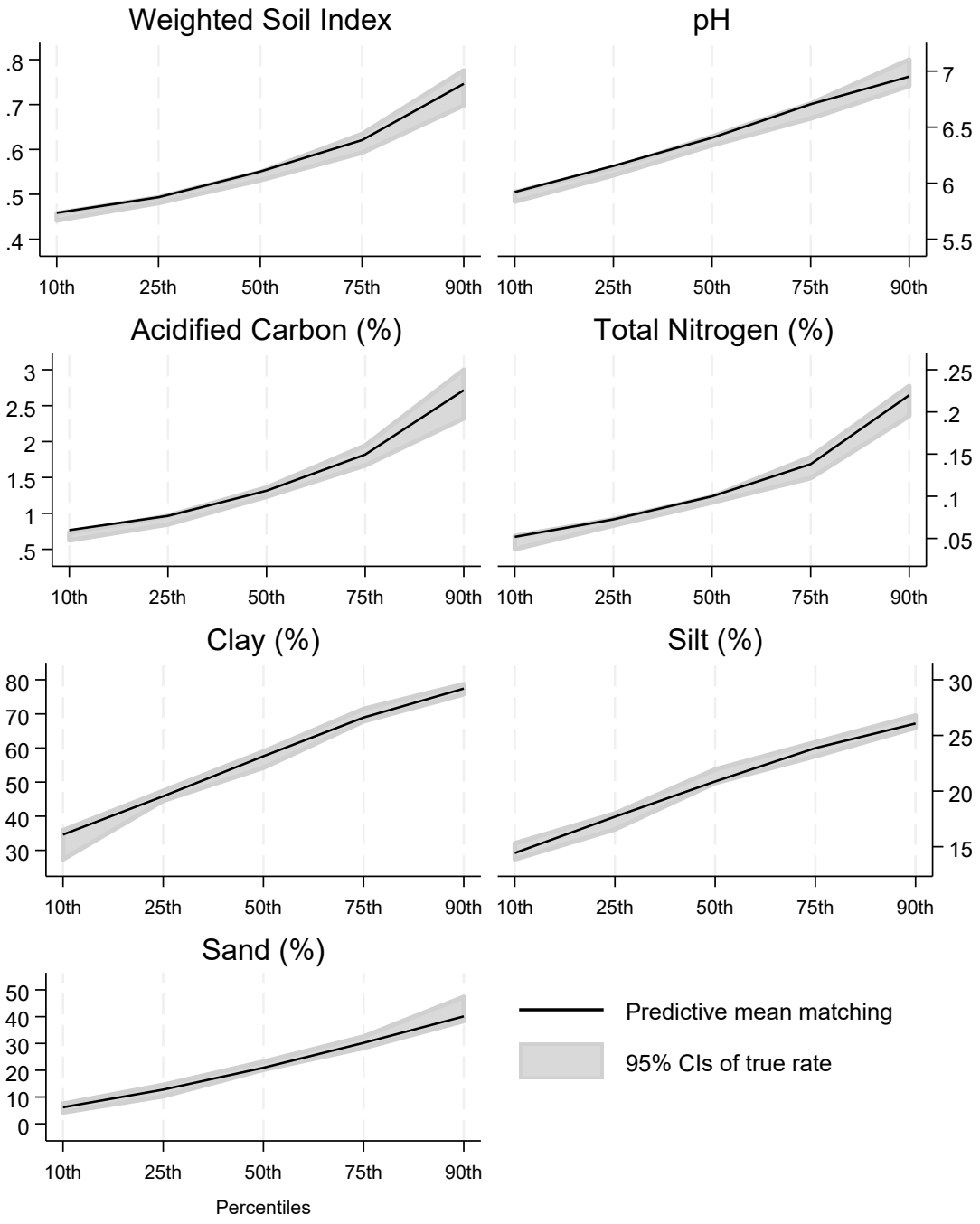


Figure 3. Imputation-based Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Sample (Base Survey), Top Soil Analysis, Ethiopia (2013/14)



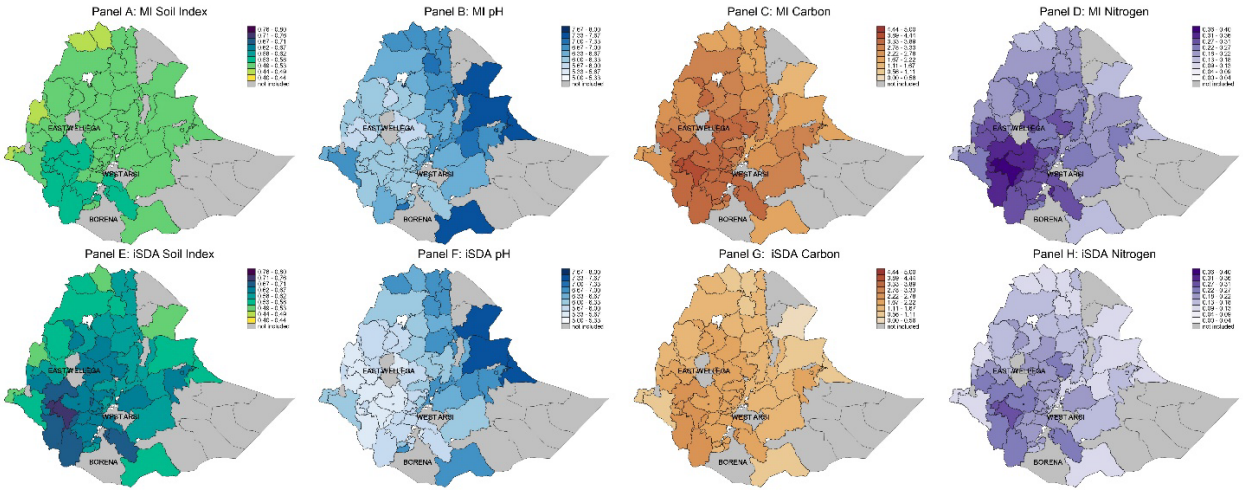
Note: The estimation sample is generated by splitting LASER data into two random samples: 50% in-sample, 50% out-of-sample. Estimates are obtained using 50 iterations using "out-of-sample" data as the target sample. Simultaneous quintile regression with bootstrapping SEs was used to get multiple imputed quintiles in the target sample. The total "in-sample" size is 836 plots, the "out-of-sample" size is 836 plots.

Figure 4. Imputation-based Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Sample (Base Survey), Top Soil Analysis, Uganda (2015/16)



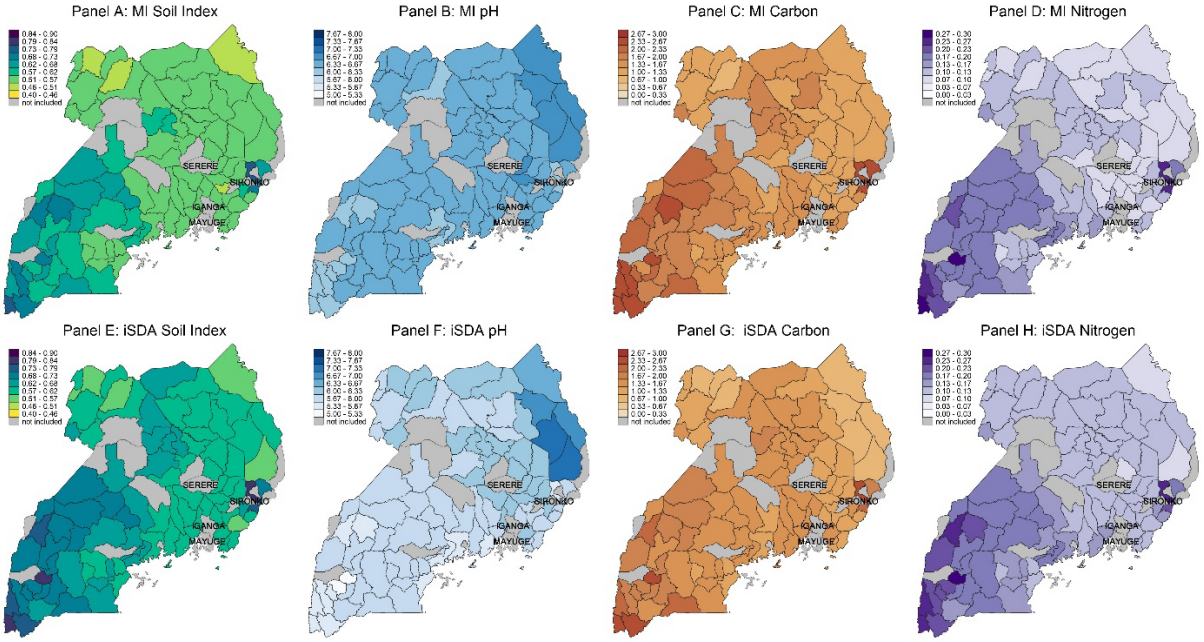
Note: The estimation sample is generated by splitting MAPS1 data into two random samples: 50% in-sample, 50% out-of-sample. Estimates are obtained using 50 iterations using "out-of-sample" data as the target sample. Simultaneous quintile regression with bootstrapping SEs was used to get multiple imputed quintiles in the target sample. The total "in-sample" size is 438 plots, the "out-of-sample" size is 439 plots.

Figure 5. Imputation-Based Estimates vs. iSDA Estimates, Top Soil Analysis, Ethiopia



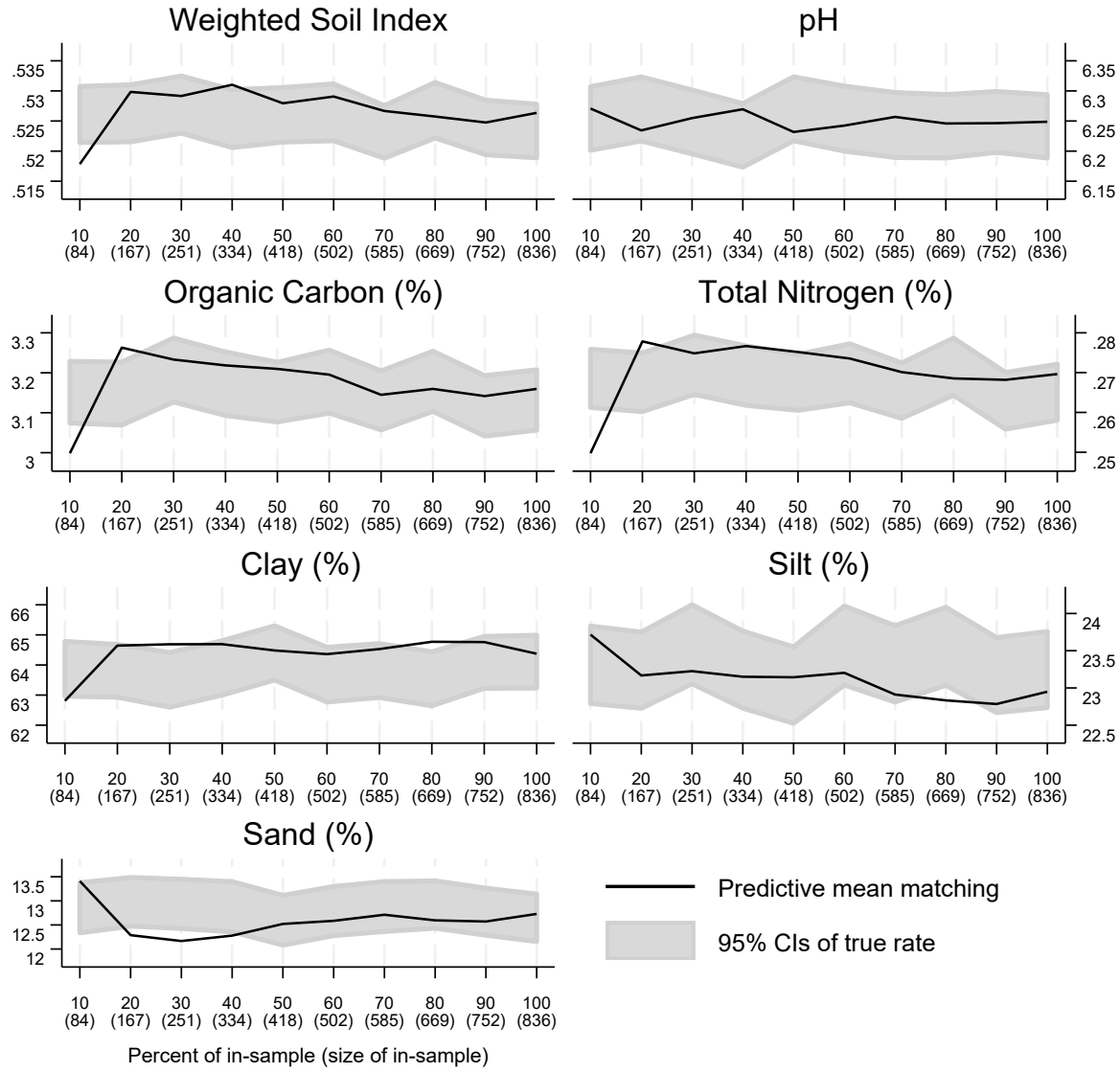
Notes: Estimates are aggregated by regions (where households were interviewed). Panels A-D show imputed estimates where the prediction model includes iSDA geospatial variables. Panels E-H show geospatial iSDA estimates. Organic carbon is used in Panel C and Panel G. The coefficient of variation for Panel A is 12.7%, for Panel B is 12.9%, for Panel C is 37.8%, for Panel D is 40.9%, for Panel E is 9.6%, for Panel F is 9.5%, for Panel G is 32.7%, for Panel H is 31.8%.

Figure 6. Imputation-Based Estimates vs. iSDA Estimates, Top Soil Analysis, Uganda



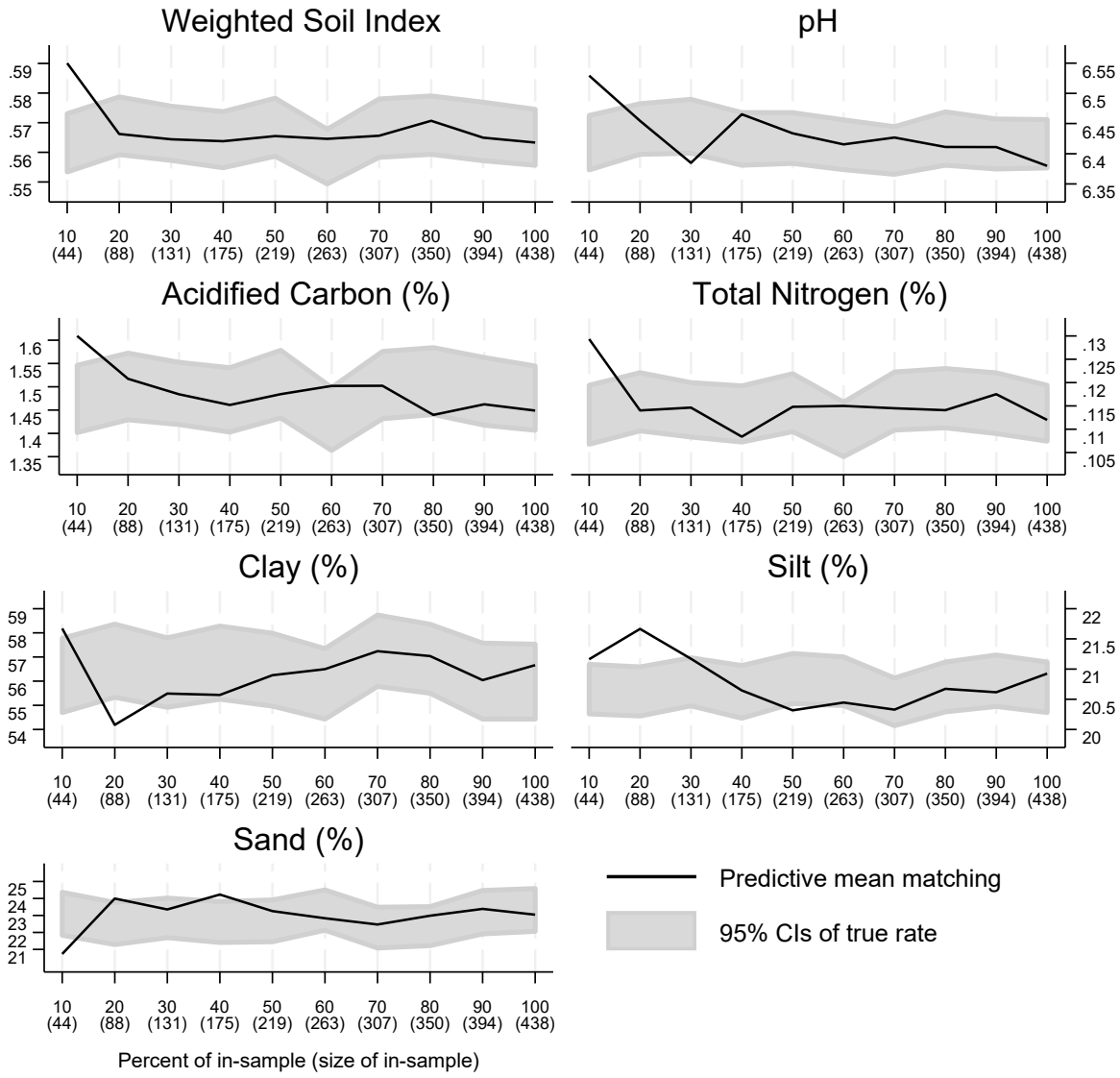
Note: Estimates are aggregated by districts (where households were interviewed). Benchmark survey districts (Iganga, Sironko and Mayge) and districts with missing imputed estimates (Bundibugyo, Nakasongola, Kaberamaido, Kalangala, Kampala, Kapchorwa) are omitted. Panels A-D show imputed estimates where the prediction model includes iSDA geospatial variables. Acidified carbon is used in Panel C and Panel G. Panels E-H show geospatial iSDA estimates. The coefficient of variation for Panel A is 18.3%, for Panel B is 7.1%, for Panel C is 51.2%, for Panel D is 51.5%, for Panel E is 9.9%. for Panel F is 4.1%, for Panel G is 24%, for Panel H is 28.6%.

Figure 7. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Sample, Top Soil Analysis, Ethiopia (2013/14)



Note: The estimation sample is generated by splitting LASER data into two random samples: 50% in-sample, and 50% out-of-sample. Benchmark sample is selected as a percentage varying from (randomly selected) 10% to 100% of the "in-sample" data (with the number of observations shown in parentheses). Estimates are obtained with 50 iterations using 100% of the "out-of-sample" data as the target sample. The total "in-sample" size is 836 plots, the "out-of-sample" size is 836 plots.

Figure 8. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Sample, Top Soil Analysis, Uganda (2015/16)



Note: The estimation sample is generated by splitting MAPS1 data into two random samples: 50% in-sample, 50% out-of-sample. Benchmark sample is selected as a percentage varying from (randomly selected) 10% to 100% of the "in-sample" data (with the number of observations shown in parentheses). Estimates are obtained with 50 iterations using 100% of the "out-of-sample" data as the target sample. The total "in-sample" size is 438 plots, the "out-of-sample" size is 439 plots.

Appendix A. Additional Tables and Figures

Table A.1. Summary Statistics, Ethiopia (2013/14)

Variables	LASER		ESS2								
			Full Sample			Benchmark Zones			Remaining Zones		
	Mean	*SD*	Mean	*SD*	KS p-value	Mean	*SD*	KS p-value	Mean	*SD*	KS p-value
Panel A: Sample without geospatial variables											
Weighted Soil Index	0.53	(0.07)									
pH	6.25	(0.78)									
Organic Carbon (%)	3.16	(1.14)									
Total Nitrogen (%)	0.27	(0.11)									
Clay (%)	63.91	(13.24)									
Silt (%)	23.29	(7.63)									
Sand (%)	12.82	(7.56)									
Control variables											
Head`s age	43.68	(15.64)	47.20	(14.29)	0.000	44.97	(15.05)	0.000	47.27	(14.27)	0.000
Head`s gender (male==1)	0.83	(0.38)	0.82	(0.38)	1.000	0.83	(0.38)	1.000	0.82	(0.38)	1.000
Head`s education											
Primary	0.40	(0.49)	0.31	(0.46)	0.000	0.52	(0.50)	0.000	0.30	(0.46)	0.000
Secondary	0.05	(0.22)	0.03	(0.17)	0.577	0.05	(0.21)	1.000	0.03	(0.17)	0.546
Higher	0.01	(0.11)	0.01	(0.08)	1.000	0.01	(0.11)	1.000	0.01	(0.08)	1.000
Log of household size	1.68	(0.45)	1.75	(0.43)	0.000	1.88	(0.41)	0.000	1.74	(0.43)	0.000
Plot characteristics											
Type of crop stand (1 – pure stand, 0 – mixed stand)	0.77	(0.42)	0.78	(0.42)	1.000	0.88	(0.32)	0.000	0.77	(0.42)	1.000
Crop rotation	0.67	(0.47)	0.85	(0.36)	0.000	0.88	(0.32)	0.000	0.85	(0.36)	0.000
Plot is prevented from erosion	0.39	(0.49)	0.57	(0.50)	0.000	0.30	(0.46)	0.001	0.57	(0.49)	0.000
Plot is tilled	0.94	(0.23)	0.88	(0.33)	0.000	0.85	(0.36)	0.001	0.88	(0.33)	0.000
Fertilizer (inorganic) is used on plot	0.25	(0.44)	0.24	(0.43)	0.966	0.24	(0.43)	1.000	0.24	(0.43)	0.968
Fertilizer (organic) is used on the plot	0.18	(0.39)	0.32	(0.47)	0.000	0.27	(0.45)	0.001	0.33	(0.47)	0.000
Pesticides are used in the plot	0.07	(0.26)	0.10	(0.31)	0.117	0.26	(0.44)	0.000	0.10	(0.30)	0.263
Hired labor	0.04	(0.19)	0.08	(0.28)	0.002	0.12	(0.33)	0.004	0.08	(0.28)	0.003
Log of the plot area	6.43	(1.62)	5.89	(2.09)	0.000	6.00	(2.53)	0.000	5.89	(2.08)	0.000
Number of plots	1,672		21,278			608			20,670		
Panel B: Sample with geospatial variables											
Weighted Soil Index	0.53	(0.07)									
pH	6.23	(0.77)									
Organic Carbon (%)	3.17	(1.14)									
Total Nitrogen (%)	0.27	(0.11)									

Clay (%)	63.84	(13.44)									
Silt (%)	23.37	(7.74)									
Sand (%)	12.82	(7.62)									
			Control variables								
Head`s age	43.76	(15.51)	47.17	(14.29)	0.000	44.97	(15.05)	0.000	47.24	(14.26)	0.000
Head`s gender (male==1)	0.82	(0.39)	0.83	(0.38)	1.000	0.83	(0.38)	1.000	0.83	(0.38)	1.000
Head`s education											
Primary	0.39	(0.49)	0.31	(0.46)	0.000	0.52	(0.50)	0.000	0.30	(0.46)	0.000
Secondary	0.05	(0.22)	0.03	(0.17)	0.579	0.05	(0.21)	1.000	0.03	(0.17)	0.549
Higher	0.01	(0.10)	0.01	(0.08)	1.000	0.01	(0.11)	1.000	0.01	(0.08)	1.000
Log of household size	1.67	(0.45)	1.75	(0.43)	0.000	1.88	(0.41)	0.000	1.74	(0.43)	0.000
Plot characteristics											
Type of crop stand (1 – pure stand, 0 – mixed stand)	0.79	(0.41)	0.78	(0.42)	0.989	0.88	(0.32)	0.001	0.77	(0.42)	0.910
Crop rotation	0.69	(0.46)	0.85	(0.36)	0.000	0.88	(0.32)	0.000	0.85	(0.36)	0.000
Plot is prevented from erosion	0.39	(0.49)	0.57	(0.50)	0.000	0.30	(0.46)	0.003	0.57	(0.49)	0.000
Plot is tilled	0.95	(0.23)	0.88	(0.33)	0.000	0.85	(0.36)	0.001	0.88	(0.33)	0.000
Fertilizer (inorganic) is used on plot	0.26	(0.44)	0.24	(0.43)	0.414	0.24	(0.43)	1.000	0.24	(0.43)	0.418
Fertilizer (organic) is used on the plot	0.19	(0.40)	0.32	(0.47)	0.000	0.27	(0.45)	0.003	0.33	(0.47)	0.000
Pesticides are used in the plot	0.08	(0.27)	0.10	(0.31)	0.236	0.26	(0.44)	0.000	0.10	(0.30)	0.427
Hired labor	0.04	(0.19)	0.08	(0.28)	0.011	0.12	(0.33)	0.000	0.08	(0.28)	0.007
Log of the plot area	6.39	(1.65)	5.89	(2.09)	0.000	6.00	(2.53)	0.000	5.89	(2.08)	0.000
iSDA variables											
Weighted Soil Index	0.63	(0.07)	0.62	(0.06)	0.000	0.64	(0.03)	0.000	0.62	(0.06)	0.000
Ph	5.89	(0.46)	6.11	(0.58)	0.000	5.81	(0.43)	0.000	6.12	(0.58)	0.000
Organic carbon	1.86	(0.54)	1.78	(0.57)	0.000	1.97	(0.35)	0.000	1.77	(0.58)	0.000
Total nitrogen	0.19	(0.06)	0.18	(0.06)	0.000	0.20	(0.04)	0.000	0.18	(0.06)	0.000
Clay	34.66	(5.65)	35.75	(4.79)	0.000	36.34	(4.25)	0.000	35.73	(4.80)	0.000
Silt	24.01	(3.03)	25.06	(2.66)	0.000	25.90	(1.83)	0.000	25.04	(2.68)	0.000
Sand	39.46	(8.00)	37.86	(6.40)	0.000	36.98	(4.58)	0.000	37.88	(6.45)	0.000
SoilGrids variables											
Weighted Soil Index	0.68	(0.07)	0.64	(0.06)	0.000	0.65	(0.04)	0.000	0.64	(0.06)	0.000
Ph	6.07	(0.53)	6.31	(0.72)	0.000	5.93	(0.54)	0.000	6.32	(0.72)	0.000
Organic carbon	3.40	(0.88)	3.27	(0.92)	0.000	3.64	(0.75)	0.000	3.26	(0.92)	0.000
Total nitrogen	0.28	(0.08)	0.27	(0.08)	0.000	0.30	(0.06)	0.000	0.27	(0.08)	0.000
Clay	36.07	(6.62)	37.20	(5.83)	0.000	37.64	(4.76)	0.000	37.19	(5.86)	0.000
Silt	30.98	(6.12)	32.14	(4.46)	0.000	33.91	(5.19)	0.000	32.09	(4.43)	0.000
Sand	32.95	(10.71)	30.55	(6.78)	0.000	28.45	(7.17)	0.000	30.62	(6.76)	0.000
Number of plots	1,529		21,183			608			20,575		

Note: Standard errors are in parentheses. Kolmogorov-Smirnov (KS) p-values test if there are differences in the distribution of variables from the LASER survey. The null hypothesis is that the distributions are equal. To be consistent with the benchmark surveys, we restrict the target surveys to cultivated plots only.

Table A.2. Summary Statistics, Uganda (2015/16)

Variables	MAPS1		UNPS5								
			Full Sample			Benchmark Districts			Remaining Districts		
	Mean	*SD*	Mean	*SD*	KS p-value	Mean	*SD*	KS p-value	Mean	*SD*	KS p-value
Panel A: Sample without geospatial variables											
Weighted Soil Index	0.57	(0.10)									
pH	6.42	(0.46)									
Acidified Carbon (%)	1.49	(0.76)									
Total Nitrogen (%)	0.11	(0.07)									
Clay (%)	56.48	(16.19)									
Silt (%)	20.67	(4.43)									
Sand (%)	22.85	(13.15)									
Control variables											
Head`s age	43.59	(14.98)	47.67	(15.54)	0.000	48.74	(14.07)	0.000	47.63	(15.59)	0.000
Head`s gender (male==1)	0.79	(0.40)	0.71	(0.45)	0.000	0.77	(0.42)	1.000	0.71	(0.45)	0.000
Head`s education											
Primary education	0.61	(0.49)	0.59	(0.49)	0.689	0.52	(0.50)	0.142	0.59	(0.49)	0.810
Secondary education	0.24	(0.43)	0.19	(0.39)	0.038	0.21	(0.41)	1.000	0.19	(0.39)	0.033
Vocational education	0.03	(0.16)	0.09	(0.28)	0.015	0.13	(0.34)	0.083	0.08	(0.28)	0.021
Higher education	0.01	(0.09)	0.01	(0.09)	1.000	0.02	(0.13)	1.000	0.01	(0.09)	1.000
Log of household size	1.66	(0.61)	1.52	(0.63)	0.000	1.63	(0.63)	0.142	1.51	(0.63)	0.000
Plot characteristics											
Type of crop stand (1 – pure stand, 0 – mixed stand)	0.43	(0.49)	0.62	(0.49)	0.000	0.42	(0.50)	0.185	0.63	(0.49)	0.000
Fertilizer (inorganic) is used on plot	0.09	(0.29)	0.02	(0.13)	0.000	0.01	(0.08)	1.000	0.02	(0.13)	0.000
Fertilizer (organic) is used on the plot	0.16	(0.36)	0.01	(0.11)	0.000	0.03	(0.17)	0.245	0.01	(0.11)	0.000
Pesticides are used in the plot	0.04	(0.20)	0.03	(0.18)	1.000	0.01	(0.11)	0.028	0.03	(0.18)	1.000
Cultivated in the previous season	0.81	(0.40)	0.96	(0.20)	0.000	0.96	(0.20)	1.000	0.96	(0.20)	0.000
Log of the parcel area	0.74	(0.49)	0.87	(0.55)	0.000	0.65	(0.38)	0.004	0.88	(0.56)	0.000
Problems with erosion	0.34	(0.47)	0.10	(0.30)	0.000	0.09	(0.29)	0.448	0.10	(0.30)	0.000
Hired labor	0.33	(0.47)	0.20	(0.40)	0.000	0.17	(0.38)	0.000	0.20	(0.40)	0.000
Number of plots	877		4,466			163			4,303		
Panel B: Sample with geospatial variables											
Weighted Soil Index	0.57	(0.10)									
pH	6.42	(0.46)									
Acidified Carbon (%)	1.49	(0.76)									
Total Nitrogen (%)	0.11	(0.07)									
Clay (%)	56.46	(16.20)									
Silt (%)	20.68	(4.43)									

Sand (%)	22.86	(13.16)									
				Control variables							
Head`s age	43.59	(14.99)	47.84	(15.51)	0.000	49.37	(13.85)	0.000	47.78	(15.56)	0.000
Head`s gender (male==1)	0.79	(0.40)	0.72	(0.45)	0.000	0.78	(0.41)	1.000	0.71	(0.45)	0.000
Head`s education											
Primary education	0.61	(0.49)	0.59	(0.49)	0.881	0.53	(0.50)	0.291	0.59	(0.49)	0.950
Secondary education	0.24	(0.43)	0.18	(0.39)	0.033	0.20	(0.40)	0.998	0.18	(0.39)	0.030
Vocational education	0.03	(0.16)	0.08	(0.27)	0.031	0.13	(0.33)	0.140	0.08	(0.27)	0.042
Higher education	0.01	(0.09)	0.01	(0.09)	1.000	0.02	(0.14)	1.000	0.01	(0.09)	1.000
Log of household size	1.66	(0.61)	1.52	(0.63)	0.000	1.63	(0.63)	0.140	1.51	(0.63)	0.000
Plot characteristics											
Type of crop stand (1 – pure stand, 0 – mixed stand)	0.43	(0.49)	0.62	(0.49)	0.000	0.41	(0.49)	1.000	0.63	(0.49)	0.000
Fertilizer (inorganic) is used on plot	0.09	(0.29)	0.02	(0.13)	0.000	0.01	(0.08)	0.262	0.02	(0.13)	0.000
Fertilizer (organic) is used on the plot	0.16	(0.36)	0.01	(0.12)	0.000	0.03	(0.18)	0.034	0.01	(0.11)	0.000
Pesticides are used in the plot	0.04	(0.20)	0.03	(0.18)	1.000	0.01	(0.08)	0.993	0.03	(0.18)	1.000
Cultivated in the previous season	0.81	(0.40)	0.96	(0.20)	0.000	0.96	(0.21)	0.005	0.96	(0.20)	0.000
Log of the parcel area	0.74	(0.49)	0.89	(0.56)	0.000	0.65	(0.39)	0.489	0.89	(0.56)	0.000
Problems with erosion	0.34	(0.47)	0.10	(0.30)	0.000	0.10	(0.29)	0.000	0.10	(0.30)	0.000
Hired labor	0.33	(0.47)	0.20	(0.40)	0.000	0.17	(0.38)	0.003	0.21	(0.40)	0.000
iSDA variables											
Weighted Soil Index	0.64	(0.07)	0.64	(0.06)	0.000	0.64	(0.07)	0.000	0.64	(0.06)	0.000
Ph	5.93	(0.19)	5.89	(0.24)	0.000	5.86	(0.16)	0.000	5.89	(0.24)	0.000
Organic carbon	1.48	(0.37)	1.52	(0.36)	0.000	1.51	(0.32)	0.000	1.52	(0.36)	0.000
Total nitrogen	0.13	(0.04)	0.15	(0.04)	0.000	0.15	(0.04)	0.000	0.15	(0.04)	0.000
Clay	31.93	(5.11)	31.59	(5.73)	0.001	33.87	(4.32)	0.000	31.51	(5.76)	0.000
Silt	22.79	(3.42)	20.19	(2.31)	0.000	22.82	(2.85)	0.014	20.08	(2.22)	0.000
Sand	45.40	(6.91)	46.76	(7.18)	0.000	41.20	(6.29)	0.000	46.97	(7.12)	0.000
SoilGrids variables											
Weighted Soil Index	0.65	(0.06)	0.63	(0.08)	0.000	0.58	(0.12)	0.000	0.63	(0.07)	0.000
Ph	5.82	(0.21)	5.87	(0.49)	0.000	5.48	(1.05)	0.000	5.89	(0.45)	0.000
Organic carbon	3.11	(0.86)	3.28	(0.79)	0.000	3.09	(0.84)	0.000	3.29	(0.79)	0.000
Total nitrogen	0.21	(0.06)	0.24	(0.06)	0.000	0.21	(0.05)	0.000	0.25	(0.06)	0.000
Clay	37.66	(1.85)	35.66	(4.62)	0.000	35.24	(7.11)	0.000	35.67	(4.50)	0.000
Silt	23.13	(1.78)	24.88	(3.16)	0.000	22.14	(4.33)	0.104	24.99	(3.06)	0.000
Sand	39.32	(2.31)	38.70	(4.74)	0.000	37.33	(6.89)	0.000	38.75	(4.63)	0.000
Number of plots	875		4,222			157			4,065		

Note: Standard errors are in parentheses. Kolmogorov-Smirnov (KS) p-values test if there are differences in the distribution of variables from the MAPS1 survey. The null hypothesis is that the distributions are equal. To be consistent with the benchmark surveys, we restrict the target surveys to cultivated plots only.

Table A.3. Benchmark Survey vs. Imputation-Based Estimates (for the zones/districts of the benchmark surveys) using Linear Regression Method, Top Soil Analysis

	Ethiopia (2013/14)		*Uganda (2015/16)*	
	Benchmark LASER	**Imputed ESS2**	**Benchmark MAPS1**	**Imputed UNPS5**
Weighted Soil Index	0.53	**0.53**[a]	0.57	**0.56**
	(0.00)	(0.00)	(0.00)	(0.01)
pH	6.25	**6.22**	6.42	6.46
	(0.02)	(0.05)	(0.02)	(0.06)
Carbon (%)	3.16	**3.18**[a]	1.49	**1.44**
	(0.03)	(0.08)	(0.03)	(0.09)
Total Nitrogen (%)	0.27	**0.27**[a]	0.11	0.11
	(0.00)	(0.01)	(0.00)	(0.01)
Clay (%)	63.91	**64.31**	56.48	**56.01**[a]
	(0.32)	(0.89)	(0.55)	(2.00)
Silt (%)	23.29	**23.21**[a]	20.67	**20.51**
	(0.19)	(0.51)	(0.15)	(0.56)
Sand (%)	12.82	**12.50**[a]	22.85	**23.08**[a]
	(0.18)	(0.51)	(0.44)	(1.63)
Number of plots	1,672	608	877	163

Note: Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of benchmark rate. Standard errors are in parentheses. Estimates are obtained using the linear regression method. The target survey is restricted to the same zones (Ethiopia) or districts (Uganda) as the benchmark survey.

Table A.4. Benchmark Survey vs. Imputation-Based Estimates with Additional Geospatial Soil Quality Information (for the zones/districts of the benchmark surveys) using Linear Regression Method, Top Soil Analysis

	Ethiopia (2013/14)			*Uganda (2015/16)*		
	Benchmark (LASER)	**Imputed with iSDAsoil (ESS2)**	**Imputed with SoilGrids (ESS2)**	**Benchmark (MAPS1)**	**Imputed with iSDAsoil (UNPS5)**	**Imputed with SoilGrids (UNPS5)**
Weighted Soil Index	0.53	**0.53**[a]	**0.53**[a]	0.57	**0.56**[a]	**0.56**[a]
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
		Components of soil index				
pH	6.23	**6.21**[a]	**6.20**	6.42	6.47	6.47
	(0.02)	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)
Carbon (%)	3.17	**3.19**[a]	**3.18**[a]	1.49	**1.46**	**1.46**
	(0.03)	(0.07)	(0.07)	(0.03)	(0.09)	(0.09)
Total Nitrogen (%)	0.27	**0.27**[a]	**0.27**[a]	0.11	**0.11**	**0.11**
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
		Soil particle composition				
Clay (%)	63.84	**64.36**	**64.14**[a]	56.46	**56.37**[a]	**56.35**[a]
	(0.34)	(0.92)	(0.96)	(0.55)	(1.73)	(1.87)
Silt (%)	23.37	**23.34**[a]	**23.20**[a]	20.68	**20.59**[a]	**20.60**[a]
	(0.20)	(0.51)	(0.54)	(0.15)	(0.53)	(0.53)
Sand (%)	12.82	**12.63**[a]	**12.52**	22.86	**23.20**[a]	**23.20**[a]
	(0.19)	(0.52)	(0.55)	(0.44)	(1.52)	(1.52)
Number of plots	1,529	608		875	157	

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Target survey is restricted to the remaining zones (Ethiopia) or districts (Uganda), other than the zones/districts of benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.4 and A.5, Appendix A.

Table A.5. Imputation-Based Estimates with Additional Geospatial Soil Quality Information vs. Geospatial Estimates (for the remaining areas) using Linear Regression Method, Top Soil Analysis

Soil Properties	*Ethiopia (2013/14)*				*Uganda (2015/16)*			
	iSDAsoil		**SoilGrids**		**iSDAsoil**		**SoilGrids**	
	Imputed with iSDAsoil (ESS2)	**iSDAsoil**	**Imputed with SoilGrids (ESS2)**	**SoilGrids**	**Imputed with iSDAsoil (UNPS5)**	**iSDAsoil**	**Imputed with SoilGrids (UNPS5)**	**SoilGrids**
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
Weighted Soil Index	0.52	0.62	0.51	0.64	0.57	0.64	0.55	0.63
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
			Components of soil index					
pH	6.42	6.12	6.42	6.32	6.42	5.89	6.47	5.89
	(0.03)	(0.00)	(0.03)	(0.01)	(0.02)	(0.00)	(0.03)	(0.01)
Carbon (%)	3.05	1.77	3.05	3.26	1.56	**1.52**	1.40	3.29
	(0.04)	(0.00)	(0.04)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)
Total Nitrogen (%)	0.26	0.18	0.27	**0.27**	0.14	0.15	0.11	0.25
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
			Soil particle composition					
Clay (%)	65.61	35.73	65.12	37.19	54.55	31.51	54.09	35.67
	(0.48)	(0.03)	(0.54)	(0.04)	(0.73)	(0.09)	(0.86)	(0.07)
Silt (%)	23.44	25.04	22.78	32.09	21.77	20.08	21.11	24.99
	(0.31)	(0.02)	(0.31)	(0.03)	(0.22)	(0.03)	(0.24)	(0.05)
Sand (%)	12.08	37.88	12.10	30.62	24.58	46.97	24.86	38.75
	(0.26)	(0.04)	(0.29)	(0.05)	(0.61)	(0.11)	(0.69)	(0.07)
Number of plots	20,575				4,065			

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Target survey is restricted to the remaining zones (Ethiopia) or districts (Uganda), other than the zones/districts of benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.1 and A.2, Appendix A.

Table A.6. Imputation Model, Top Soil, Ethiopia (2013/14)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	-0.000	-0.001	-0.002	-0.000	0.012	-0.009	-0.003
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.01)
Head`s gender (male==1)	-0.013***	-0.031	-0.213***	-0.014*	2.557***	-2.382***	-0.226
	(0.00)	(0.05)	(0.08)	(0.01)	(0.94)	(0.52)	(0.54)
Head`s education							
Primary school	0.004	-0.089**	0.110*	0.010*	-0.121	0.492	-0.319
	(0.00)	(0.04)	(0.06)	(0.01)	(0.74)	(0.41)	(0.43)
Secondary education (+vocational)	0.022***	-0.027	0.402***	0.037***	0.163	1.066	-1.153
	(0.01)	(0.09)	(0.13)	(0.01)	(1.55)	(0.86)	(0.89)
Higher education	-0.025*	-0.454***	-0.058	-0.010	1.430	0.961	-2.329
	(0.02)	(0.17)	(0.25)	(0.02)	(2.99)	(1.67)	(1.72)
Log of household size	0.012***	0.053	0.168***	0.011*	-3.137***	1.908***	1.239***
	(0.00)	(0.04)	(0.06)	(0.01)	(0.76)	(0.43)	(0.44)
Plot characteristics							
Type of crop stand (1 – pure stand, 0 – mixed stand)	-0.014***	-0.160***	-0.056	-0.013**	0.601	0.277	-0.892**
	(0.00)	(0.05)	(0.07)	(0.01)	(0.79)	(0.44)	(0.45)
Crop rotation (yes/no)	-0.022***	-0.024	-0.359***	-0.038***	1.978***	-1.484***	-0.478
	(0.00)	(0.04)	(0.06)	(0.01)	(0.74)	(0.41)	(0.43)
Plot is prevented from erosion	-0.003	-0.214***	0.132**	0.013**	1.853***	-1.271***	-0.554
	(0.00)	(0.04)	(0.06)	(0.01)	(0.67)	(0.37)	(0.39)
Plot is tilled	-0.001	-0.105	0.121	0.010	-4.921***	4.287***	0.631
	(0.01)	(0.09)	(0.12)	(0.01)	(1.50)	(0.84)	(0.86)
Fertilizer (inorganic) is used on plot	0.005	-0.260***	0.310***	0.021***	-4.063***	4.746***	-0.673
	(0.00)	(0.05)	(0.07)	(0.01)	(0.86)	(0.48)	(0.50)
Fertilizer (organic) is used on the plot	0.000	-0.019	-0.033	0.003	-2.032**	0.502	1.547***
	(0.00)	(0.05)	(0.07)	(0.01)	(0.84)	(0.47)	(0.48)
Pesticides are used in plot	-0.007	-0.080	-0.030	-0.012	0.241	0.127	-0.413
	(0.01)	(0.08)	(0.11)	(0.01)	(1.31)	(0.73)	(0.75)
Hired labor	0.008	0.089	0.031	0.003	1.185	0.037	-1.145
	(0.01)	(0.10)	(0.15)	(0.01)	(1.75)	(0.98)	(1.01)
Log of the plot area	-0.009***	0.072***	-0.221***	-0.019***	-0.526**	-0.173	0.695***
	(0.00)	(0.01)	(0.02)	(0.00)	(0.22)	(0.12)	(0.13)
_cons	0.603***	6.180***	4.560***	0.400***	73.390***	19.222***	7.390***
	(0.01)	(0.14)	(0.20)	(0.02)	(2.46)	(1.37)	(1.41)
r2_a	0.09	0.07	0.11	0.11	0.04	0.10	0.03
N	1672	1672	1672	1672	1672	1672	1672

Table A.7. Imputation Model, Top Soil, Uganda (2015/16)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.001***	0.002*	0.005***	0.000**	0.080**	-0.010	-0.071**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.01)	(0.03)
Head`s gender (male==1)	0.018**	-0.042	0.138**	0.013**	2.112	-0.122	-1.990*
	(0.01)	(0.04)	(0.06)	(0.01)	(1.35)	(0.38)	(1.09)
Head`s education							
Primary education	0.006	0.048	0.055	-0.001	-0.439	0.221	0.217
	(0.01)	(0.05)	(0.08)	(0.01)	(1.80)	(0.51)	(1.46)
Secondary education	-0.006	0.028	-0.026	-0.007	-1.128	-0.039	1.167
	(0.01)	(0.06)	(0.09)	(0.01)	(2.02)	(0.57)	(1.64)
Vocational education and training	-0.040*	0.058	-0.275*	-0.029**	-6.968**	0.674	6.294**
	(0.02)	(0.11)	(0.16)	(0.01)	(3.55)	(1.01)	(2.88)
Higher education	-0.028	-0.049	-0.211	-0.016	1.673	0.127	-1.800
	(0.04)	(0.18)	(0.28)	(0.02)	(5.99)	(1.71)	(4.86)
Log of household size	-0.017***	-0.041	-0.110***	-0.010***	-0.995	-0.033	1.028
	(0.01)	(0.03)	(0.04)	(0.00)	(0.89)	(0.25)	(0.72)
Plot characteristics							
Type of crop stand (1 – pure stand, 0 – mixed stand)	-0.009	0.022	-0.062	-0.006	-3.335***	0.818***	2.517***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.05)	(0.30)	(0.85)
Fertilizer (organic) is used on the plot	-0.009	0.025	-0.061	-0.008	-0.711	-0.065	0.776
	(0.01)	(0.05)	(0.08)	(0.01)	(1.81)	(0.52)	(1.47)
Fertilizer (inorganic) is used on plot	0.067***	-0.182***	0.539***	0.052***	10.421***	-1.822***	-8.600***
	(0.01)	(0.04)	(0.07)	(0.01)	(1.49)	(0.42)	(1.20)
Pesticides are used in plot	-0.008	-0.074	-0.003	-0.001	2.846	-1.108	-1.738
	(0.02)	(0.08)	(0.12)	(0.01)	(2.60)	(0.74)	(2.11)
Plot was cultivated in the previous season	-0.026***	-0.062	-0.187***	-0.015***	-1.303	-0.187	1.490
	(0.01)	(0.04)	(0.06)	(0.01)	(1.34)	(0.38)	(1.08)
Log of the plot area	-0.029***	-0.029	-0.216***	-0.020***	-6.389***	1.408***	4.980***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.14)	(0.33)	(0.93)
Problems with erosion	-0.024***	-0.066**	-0.193***	-0.012***	-3.981***	0.956***	3.025***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.10)	(0.31)	(0.89)
Hired labor	0.002	-0.008	0.015	-0.001	0.348	-0.107	-0.240
	(0.01)	(0.03)	(0.05)	(0.00)	(1.13)	(0.32)	(0.92)
_cons	0.594***	6.519***	1.654***	0.135***	60.394***	19.910***	19.696***
	(0.02)	(0.09)	(0.14)	(0.01)	(3.01)	(0.86)	(2.44)
r2_a	0.14	0.03	0.15	0.16	0.13	0.06	0.14
N	877	877	877	877	877	877	877

Table A.8. Imputation Model with iSDA Soil Quality Variables as Dependent Variables, Top Soil, Ethiopia (2013/14)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	-0.000	0.000	0.000	-0.000	-0.001	0.014***	-0.005
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Head`s gender (male==1)	-0.010**	-0.059*	-0.046	-0.004	0.125	-0.641***	0.254
	(0.00)	(0.03)	(0.04)	(0.00)	(0.39)	(0.21)	(0.54)
Head`s education							
Primary school	0.011***	-0.060**	0.106***	0.009***	0.948***	0.934***	-1.925***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Secondary education (+vocational)	0.025***	-0.019	0.164***	0.019***	0.216	1.213***	-1.078
	(0.01)	(0.05)	(0.06)	(0.01)	(0.65)	(0.35)	(0.90)
Higher education	0.026	-0.166	0.214*	0.026**	2.301*	0.603	-3.416*
	(0.02)	(0.11)	(0.13)	(0.01)	(1.31)	(0.70)	(1.81)
Log of household size	0.004	0.090***	-0.015	-0.000	-1.533***	0.134	1.943***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Plot characteristics							
Cropping method	0.002	0.064**	-0.013	0.001	-0.459	0.336*	-0.226
	(0.00)	(0.03)	(0.03)	(0.00)	(0.34)	(0.18)	(0.47)
Crop rotation (yes/no)	-0.019***	0.038	-0.125***	-0.016***	0.336	-0.186	-0.401
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Plot is prevented from erosion	0.003	-0.192***	0.099***	0.008***	2.373***	-0.231	-3.485***
	(0.00)	(0.02)	(0.03)	(0.00)	(0.28)	(0.15)	(0.39)
Plot is tilled	0.022***	0.125**	0.054	0.014**	0.467	1.893***	-2.536***
	(0.01)	(0.05)	(0.06)	(0.01)	(0.64)	(0.34)	(0.89)
Fertilizer (inorganic) is used on plot	0.023***	-0.014	0.148***	0.016***	0.963***	2.021***	-2.126***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.36)	(0.19)	(0.50)
Fertilizer (organic) is used on plot	0.003	0.077***	-0.027	0.001	-2.141***	0.166	1.945***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.35)	(0.18)	(0.48)
Pesticides are used on plot	0.006	-0.051	0.046	0.007	0.103	0.361	0.563
	(0.01)	(0.04)	(0.05)	(0.01)	(0.54)	(0.29)	(0.75)
Hired labor	0.005	0.010	0.021	0.002	0.922	0.608	-1.052
	(0.01)	(0.06)	(0.07)	(0.01)	(0.73)	(0.39)	(1.00)
Log of plot area	-0.013***	0.103***	-0.120***	-0.012***	-1.023***	-0.436***	1.758***
	(0.00)	(0.01)	(0.01)	(0.00)	(0.09)	(0.05)	(0.13)
_cons	0.692***	5.025***	2.603***	0.256***	42.200***	23.611***	30.220***
	(0.01)	(0.08)	(0.10)	(0.01)	(1.04)	(0.55)	(1.44)
r2_a	0.11	0.17	0.14	0.14	0.14	0.15	0.18
N	1529	1529	1529	1529	1529	1529	1529

Table A.9. Imputation Model with Additional iSDA Soil Quality Variables, Top Soil, Ethiopia (2013/14)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	-0.000	-0.001	-0.002	-0.000	0.015	-0.016	-0.004
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.01)
Head`s gender (male==1)	-0.008*	0.021	-0.147**	-0.008	2.613***	-1.928***	-0.561
	(0.00)	(0.05)	(0.07)	(0.01)	(0.83)	(0.52)	(0.45)
Head`s education							
Primary school	-0.000	-0.051	-0.010	0.001	-1.440**	-0.187	0.893**
	(0.00)	(0.04)	(0.05)	(0.00)	(0.68)	(0.43)	(0.37)
Secondary education (+vocational)	0.010	-0.007	0.151	0.010	0.573	-0.197	-0.862
	(0.01)	(0.08)	(0.11)	(0.01)	(1.39)	(0.87)	(0.75)
Higher education	-0.036**	-0.265	-0.350	-0.049**	1.184	-1.351	-1.291
	(0.02)	(0.16)	(0.22)	(0.02)	(2.80)	(1.75)	(1.52)
Log of household size	0.010***	-0.022	0.179***	0.011**	-1.270*	1.773***	0.217
	(0.00)	(0.04)	(0.05)	(0.01)	(0.69)	(0.43)	(0.37)
Plot characteristics							
Cropping method	-0.016***	-0.220***	-0.029	-0.013**	1.333*	0.070	-0.924**
	(0.00)	(0.04)	(0.06)	(0.01)	(0.73)	(0.46)	(0.40)
Crop rotation (yes/no)	-0.017***	-0.079**	-0.235***	-0.023***	1.837***	-1.543***	-0.344
	(0.00)	(0.04)	(0.05)	(0.01)	(0.68)	(0.42)	(0.37)
Plot is prevented from erosion	-0.005	-0.055	-0.005	0.002	-1.032*	-1.318***	1.555***
	(0.00)	(0.04)	(0.05)	(0.00)	(0.62)	(0.38)	(0.34)
Plot is tilled	-0.007	-0.201**	0.115	-0.002	-5.634***	2.983***	2.156***
	(0.01)	(0.08)	(0.11)	(0.01)	(1.37)	(0.86)	(0.74)
Fertilizer (inorganic) is used on plot	-0.005	-0.245***	0.121**	0.002	-5.195***	3.200***	0.479
	(0.00)	(0.04)	(0.06)	(0.01)	(0.77)	(0.50)	(0.42)
Fertilizer (organic) is used on plot	-0.000	-0.067	0.007	0.002	0.783	0.412	0.318
	(0.00)	(0.04)	(0.06)	(0.01)	(0.75)	(0.46)	(0.40)
Pesticides are used on plot	-0.007	-0.027	-0.061	-0.017**	-0.087	0.001	-0.717
	(0.01)	(0.07)	(0.09)	(0.01)	(1.16)	(0.73)	(0.63)
Hired labor	0.010	0.109	0.044	0.004	-0.579	-0.151	-0.241
	(0.01)	(0.09)	(0.12)	(0.01)	(1.55)	(0.97)	(0.84)
Log of plot area	-0.004***	-0.012	-0.075***	-0.005***	0.807***	0.152	-0.354***
	(0.00)	(0.01)	(0.02)	(0.00)	(0.20)	(0.13)	(0.11)
iSDA soil quality variables							
Weighted soil index	0.395***						
	(0.02)						
pH		0.848***					
		(0.04)					
Organic carbon			1.219***				
			(0.05)				
Total nitrogen				1.119***			
				(0.04)			
Clay					1.313***		
					(0.05)		
Silt						0.769***	
						(0.06)	
Sand							0.607***
							(0.02)
_cons	0.327***	1.930***	1.335***	0.108***	17.722***	1.118	-10.718***
	(0.02)	(0.24)	(0.21)	(0.02)	(3.21)	(2.06)	(1.37)
r2_a	0.22	0.29	0.40	0.41	0.31	0.18	0.37
N	1529	1529	1529	1529	1529	1529	1529

Table A.10. Imputation Model with SoilGrids Soil Quality Variables as Dependent Variables, Top Soil, Ethiopia (2013/14)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	-0.000	0.000	0.000	-0.000	-0.001	0.014***	-0.005
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Head`s gender (male==1)	-0.010**	-0.059*	-0.046	-0.004	0.125	-0.641***	0.254
	(0.00)	(0.03)	(0.04)	(0.00)	(0.39)	(0.21)	(0.54)
Head`s education							
Primary school	0.011***	-0.060**	0.106***	0.009***	0.948***	0.934***	-1.925***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Secondary education (+vocational)	0.025***	-0.019	0.164***	0.019***	0.216	1.213***	-1.078
	(0.01)	(0.05)	(0.06)	(0.01)	(0.65)	(0.35)	(0.90)
Higher	0.026	-0.166	0.214*	0.026**	2.301*	0.603	-3.416*
	(0.02)	(0.11)	(0.13)	(0.01)	(1.31)	(0.70)	(1.81)
Log of household size	0.004	0.090***	-0.015	-0.000	-1.533***	0.134	1.943***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Plot characteristics							
Cropping method	0.002	0.064**	-0.013	0.001	-0.459	0.336*	-0.226
	(0.00)	(0.03)	(0.03)	(0.00)	(0.34)	(0.18)	(0.47)
Crop rotation (yes/no)	-0.019***	0.038	-0.125***	-0.016***	0.336	-0.186	-0.401
	(0.00)	(0.03)	(0.03)	(0.00)	(0.32)	(0.17)	(0.44)
Plot is prevented from erosion	0.003	-0.192***	0.099***	0.008***	2.373***	-0.231	-3.485***
	(0.00)	(0.02)	(0.03)	(0.00)	(0.28)	(0.15)	(0.39)
Plot is tilled	0.022***	0.125**	0.054	0.014**	0.467	1.893***	-2.536***
	(0.01)	(0.05)	(0.06)	(0.01)	(0.64)	(0.34)	(0.89)
Fertilizer (inorganic) is used on plot	0.023***	-0.014	0.148***	0.016***	0.963***	2.021***	-2.126***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.36)	(0.19)	(0.50)
Fertilizer (organic) is used on plot	0.003	0.077***	-0.027	0.001	-2.141***	0.166	1.945***
	(0.00)	(0.03)	(0.03)	(0.00)	(0.35)	(0.18)	(0.48)
Pesticides are used on plot	0.006	-0.051	0.046	0.007	0.103	0.361	0.563
	(0.01)	(0.04)	(0.05)	(0.01)	(0.54)	(0.29)	(0.75)
Hired labor	0.005	0.010	0.021	0.002	0.922	0.608	-1.052
	(0.01)	(0.06)	(0.07)	(0.01)	(0.73)	(0.39)	(1.00)
Log of plot area	-0.013***	0.103***	-0.120***	-0.012***	-1.023***	-0.436***	1.758***
	(0.00)	(0.01)	(0.01)	(0.00)	(0.09)	(0.05)	(0.13)
_cons	0.692***	5.025***	2.603***	0.256***	42.200***	23.611***	30.220***
	(0.01)	(0.08)	(0.10)	(0.01)	(1.04)	(0.55)	(1.44)
r2_a	0.11	0.17	0.14	0.14	0.14	0.15	0.18
N	1529	1529	1529	1529	1529	1529	1529

Table A.11. Imputation Model with Additional SoilGrids Soil Quality Variables, Top Soil, Ethiopia (2013/14)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	-0.000	-0.001	-0.002	-0.000	-0.006	-0.006	0.004
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.01)
Head`s gender (male==1)	-0.008*	0.016	-0.186***	-0.005	2.950***	-2.282***	-0.733
	(0.00)	(0.05)	(0.07)	(0.01)	(0.93)	(0.54)	(0.51)
Head`s education							
Primary school	-0.000	-0.058	0.003	-0.002	-1.373*	0.390	0.542
	(0.00)	(0.04)	(0.05)	(0.01)	(0.77)	(0.44)	(0.42)
Secondary education (+vocational)	0.012	-0.034	0.215*	0.014	0.448	0.497	-0.903
	(0.01)	(0.08)	(0.11)	(0.01)	(1.56)	(0.90)	(0.86)
Higher	-0.031**	-0.350**	-0.178	-0.039*	4.035	-1.112	-2.893*
	(0.02)	(0.17)	(0.22)	(0.02)	(3.14)	(1.82)	(1.73)
Log of household size	0.010***	-0.018	0.193***	0.010*	-2.265***	2.100***	0.512
	(0.00)	(0.04)	(0.05)	(0.01)	(0.77)	(0.45)	(0.43)
Plot characteristics							
Cropping method	-0.014***	-0.193***	-0.008	-0.009	1.046	0.069	-0.763*
	(0.00)	(0.04)	(0.06)	(0.01)	(0.82)	(0.48)	(0.45)
Crop rotation (yes/no)	-0.016***	-0.025	-0.235***	-0.023***	1.697**	-1.752***	-0.189
	(0.00)	(0.04)	(0.05)	(0.01)	(0.76)	(0.44)	(0.42)
Plot is prevented from erosion	-0.007**	-0.060	-0.073	-0.003	0.061	-1.781***	0.920**
	(0.00)	(0.04)	(0.05)	(0.00)	(0.70)	(0.40)	(0.39)
Plot is tilled	-0.013*	-0.225***	-0.026	-0.013	-7.123***	3.553***	3.183***
	(0.01)	(0.08)	(0.11)	(0.01)	(1.54)	(0.90)	(0.86)
Fertilizer (inorganic) is used on plot	-0.008*	-0.347***	0.116*	-0.003	-5.807***	4.062***	1.306***
	(0.00)	(0.05)	(0.06)	(0.01)	(0.88)	(0.52)	(0.49)
Fertilizer (organic) is used on plot	0.002	-0.041	0.035	0.007	-1.363	0.695	0.905**
	(0.00)	(0.04)	(0.06)	(0.01)	(0.83)	(0.48)	(0.46)
Pesticides are used on plot	-0.010	-0.106	-0.051	-0.019**	-0.133	0.215	-0.176
	(0.01)	(0.07)	(0.09)	(0.01)	(1.30)	(0.75)	(0.72)
Hired labor	0.009	0.143	0.017	-0.001	0.100	0.276	-0.553
	(0.01)	(0.09)	(0.12)	(0.01)	(1.74)	(1.01)	(0.96)
Log of plot area	-0.005***	-0.001	-0.075***	-0.007***	0.185	0.020	0.008
	(0.00)	(0.01)	(0.02)	(0.00)	(0.23)	(0.13)	(0.13)
SoilGrids soil quality variables							
Weighted soil index	0.357***						
	(0.02)						
pH		0.712***					
		(0.04)					
Organic carbon			0.742***				
			(0.03)				
Total nitrogen				0.801***			
				(0.03)			
Clay					0.646***		
					(0.05)		
Silt						0.181***	
						(0.03)	
Sand							0.315***
							(0.02)
_cons	0.345***	2.511***	1.206***	0.123***	48.005***	13.360***	-1.307
	(0.02)	(0.23)	(0.22)	(0.02)	(3.25)	(1.83)	(1.48)
r2_a	0.22	0.27	0.39	0.39	0.13	0.12	0.18
N	1529	1529	1529	1529	1529	1529	1529

Table A.12. Imputation Model with iSDA Soil Quality Variables as Dependent Variables, Top Soil, Uganda (2015/16)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.000**	0.001	0.001*	0.000**	-0.014	-0.006	0.020
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)
Head`s gender (male==1)	0.016***	0.005	0.080***	0.009***	-0.049	0.405	-0.089
	(0.01)	(0.02)	(0.03)	(0.00)	(0.44)	(0.27)	(0.57)
Head`s education							
Primary education	-0.004	0.026	-0.029	-0.003	-1.250**	-0.881**	1.710**
	(0.01)	(0.02)	(0.04)	(0.00)	(0.58)	(0.36)	(0.76)
Secondary education	-0.009	0.022	-0.044	-0.007	-0.735	-0.709*	1.417*
	(0.01)	(0.02)	(0.04)	(0.00)	(0.65)	(0.40)	(0.85)
Vocational education and training	-0.027**	0.111***	-0.165**	-0.021**	-3.126***	-2.401***	6.604***
	(0.01)	(0.04)	(0.08)	(0.01)	(1.15)	(0.70)	(1.50)
Higher education	-0.034	0.033	-0.177	-0.022	-3.067	-1.574	4.060
	(0.02)	(0.07)	(0.13)	(0.01)	(1.93)	(1.19)	(2.53)
Log of household size	-0.011***	-0.006	-0.046**	-0.008***	0.167	-0.179	-0.000
	(0.00)	(0.01)	(0.02)	(0.00)	(0.29)	(0.18)	(0.38)
Plot characteristics							
Cropping method	-0.016***	0.036***	-0.090***	-0.011***	-0.392	-0.601***	1.216***
	(0.00)	(0.01)	(0.02)	(0.00)	(0.34)	(0.21)	(0.44)
Fertilizer (organic) is used on plot	0.004	0.034	-0.005	0.003	-1.449**	-0.604*	0.746
	(0.01)	(0.02)	(0.04)	(0.00)	(0.59)	(0.36)	(0.77)
Fertilizer (inorganic) is used on plot	0.077***	-0.126***	0.403***	0.054***	3.125***	3.781***	-5.730***
	(0.01)	(0.02)	(0.03)	(0.00)	(0.48)	(0.29)	(0.63)
Pesticides are used on plot	-0.007	0.004	-0.054	-0.003	0.292	0.534	-0.357
	(0.01)	(0.03)	(0.06)	(0.01)	(0.84)	(0.52)	(1.10)
Plot was cultivated in previous season	-0.019***	0.009	-0.092***	-0.012***	1.338***	0.107	-0.653
	(0.01)	(0.02)	(0.03)	(0.00)	(0.43)	(0.27)	(0.56)
Log size of the parcel in acres	-0.021***	0.021	-0.107***	-0.014***	-1.490***	-1.101***	2.236***
	(0.00)	(0.01)	(0.02)	(0.00)	(0.37)	(0.23)	(0.48)
Problems with erosion	0.006	-0.049***	0.057**	0.004	1.502***	1.025***	-1.713***
	(0.00)	(0.01)	(0.02)	(0.00)	(0.35)	(0.22)	(0.46)
Hired labor	-0.001	0.032**	-0.009	-0.003	-0.496	-0.593***	0.746
	(0.00)	(0.01)	(0.02)	(0.00)	(0.37)	(0.22)	(0.48)
_cons	0.663***	5.869***	1.586***	0.150***	32.834***	24.092***	42.508***
	(0.01)	(0.04)	(0.06)	(0.01)	(0.97)	(0.60)	(1.27)
r2_a	0.28	0.09	0.25	0.32	0.09	0.24	0.15
N	875	875	875	875	875	875	875

Table A.13. Imputation Model with Additional iSDA Soil Quality Variables, Top Soil, Uganda (2015/16)

	Weighted soil index	pH	Acidified Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.000**	0.002	0.003**	0.000	0.097***	-0.012	-0.083***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)	(0.01)	(0.03)
Head`s gender (male==1)	0.003	-0.045	0.053	0.002	2.150*	0.057	-1.928*
	(0.01)	(0.04)	(0.05)	(0.00)	(1.26)	(0.37)	(1.04)
Head`s education							
Primary education	0.009	0.033	0.086	0.003	0.969	-0.143	-0.784
	(0.01)	(0.05)	(0.07)	(0.01)	(1.68)	(0.49)	(1.39)
Secondary education	0.003	0.017	0.023	0.001	-0.278	-0.333	0.311
	(0.01)	(0.06)	(0.08)	(0.01)	(1.88)	(0.55)	(1.56)
Vocational education and training	-0.014	-0.005	-0.097	-0.006	-3.411	-0.336	2.391
	(0.02)	(0.10)	(0.14)	(0.01)	(3.32)	(0.97)	(2.77)
Higher education	0.004	-0.067	-0.020	0.009	5.170	-0.538	-4.205
	(0.03)	(0.17)	(0.24)	(0.02)	(5.59)	(1.64)	(4.64)
Log of household size	-0.006	-0.036	-0.059*	-0.002	-1.175	-0.108	1.019
	(0.00)	(0.03)	(0.04)	(0.00)	(0.83)	(0.24)	(0.69)
Plot characteristics							
Cropping pattern	0.006	0.002	0.035	0.006**	-2.861***	0.558*	1.776**
	(0.01)	(0.03)	(0.04)	(0.00)	(0.98)	(0.29)	(0.82)
Fertilizer (organic) is used on plot	-0.013	0.005	-0.056	-0.011*	0.953	-0.327	0.328
	(0.01)	(0.05)	(0.07)	(0.01)	(1.70)	(0.50)	(1.40)
Fertilizer (inorganic) is used on plot	-0.006	-0.110**	0.105	-0.007	6.876***	-0.232	-5.219***
	(0.01)	(0.04)	(0.06)	(0.01)	(1.42)	(0.44)	(1.20)
Pesticides are used on plot	-0.001	-0.077	0.055	0.003	2.513	-0.884	-1.524
	(0.01)	(0.08)	(0.10)	(0.01)	(2.42)	(0.71)	(2.01)
Plot was cultivated in previous season	-0.008	-0.066*	-0.087	-0.002	-2.849**	-0.131	1.888*
	(0.01)	(0.04)	(0.05)	(0.00)	(1.25)	(0.36)	(1.03)
Log size of the parcel in acres	-0.010	-0.042	-0.101**	-0.005	-4.667***	0.931***	3.646***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.08)	(0.32)	(0.89)
Problems with erosion	-0.030***	-0.039	-0.255***	-0.016***	-5.674***	1.379***	4.029***
	(0.01)	(0.03)	(0.04)	(0.00)	(1.04)	(0.30)	(0.86)
Hired labor	0.003	-0.027	0.024	0.001	0.918	-0.364	-0.681
	(0.01)	(0.03)	(0.05)	(0.00)	(1.06)	(0.31)	(0.88)
iSDA soil quality variables							
Weighted soil index	0.952***						
	(0.05)						
pH		0.573***					
		(0.08)					
Organic carbon			1.076***				
			(0.06)				
Total nitrogen				1.095***			
				(0.04)			
Clay					1.138***		
					(0.10)		
Silt						-0.422***	
						(0.05)	
Sand							0.591***
							(0.06)
_cons	-0.038	3.153***	-0.054	-0.029***	22.959***	30.079***	-5.386
	(0.03)	(0.50)	(0.16)	(0.01)	(4.28)	(1.40)	(3.53)
r2_a	0.42	0.08	0.36	0.53	0.25	0.14	0.22
N	875	875	875	875	875	875	875

Table A.14. Imputation Model with SoilGrids Soil Quality Variables as Dependent Variables, Top Soil, Uganda (2015/16)

	Weighted soil index	pH	Organic Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.000*	0.001*	0.003	0.000	0.006	-0.003	-0.005
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Head`s gender (male==1)	0.015***	-0.008	0.195***	0.011**	0.203	0.029	-0.568***
	(0.01)	(0.02)	(0.07)	(0.00)	(0.16)	(0.15)	(0.19)
Head`s education							
Primary education	-0.003	0.027	-0.103	-0.004	0.202	-0.212	-0.205
	(0.01)	(0.02)	(0.09)	(0.01)	(0.22)	(0.20)	(0.25)
Secondary education	-0.014*	-0.006	-0.231**	-0.010	-0.213	-0.607***	0.479*
	(0.01)	(0.03)	(0.10)	(0.01)	(0.24)	(0.23)	(0.28)
Vocational education and training	-0.018	0.052	-0.373**	-0.015	-0.363	-0.890**	0.973*
	(0.01)	(0.05)	(0.18)	(0.01)	(0.43)	(0.40)	(0.50)
Higher education	-0.033	0.033	-0.584*	-0.023	0.965	-1.584**	0.458
	(0.02)	(0.08)	(0.31)	(0.02)	(0.72)	(0.67)	(0.84)
Log of household size	-0.010***	-0.018	-0.123***	-0.007**	-0.324***	-0.240**	0.572***
	(0.00)	(0.01)	(0.05)	(0.00)	(0.11)	(0.10)	(0.13)
Plot characteristics							
Cropping pattern	-0.012***	0.019	-0.151***	-0.010***	0.180	-0.442***	0.618***
	(0.00)	(0.01)	(0.05)	(0.00)	(0.13)	(0.12)	(0.15)
Fertilizer (organic) is used on plot	-0.005	0.030	-0.067	-0.005	0.247	-0.212	0.443*
	(0.01)	(0.02)	(0.09)	(0.01)	(0.22)	(0.20)	(0.26)
Fertilizer (inorganic) is used on plot	0.056***	-0.087***	0.757***	0.053***	0.460***	1.196***	-1.853***
	(0.01)	(0.02)	(0.08)	(0.00)	(0.18)	(0.17)	(0.21)
Pesticides is used on plot	-0.007	-0.012	-0.027	-0.009	0.187	-0.025	-0.221
	(0.01)	(0.03)	(0.13)	(0.01)	(0.31)	(0.29)	(0.37)
Plot was cultivated in previous season	-0.018***	-0.071***	-0.175**	-0.012***	-0.422***	-0.095	0.377**
	(0.01)	(0.02)	(0.07)	(0.00)	(0.16)	(0.15)	(0.19)
Log size of the parcel in acres	-0.007	0.051***	-0.162***	-0.007*	-0.086	0.007	-0.047
	(0.00)	(0.02)	(0.06)	(0.00)	(0.14)	(0.13)	(0.16)
Problems with erosion	0.005	-0.072***	0.123**	0.008**	-0.316**	0.111	0.237
	(0.00)	(0.01)	(0.06)	(0.00)	(0.13)	(0.12)	(0.15)
Hired labor	-0.003	0.035**	-0.054	-0.005	-0.111	0.125	-0.028
	(0.00)	(0.02)	(0.06)	(0.00)	(0.14)	(0.13)	(0.16)
_cons	0.665***	5.834***	3.366***	0.225***	38.055***	23.954***	38.680***
	(0.01)	(0.04)	(0.16)	(0.01)	(0.36)	(0.34)	(0.42)
r2_a	0.16	0.07	0.17	0.17	0.04	0.09	0.16
N	875	875	875	875	875	875	875

Table A.15. Imputation Model with Additional SoilGrids Soil Quality Variables, Top Soil, Uganda (2015/16)

	Weighted soil index	pH	Acidified Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.001**	0.001	0.004**	0.000**	0.077**	-0.009	-0.063**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.01)	(0.03)
Head`s gender (male==1)	0.006	-0.037	0.063	0.005	1.968	-0.120	-1.090
	(0.01)	(0.04)	(0.06)	(0.00)	(1.35)	(0.38)	(1.06)
Head`s education							
Primary education	0.008	0.032	0.095	0.002	-0.580	0.274	0.549
	(0.01)	(0.05)	(0.08)	(0.01)	(1.80)	(0.51)	(1.41)
Secondary education	0.006	0.033	0.065	-0.000	-0.982	0.098	0.397
	(0.01)	(0.06)	(0.08)	(0.01)	(2.02)	(0.58)	(1.58)
Vocational education and training	-0.025	0.027	-0.131	-0.019*	-6.745*	0.870	4.768*
	(0.02)	(0.10)	(0.15)	(0.01)	(3.55)	(1.01)	(2.78)
Higher education	-0.001	-0.069	0.015	0.001	1.080	0.472	-2.523
	(0.03)	(0.17)	(0.25)	(0.02)	(5.99)	(1.71)	(4.69)
Log of household size	-0.009*	-0.029	-0.062*	-0.005*	-0.785	0.019	0.121
	(0.00)	(0.03)	(0.04)	(0.00)	(0.90)	(0.25)	(0.71)
Plot characteristics							
Cropping pattern	0.000	0.011	-0.004	0.001	-3.419***	0.908***	1.525*
	(0.01)	(0.03)	(0.04)	(0.00)	(1.05)	(0.30)	(0.83)
Fertilizer (organic) is used on plot	-0.004	0.006	-0.036	-0.004	-0.850	-0.026	0.073
	(0.01)	(0.05)	(0.08)	(0.01)	(1.81)	(0.52)	(1.42)
Fertilizer (inorganic) is used on plot	0.022**	-0.130***	0.246***	0.014***	10.147***	-2.088***	-5.699***
	(0.01)	(0.04)	(0.07)	(0.01)	(1.49)	(0.43)	(1.21)
Pesticides is used on plot	-0.003	-0.067	0.008	0.005	2.729	-1.104	-1.389
	(0.01)	(0.07)	(0.11)	(0.01)	(2.60)	(0.74)	(2.03)
Plot was cultivated in previous season	-0.011	-0.018	-0.119**	-0.006	-1.064	-0.155	0.910
	(0.01)	(0.04)	(0.06)	(0.00)	(1.34)	(0.38)	(1.05)
Log size of the parcel in acres	-0.024***	-0.061*	-0.154***	-0.015***	-6.310***	1.394***	5.042***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.14)	(0.32)	(0.89)
Problems with erosion	-0.028***	-0.024	-0.241***	-0.017***	-3.767***	0.922***	2.645***
	(0.01)	(0.03)	(0.05)	(0.00)	(1.10)	(0.31)	(0.86)
Hired labor	0.004	-0.030	0.035	0.002	0.423	-0.141	-0.196
	(0.01)	(0.03)	(0.05)	(0.00)	(1.13)	(0.32)	(0.89)
SoilGrids soil quality variables							
Weighted soil index	0.822***						
	(0.05)						
pH		0.604***					
		(0.07)					
Organic carbon			0.387***				
			(0.03)				
Total nitrogen				0.713***			
				(0.03)			
Clay					0.621**		
					(0.28)		
Silt						0.218**	
						(0.09)	
Sand							1.569***
							(0.19)
_cons	0.047	2.992***	0.351**	-0.025**	36.707***	14.691***	-40.941***
	(0.04)	(0.44)	(0.16)	(0.01)	(11.23)	(2.23)	(7.70)
r2_a	0.35	0.10	0.31	0.45	0.14	0.06	0.20
N	875	875	875	875	875	875	875

Figure A.1. Imputation-based Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Survey using Linear Regression Method, Top Soil Analysis, Ethiopia

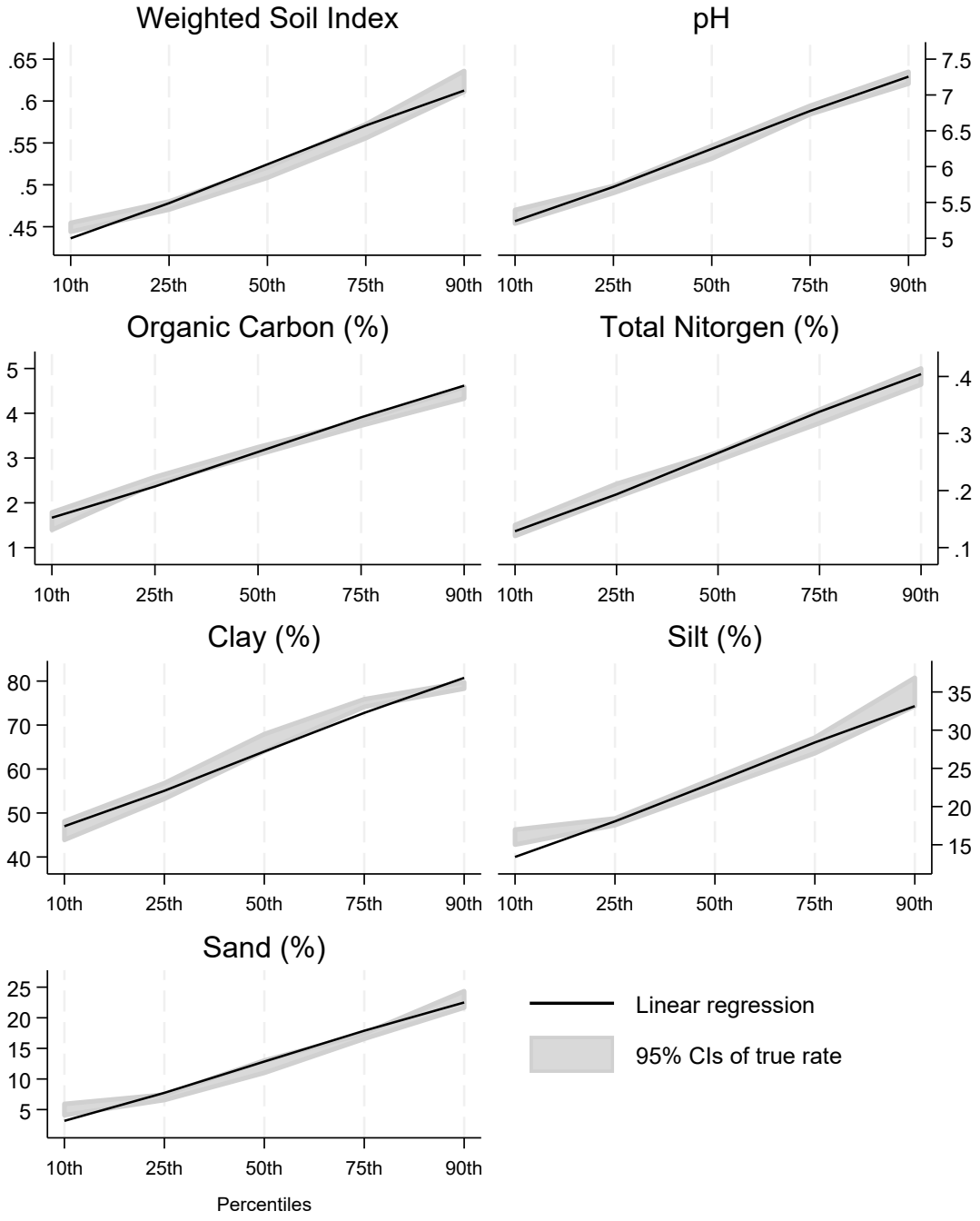


Figure A.2. Imputation-based Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Survey Using Linear Regression Method, Top Soil Analysis, Uganda (2015/16)

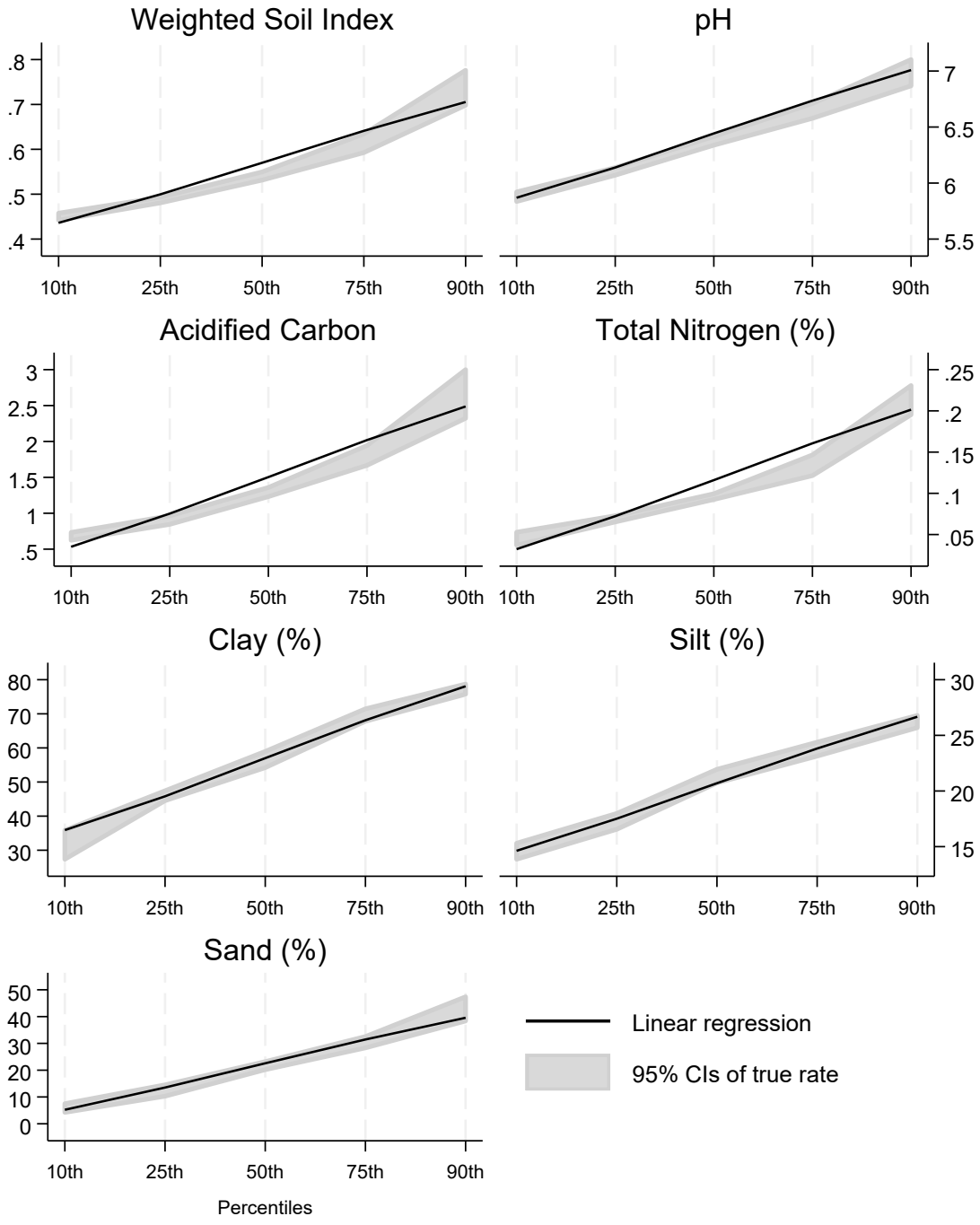
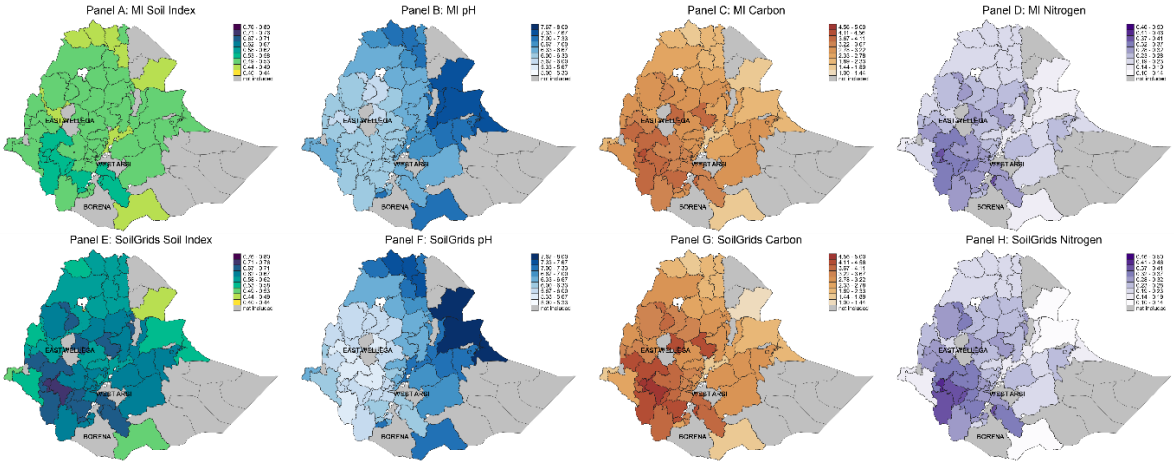
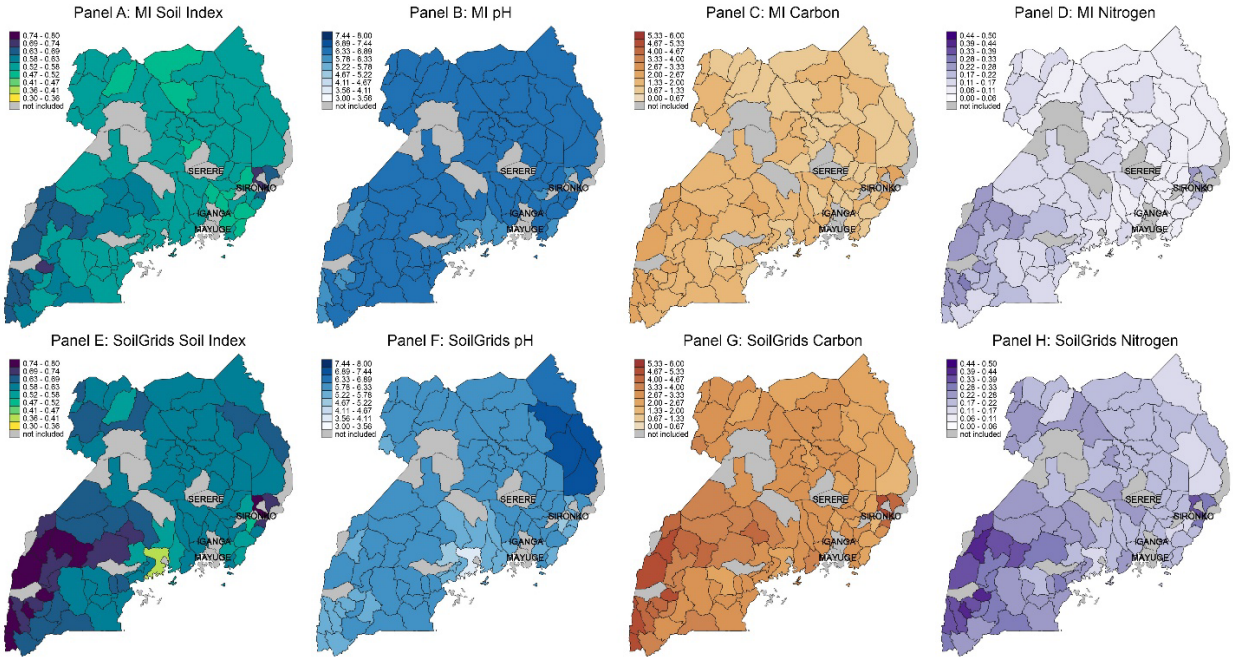


Figure A.3. Imputation-Based Estimates vs. SoilGrids Estimates, Top Soil Analysis, Ethiopia



Notes: Estimates are aggregated by regions (where households were interviewed). Panels A-D show imputed estimates where the prediction model includes SoilGrids geospatial variables. Panels E-H show geospatial SoilGrids estimates. Organic carbon is used in Panel C and Panel G. The coefficient of variation for Panel A is 12.1%, for Panel B is 13.1%, for Panel C is 37.8%, for Panel D is 40.5%, for Panel E is 8.9%, for Panel F is 11.4%, for Panel G is 28.3%, for Panel H is 27.8%.

Figure A.4. Imputation-Based Estimates vs. SoilGrids Estimates, Top Soil Analysis, Uganda



Note: Estimates are aggregated by districts (where households were interviewed). Benchmark survey districts (Iganga, Sironko and Mayge) and districts with missing imputed estimates (Bundibugyo, Nakasongola, Kaberamaido, Kalangala, Kampala, Kapchorwa) are omitted. Panels A-D show imputed estimates where the prediction model includes SoilGrids geospatial variables. Panels E-H show geospatial SoilGrids estimates. Acidified carbon is used in Panel C and Panel G. The coefficient of variation for Panel A is 18.2%, for Panel B is 7.5%, for Panel C is 50.7%, for Panel D is 55.3%, for Panel E is 11.4%. for Panel F is 7.7%, for Panel G is 24%, for Panel H is 25.7%.

Figure A.5. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Survey using Linear Regression Method, Ethiopia

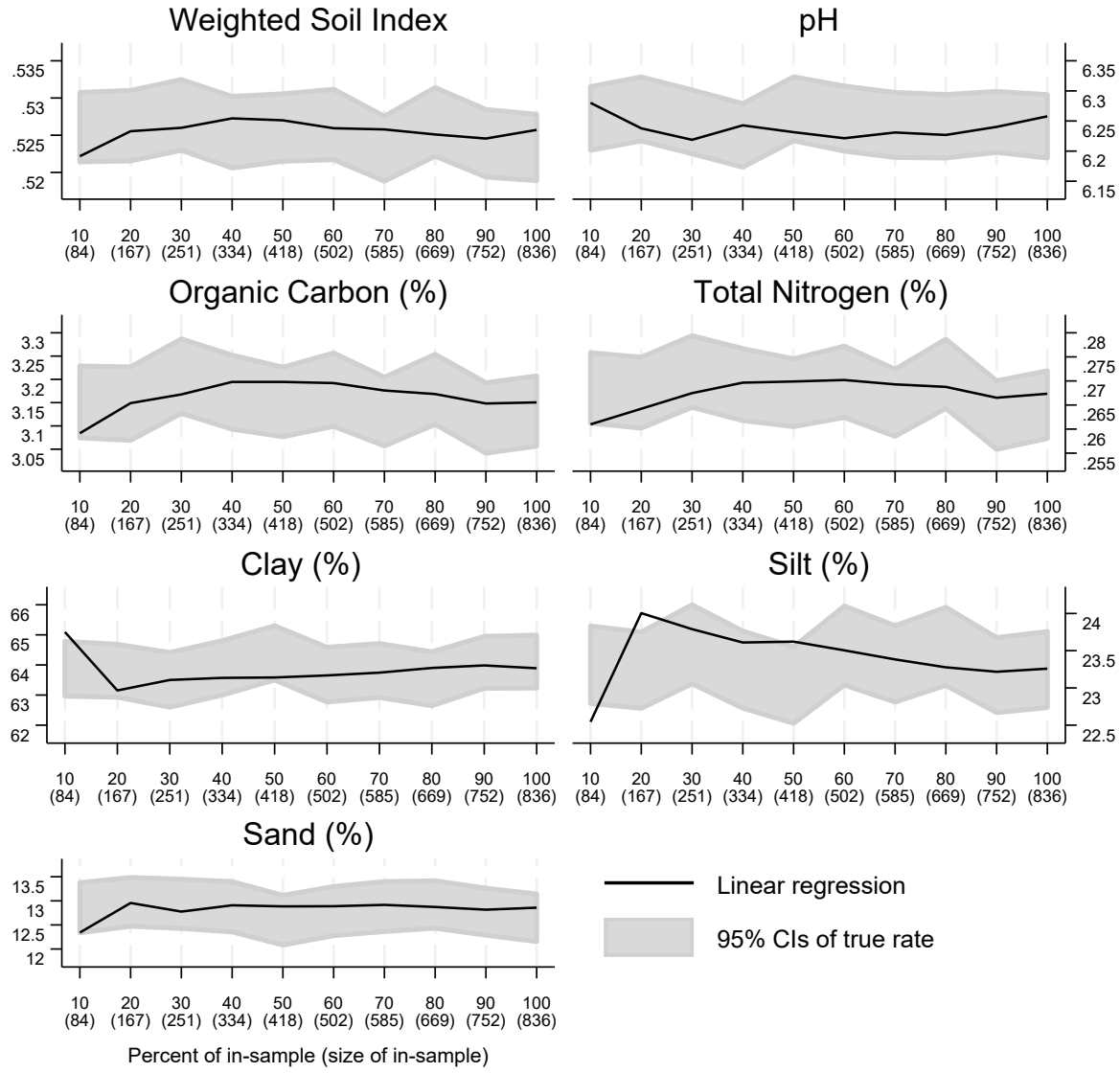


Figure A.6. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Survey using Linear Regression Method, Top Soil Analysis, Uganda (2015/16)

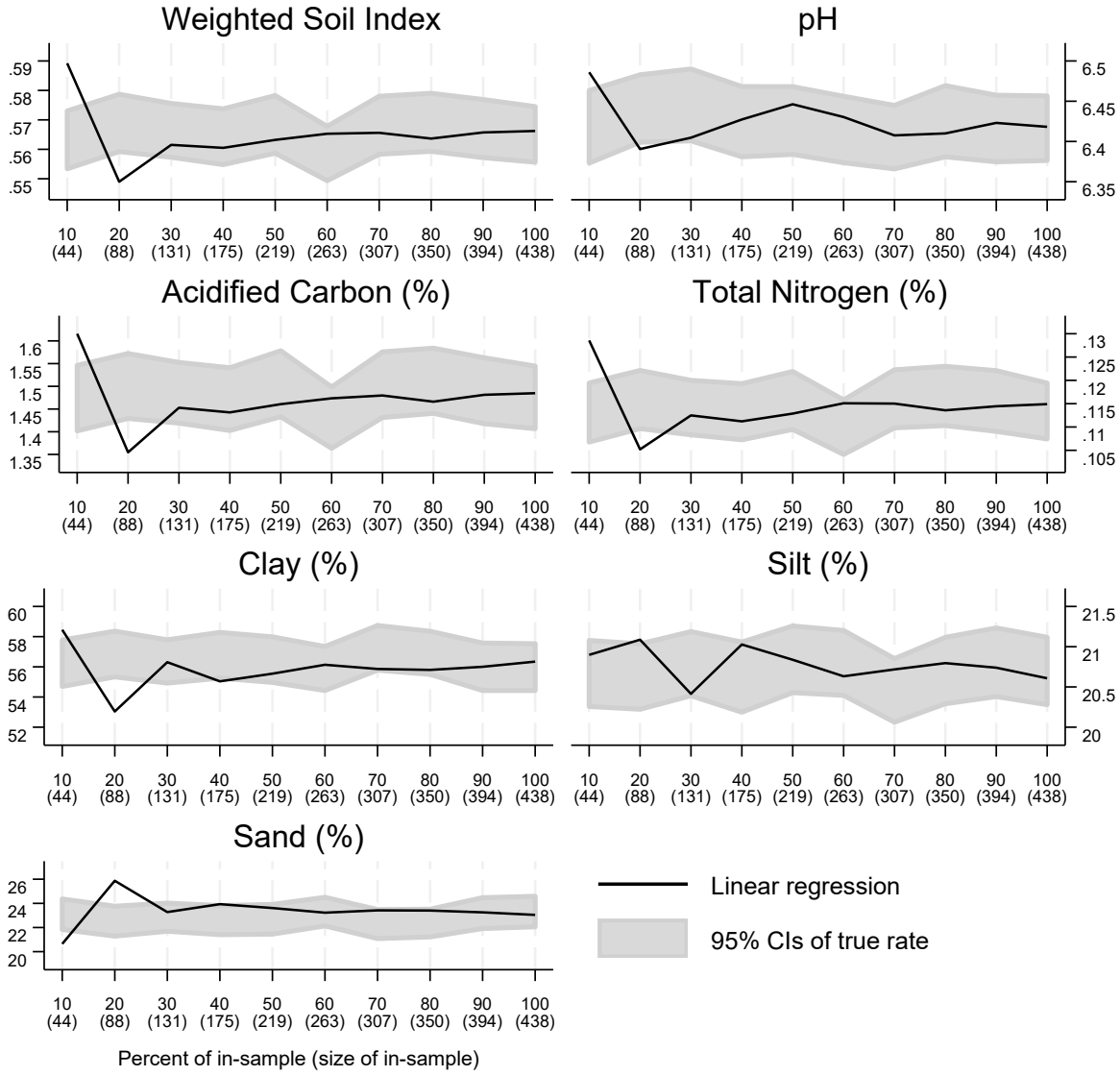
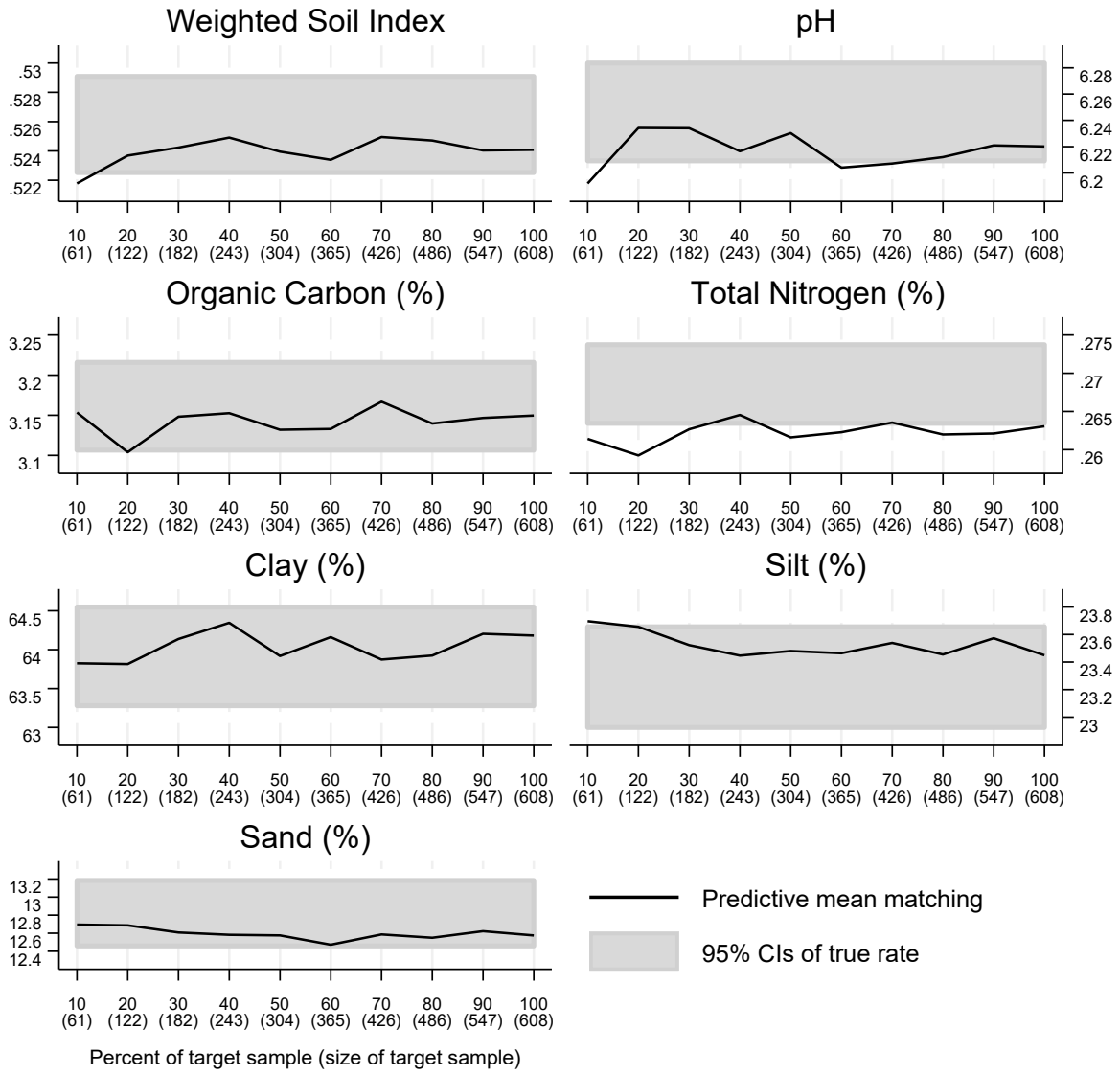
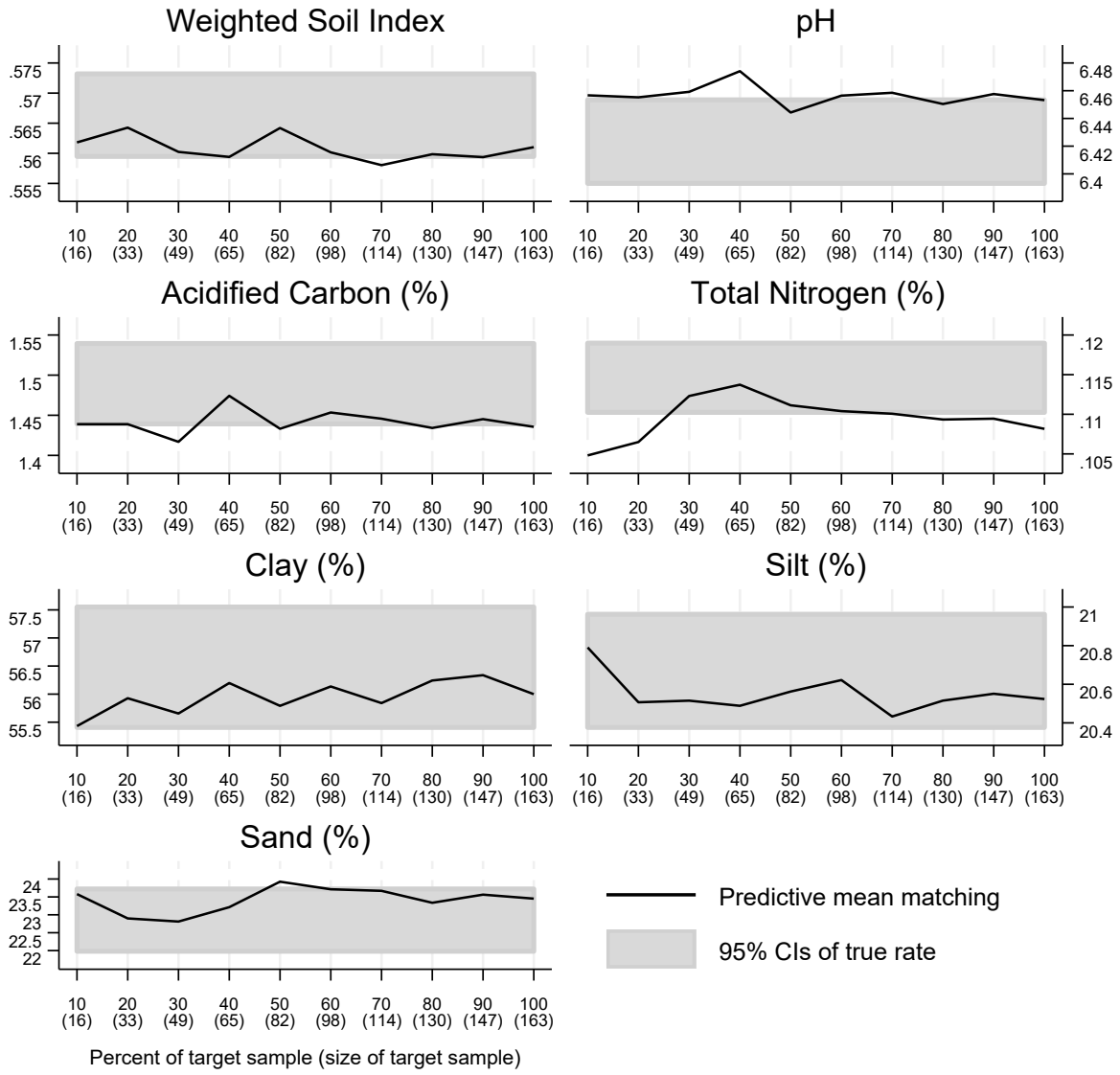


Figure A.7. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Target Sample, Top Soil Analysis, Ethiopia (2013/14)



Note: The target survey (ESS 2) is restricted to the same zones as the benchmark survey (LASER). The target sample is selected as a percentage varying from (randomly selected) 10% to 100% of the target survey (with the number of observations shown in parentheses). Estimates are obtained with 50 iterations using 100% of benchmark data. The total benchmark size is 1672 plots, the target size is 608 plots.

Figure A.8. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Target Sample, Top Soil Analysis, Uganda (2015/16)



Note: The target survey (UNPS 5) is restricted to the same zones as the benchmark survey (MAPS 1). The target sample is selected as a percentage varying from (randomly selected) 10% to 100% of the target survey (with the number of observations shown in parentheses). Estimates are obtained with 50 iterations using 100% of benchmark data. The total benchmark size is 877 plots, the target size is 163 plots.

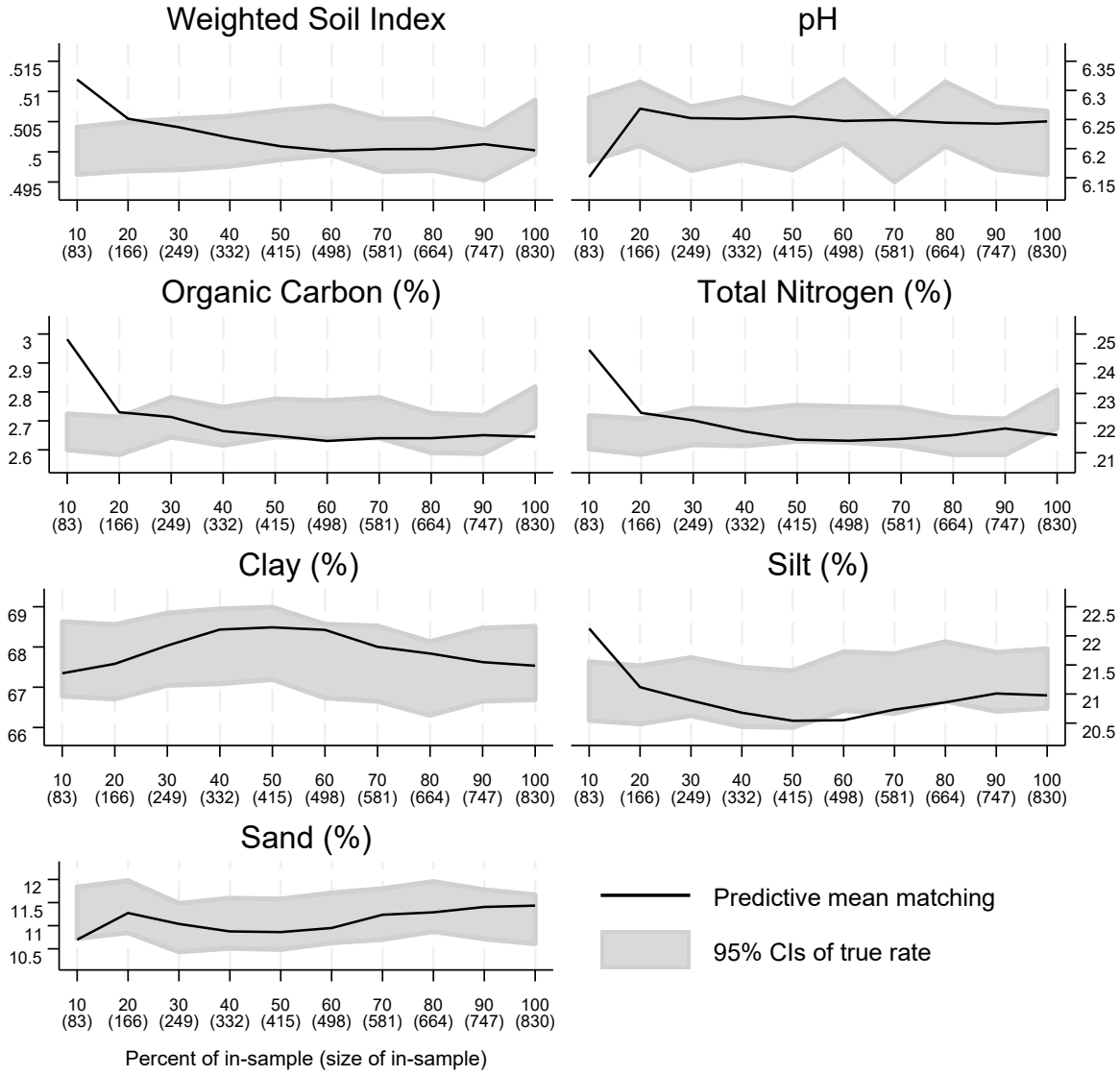
Appendix B. Analysis using sub-soil data

Table B.1. Comparison of Sub Soil Characteristics Across Sources: Plot-Level (LASER 2013/14 and MAPS1 2015/16) vs Geospatial (SoilGrids, 2020 and iSDA, 2021)

	Ethiopia (2013/14)			Uganda (2015/16)		
	LASER	**LASER-iSDAsoil**	**LASER-SoilGrids**	**MAPS1**	**MAPS1-iSDAsoil**	**MAPS1-SoilGrids**
Weighted Soil Index	0.50	-0.11***	-0.18***	0.54	-0.06***	-0.06***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Components of soil index						
pH	6.22	0.33***	0.16***	6.30	0.41***	0.50***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Carbon (%)	2.70	1.48***	0.65***	1.19	0.39***	-0.75***
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Total Nitrogen (%)	0.22	0.09***	0.04***	0.09	0.01***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Soil particle composition						
Clay (%)	67.71	28.66***	27.16***	63.13	27.52***	21.72***
	(0.35)	(0.29)	(0.31)	(0.56)	(0.52)	(0.56)
Silt (%)	21.12	-2.19***	-7.75***	17.11	-5.23***	-4.53***
	(0.19)	(0.18)	(0.21)	(0.17)	(0.23)	(0.18)
Sand (%)	11.19	-25.92***	-19.40***	19.75	-23.14***	-17.30***
	(0.21)	(0.18)	(0.25)	(0.44)	(0.42)	(0.42)
Number of plots	*1,520*			*880*		

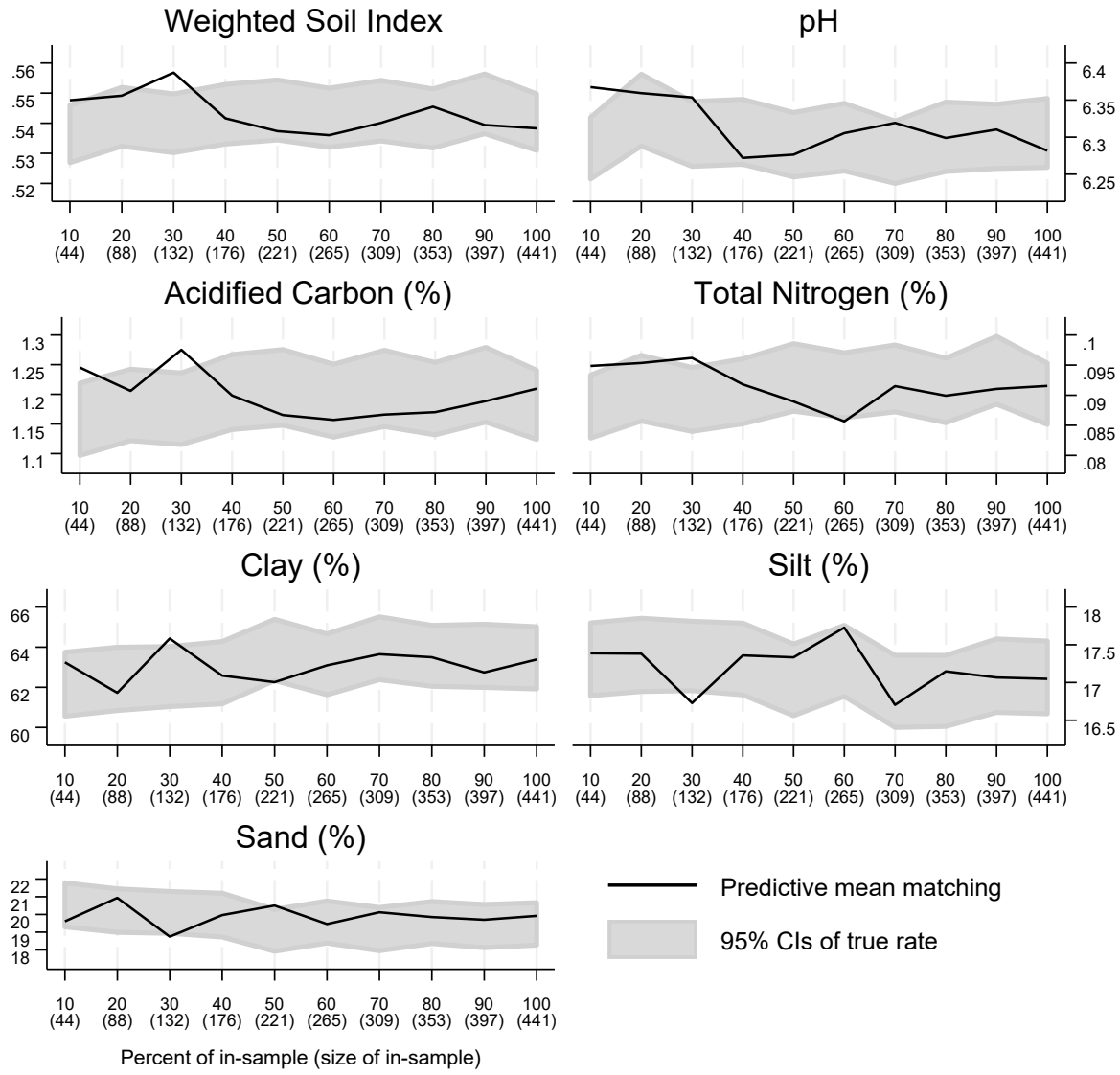
Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. LASER and MAPS 1 subsoil samples are at a soil depth of 20-50cm, SoilGrids subsoil is at a soil depth of 15-30 cm, iSDA subsoil is at soil depth of 20-50cm. The weighted soil index includes pH, carbon, and total nitrogen but excludes electrical conductivity. The sample is restricted to non-missing soil properties in household surveys and geospatial data.

Figure B.2. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Survey, Sub Soil Analysis, Ethiopia (2013/14)



Note: The estimation sample is generated by splitting LASER data into two random samples: 50% in-sample, 50% out-of-sample. Benchmark sample is selected as a percentage varying from (randomly selected) 10% to 100% of the "in-sample" data (with the number of observations shown in parentheses). Estimates are obtained using 50 iterations using 100% of the "out-of-sample" data as the target sample. Total "in-sample" size is 830 plots, the "out-of-sample" size is 830 plots.

Figure B.3. Imputation-based Estimates of Soil Quality Index and its Components for Different Sample Sizes of the Benchmark Survey, Sub Soil Analysis, Uganda (2015/16)



Note: The estimation sample is generated by splitting MAPS data into two random samples: 50% in-sample, 50% out-of-sample. Benchmark sample is selected as a percentage varying from (randomly selected) 10% to 100% of the "in-sample" data (with the number of observations shown in parentheses). Estimates are obtained using 50 iterations using 100% of the "out-of-sample" data as the target sample. Total "in-sample" size is 441 plots, the "out-of-sample" size is 441 plots.

Appendix C. Analysis using older data

Table C.1. Comparison of Top Soil Characteristics Across Sources: Plot-Level (LASER 2013/14 and MAPS1 2015/16) vs Geospatial (AFSIS, 2015)

	Ethiopia (2013/14)		Uganda (2015/16)	
	LASER	**LASER-AFSIS**	**MAPS1**	**MAPS1-AFSIS**
Weighted Soil Index	0.47	-0.01***	0.47	-0.02***
	(0.00)	(0.00)	(0.00)	(0.00)
pH	6.24	-0.09***	6.42	0.67***
	(0.02)	(0.01)	(0.02)	(0.02)
Carbon (%)	3.16	-0.19***	1.49	-0.33***
	(0.03)	(0.01)	(0.03)	(0.03)
Total Nitrogen (%)	0.27	0.01***	0.11	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)
Electrical conductivity (dS/m)	0.13	-0.01***	0.06	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)
Clay (%)	63.90	-1.14***	56.41	17.54***
	(0.32)	(0.08)	(0.55)	(0.53)
Silt (%)	23.28	0.66***	20.69	0.03
	(0.19)	(0.04)	(0.15)	(0.21)
Number of plots	*1,675*		*879*	

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in Ethiopia LASER and geospatial data sources. LASER and MAPS1 topsoil samples are at a soil depth of 0-20cm, AFSIS topsoil is a weighted avg of 0-5/5-15cm, and iSDA topsoil is at soil depths of 0-20cm. The weighted soil index includes pH, carbon, total nitrogen and electrical conductivity. The sample is restricted to plots with non-missing soil properties in household surveys and non-missing soil properties in geospatial data.

Table C.2. Imputation-Based vs. Benchmark Estimates (for the zones of the benchmark surveys), Top Soil Analysis

Soil Properties	*Ethiopia (2013/14)*		*Uganda (2015/16)*	
	Benchmark (LASER)	**Imputed (ESS 2)**	**Benchmark (MAPS 1)**	**Imputed (UNPS 5)**
	(1)	*(2)*	*(3)*	*(4)*
Weighted Soil Index	0.47	**0.47**	0.47	**0.46**
	(0.00)	(0.01)	(0.00)	(0.01)
Electrical Conductivity (dS/m)	0.13	**0.13**[a]	0.06	**0.06**[a]
	(0.01)	(0.01)	(0.00)	(0.00)
Number of plots	1,672	608	877	163

Note: Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of benchmark rate. Standard errors are in parentheses. Estimates are obtained using the PMM method, with 50 iterations. The target survey is restricted to the same zones (Ethiopia) or districts (Uganda) as the benchmark survey. The distributions of the control variables between the benchmark and target surveys are shown in Tables A.4 and A.5, Appendix A.

Appendix D: Estimation results using machine learning

Table D.1. Benchmark Survey vs. Imputation-Based Estimates (for the zones of the benchmark surveys), Ethiopia (2013/14), Top Soil Analysis

Soil Properties	Benchmark (LASER)	Imputed (ESS2)		
		LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*
Weighted Soil Index	0.53	**0.53**[a]	**0.53**[a]	**0.53**[a]
	(0.00)	(0.00)	(0.00)	(0.00)
		Components of soil index		
pH	6.25	**6.22**	**6.22**	6.20
	(0.02)	(0.01)	(0.01)	(0.01)
Organic Carbon (%)	3.16	**3.19**[a]	**3.19**[a]	**3.20**
	(0.03)	(0.02)	(0.02)	(0.02)
Total Nitrogen (%)	0.27	**0.27**[a]	**0.27**[a]	**0.27**
	(0.00)	(0.00)	(0.00)	(0.00)
		Soil particle composition		
Clay (%)	63.91	**64.27**	**64.27**	**64.10**[a]
	(0.32)	(0.16)	(0.16)	(0.19)
Silt (%)	23.29	**23.18**[a]	**23.18**[a]	**23.45**[a]
	(0.19)	(0.13)	(0.13)	(0.13)
Sand (%)	12.82	**12.51**	**12.51**	**12.54**
	(0.18)	(0.07)	(0.07)	(0.10)
Number of plots	1,672		608	

Note: Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of benchmark rate. Standard errors are in parentheses. The target survey is restricted to the same zones as the benchmark survey. Models are trained in the benchmark survey and tested against the target survey. Imputation models with statistics for Lasso and Elastic Net using postelection coefficient estimates are shown in Table D.7, Appendix D. Importance matrix of the variables is shown in Table D.8, Appendix D.

Table D.2. Benchmark Survey vs. Imputation-Based Estimates (for the districts of the benchmark surveys), Uganda (2015/16), Top Soil Analysis

Soil Properties	Benchmark (MAPS1)	Imputed (UNPS5)		
		LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*
Weighted Soil Index	0.57	**0.56**	**0.56**	0.56
	(0.00)	(0.00)	(0.00)	(0.00)
		Components of soil index		
pH	6.42	6.47	6.47	**6.45**
	(0.02)	(0.00)	(0.00)	(0.01)
Carbon (%)	1.49	**1.45**	**1.45**	1.43
	(0.03)	(0.02)	(0.02)	(0.02)
Total Nitrogen (%)	0.11	0.11	0.11	0.11
	(0.00)	(0.00)	(0.00)	(0.00)
		Soil particle composition		
Clay (%)	56.48	**56.18**[a]	**56.19**[a]	**56.35**[a]
	(0.55)	(0.41)	(0.41)	(0.55)
Silt (%)	20.67	**20.52**	**20.52**	20.16
	(0.15)	(0.07)	(0.07)	(0.15)
Sand (%)	22.85	**23.20**[a]	**23.25**[a]	**23.26**[a]
	(0.44)	(0.34)	(0.34)	(0.43)
Number of plots	877		163	

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of benchmark rate. Standard errors are in parentheses. The target survey is restricted to the same districts as the benchmark survey. Models are trained in the benchmark survey and tested against the target survey. Imputation models with statistics for Lasso and Elastic Net using postselection coefficient estimates are shown in Table D.9, Appendix D. Importance matrix of the variables is shown in Table D.10, Appendix D.

Table D.3. Benchmark Survey vs. Imputation-Based Estimates with Additional Geospatial Soil Quality Information (for the zones of the benchmark surveys), Ethiopia (2013/14), Top Soil Analysis

Soil Properties	Benchmark (LASER)	Imputed with iSDAsoil (ESS2)			Imputed with SoilGrids (ESS2)		
		LASSO	Elastic Net	Random Forest	LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Weighted Soil Index	0.53 (0.00)	**0.53**[a] (0.00)	**0.53**[a] (0.00)	**0.53**[a] (0.00)	0.52 (0.00)	0.52 (0.00)	0.52 (0.00)
			Components of soil index				
pH	6.23 (0.02)	6.15 (0.01)	6.15 (0.01)	6.12 (0.02)	6.13 (0.01)	6.13 (0.01)	6.17 (0.01)
Organic Carbon (%)	3.17 (0.03)	3.29 (0.02)	3.29 (0.02)	3.32 (0.01)	3.36 (0.02)	3.36 (0.02)	3.31 (0.02)
Total Nitrogen (%)	0.27 (0.00)	0.28 (0.00)	0.28 (0.00)	0.28 (0.00)	0.28 (0.00)	0.28 (0.00)	0.28 (0.00)
			Soil particle composition				
Clay (%)	63.84 (0.34)	66.65 (0.28)	66.60 (0.28)	**63.69**[a] (0.37)	65.35 (0.22)	65.35 (0.22)	**64.27** (0.18)
Silt (%)	23.37 (0.20)	24.55 (0.12)	24.55 (0.12)	24.22 (0.12)	**23.75** (0.11)	**23.76** (0.11)	**23.41**[a] (0.13)
Sand (%)	12.82 (0.19)	10.96 (0.11)	10.96 (0.11)	11.77 (0.10)	11.10 (0.11)	11.10 (0.11)	11.42 (0.11)
Number of plots	1,529		608			608	

Note: Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. The target survey is restricted to the same zones as the benchmark survey. Models are trained in the benchmark survey and tested against the target survey.

Table D.4. Benchmark Survey vs. Imputation-Based Estimates with Additional Geospatial Soil Quality Information (for the districts of the benchmark surveys), Uganda (2015/16), Top Soil Analysis

Soil Properties	Benchmark (MAPS1)	Imputed with iSDAsoil (UNPS5)			Imputed with SoilGrids (UNPS5)		
		LASSO	Elastic Net	Random Forest	LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Weighted Soil Index	0.57	**0.57**[a]	**0.57**[a]	0.58	0.52	0.52	0.55
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)
Components of soil index							
pH	6.42	**6.41**[a]	**6.41**[a]	6.38	6.26	6.26	**6.44**
	(0.02)	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.01)
Carbon (%)	1.49	1.57	1.57	1.60	**1.50**[a]	**1.50**[a]	1.55
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
Total Nitrogen (%)	0.11	0.14	0.14	0.13	**0.12**	**0.12**	**0.12**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Soil particle composition							
Clay (%)	56.46	59.03	59.03	58.56	54.83	54.79	**56.31**[a]
	(0.55)	(0.55)	(0.55)	(0.52)	(0.59)	(0.58)	(0.53)
Silt (%)	20.68	20.34	20.35	19.77	20.35	20.35	**20.41**
	(0.15)	(0.13)	(0.12)	(0.16)	(0.09)	(0.09)	(0.12)
Sand (%)	22.86	19.89	19.96	19.93	19.57	19.57	20.80
	(0.44)	(0.43)	(0.43)	(0.46)	(0.91)	(0.91)	(0.52)
Number of plots	876	158			158		

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. The target survey is restricted to the same districts as the benchmark survey. Models are trained in the benchmark survey and tested against the target survey.

Table D.5. Imputation-Based Estimates with Additional Geospatial Soil Quality Information vs. Geospatial Estimates (for the remaining areas), Ethiopia (2013/14), Top Soil Analysis

Soil Properties	iSDA (ESS2)	Imputed with iSDAsoil (ESS2)			SoilGrids (ESS2)	Imputed with SoilGrids (ESS2)		
		LASSO	Elastic Net	Random Forest		LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
Weighted Soil Index	0.62	0.52	0.52	0.52	0.64	0.37	0.37	0.37
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Components of soil index								
pH	6.12	6.42	6.43	6.30	6.32	6.15	6.15	6.15
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Organic Carbon (%)	1.77	3.05	3.05	3.15	3.26	3.24	3.24	3.31
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Total Nitrogen (%)	0.18	0.26	0.26	0.26	0.27	0.28	0.28	0.28
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Soil particle composition								
Clay (%)	35.73	65.65	65.61	64.16	37.19	65.53	65.56	63.75
	(0.03)	(0.05)	(0.05)	(0.06)	(0.04)	(0.05)	(0.05)	(0.06)
Silt (%)	25.04	23.42	23.42	23.48	32.09	22.40	22.40	23.16
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Sand (%)	37.88	12.08	12.08	12.36	30.62	12.08	12.08	12.36
	(0.04)	(0.03)	(0.03)	(0.02)	(0.05)	(0.03)	(0.03)	(0.02)
Number of plots	20,575				20,575			

Note: Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Target survey is restricted to the remaining zones, other than the zones/districts of benchmark survey. Models are trained in the benchmark survey and tested against the target survey.

Table D.6. Imputation-Based Estimates with Additional Geospatial Soil Quality Information vs. Geospatial Estimates (for the remaining areas), Uganda (2015/16), Top Soil Analysis

Soil Properties	iSDA (UNPS5)	Imputed with iSDAsoil (UNPS5)			SoilGrids (UNPS5)	Imputed with SoilGrids (UNPS5)		
		LASSO	Elastic Net	Random Forest		LASSO	Elastic Net	Random Forest
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
Weighted Soil Index	0.64	0.57	0.57	0.57	0.63	0.56	0.56	0.56
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Components of soil index								
pH	5.89	6.43	6.43	6.43	5.89	6.50	6.50	6.46
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Carbon (%)	1.52	1.56	1.56	1.59	3.29	1.54	1.54	1.41
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Total Nitrogen (%)	0.15	0.14	0.14	0.13	0.25	0.14	0.14	0.11
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Soil particle composition								
Clay (%)	31.51	54.63	54.63	54.82	35.67	52.84	52.83	54.39
	(0.09)	(0.13)	(0.13)	(0.10)	(0.07)	(0.09)	(0.09)	(0.09)
Silt (%)	20.08	21.79	21.75	21.65	24.99	21.51	21.51	20.98
	(0.03)	(0.02)	(0.02)	(0.02)	(0.05)	(0.02)	(0.02)	(0.03)
Sand (%)	46.97	24.57	24.65	24.42	38.75	23.43	23.43	24.46
	(0.11)	(0.09)	(0.09)	(0.07)	(0.07)	(0.13)	(0.13)	(0.07)
Number of plots		4,065				4,065		

Note: Acidified carbon is used in Uganda MAPS1 and organic carbon is used in geospatial data sources. Estimates shown in boldface or with "a" respectively fall within the 95% confidence interval or one standard error of imputation-based rate. Standard errors are in parentheses. Target survey is restricted to the districts, other than the zones/districts of benchmark survey. Models are trained in the benchmark survey and tested against the target survey.

Table D.7. The list of selected variables in Lasso and Elastic Net models with penalized standardized coefficients, Ethiopia

	LASSO							ELASTIC NET						
	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand
Head's age	-0.001	-0.001	-0.030	-0.001		-0.096		-0.001		-0.027	-0.001		-0.095	
Head's gender	-0.003	-0.007	-0.066	-0.005	0.698	-0.791	-0.008	-0.003		-0.058	-0.004	0.699	-0.785	-0.023
Primary	0.001	-0.036	0.044	0.004		0.153	-0.076	0.001	-0.028	0.040	0.004		0.152	-0.100
Secondary	0.003		0.076	0.007		0.149	-0.170	0.003		0.070	0.007		0.147	-0.187
Higher	-0.002	-0.043		-0.001	0.037	0.029	-0.194	-0.002	-0.034		-0.001	0.050	0.028	-0.209
Log of household size	0.003	0.017	0.060	0.004	-1.164	0.754	0.454	0.003	0.007	0.053	0.004	-1.158	0.746	0.451
Type of crop stand	-0.004	-0.064	-0.013	-0.005	0.101	0.062	-0.342	-0.004	-0.059	-0.008	-0.005	0.111	0.064	-0.337
Crop rotation	-0.007	-0.010	-0.156	-0.017	0.657	-0.595	-0.136	-0.007	-0.007	-0.150	-0.017	0.658	-0.587	-0.156
Plot is prevented from erosion	-0.001	-0.102	0.053	0.006	0.703	-0.558	-0.197	-0.001	-0.096	0.048	0.005	0.706	-0.554	-0.200
Plot is tilled	0.001	-0.109	0.119	0.008	-1.514	2.006	-0.236	-0.000	-0.015	0.010	0.001	-0.885	0.869	0.082
Fertilizer (inorganic)	-0.001	-0.018		-0.003			-0.062	0.001	-0.100	0.113	0.008	-1.505	1.990	-0.205
Fertilizer (organic)	0.001	0.011			0.004		-0.152	-0.001	-0.013		-0.002			-0.082
Pesticides are used in the plot	-0.010	0.112	-0.342	-0.031	-0.668	-0.211	0.998	0.001	0.001			0.018		-0.157
Hired labor		-0.021	0.016	0.002	-0.886	0.876	0.040	-0.010	0.101	-0.335	-0.030	-0.679	-0.204	0.902
Log of the plot area		-0.004	-0.002	0.001	-0.561	0.127	0.514				0.001	-0.566	0.126	0.473
MSE	0.00	0.55	1.14	0.01	166.22	51.74	54.87	0.00	0.55	1.14	0.01	166.22	51.74	54.87
Rsquared	0.10	0.08	0.12	0.12	0.05	0.11	0.04	0.10	0.08	0.12	0.12	0.05	0.11	0.04

Note: Variables are standardized

Table D.8. Variable importance scores in Random Forest, Ethiopia

	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.740	0.851	0.778	0.712	0.855	0.431	0.915
Head`s gender	0.608	0.622	0.609	0.543	0.663	0.411	0.588
Primary	0.510	0.682	0.565	0.516	0.574	0.291	0.621
Secondary	0.734	0.565	0.707	0.718	0.436	0.230	0.455
Higher	0.334	0.716	0.219	0.223	0.535	0.336	0.440
Log of household size	0.666	0.706	0.707	0.651	0.726	0.355	0.786
Type of crop stand	0.659	0.783	0.506	0.512	0.591	0.263	0.789
Crop rotation	1.000	0.595	0.907	1.000	0.606	0.311	0.600
Plot is prevented from erosion	0.570	0.940	0.574	0.526	0.668	0.346	0.652
Plot is tilled	0.743	0.614	0.626	0.618	0.675	0.450	0.628
Fertilizer (inorganic)	0.483	1.000	0.506	0.436	1.000	1.000	0.507
Fertilizer (organic)	0.558	0.577	0.594	0.560	0.546	0.262	0.655
Pesticides are used in the plot	0.426	0.554	0.309	0.382	0.540	0.352	0.351
Hired labor	0.373	0.551	0.296	0.350	0.619	0.341	0.464
Log of the plot area	0.868	0.932	1.000	0.910	0.905	0.458	1.000

Note: The values are scaled proportional to the largest value in the set.

Table D.9. The list of selected variables in Lasso and Elastic Net models with penalized standardized coefficients, Uganda

	LASSO							ELASTIC NET						
	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.007	0.012	0.062	0.005	0.903	-0.004	-0.885	0.006	0.013	0.055	0.005	0.800	-0.003	-0.733
Head`s gender	0.004	-0.007	0.044	0.004	0.496		-0.604	0.003	-0.008	0.039	0.004	0.411		-0.479
Primary education	0.002		0.029					0.002		0.026				
Secondary education	-0.001			-0.002	-0.036		0.256				-0.002			0.182
Vocational education	-0.004		-0.036	-0.004	-0.799		0.839	-0.003		-0.033	-0.004	-0.725		0.725
Higher education	-0.001		-0.010	-0.001			-0.064	-0.000		-0.008	-0.001			-0.009
Log of household size	-0.007	-0.014	-0.063	-0.006	-0.461		0.547	-0.007	-0.014	-0.060	-0.006	-0.448		0.525
Type of crop stand	-0.003		-0.024	-0.003	-1.450	0.278	1.128	-0.002		-0.021	-0.002	-1.380	0.275	1.033
Fertilizer (inorganic)	-0.001		-0.009	-0.002			0.066	-0.001		-0.005	-0.001			
Fertilizer (organic)	0.017	-0.052	0.188	0.018	3.604	-0.545	-3.014	0.016	-0.051	0.179	0.018	3.464	-0.538	-2.787
Pesticides	-0.001	-0.008			0.314	-0.106	-0.201	-0.000	-0.009			0.262	-0.105	-0.131
Cultivated in the previous season	-0.007	-0.012	-0.069	-0.006	-0.364		0.504	-0.006	-0.012	-0.066	-0.005	-0.349		0.478
Log of the parcel area	-0.010	-0.006	-0.100	-0.009	-2.924	0.542	2.309	-0.009	-0.007	-0.095	-0.009	-2.813	0.535	2.138
Problems with erosion	-0.008	-0.022	-0.085	-0.005	-1.687	0.316	1.324	-0.007	-0.022	-0.080	-0.005	-1.589	0.311	1.193
Hired labor				-0.000							-0.000			
MSE	0.00	0.20	0.48	0.00	223.46	18.17	146.90	0.00	0.20	0.48	0.00	223.55	18.17	146.94
Rsquared	0.15	0.04	0.16	0.17	0.15	0.07	0.15	0.15	0.04	0.16	0.17	0.15	0.07	0.15

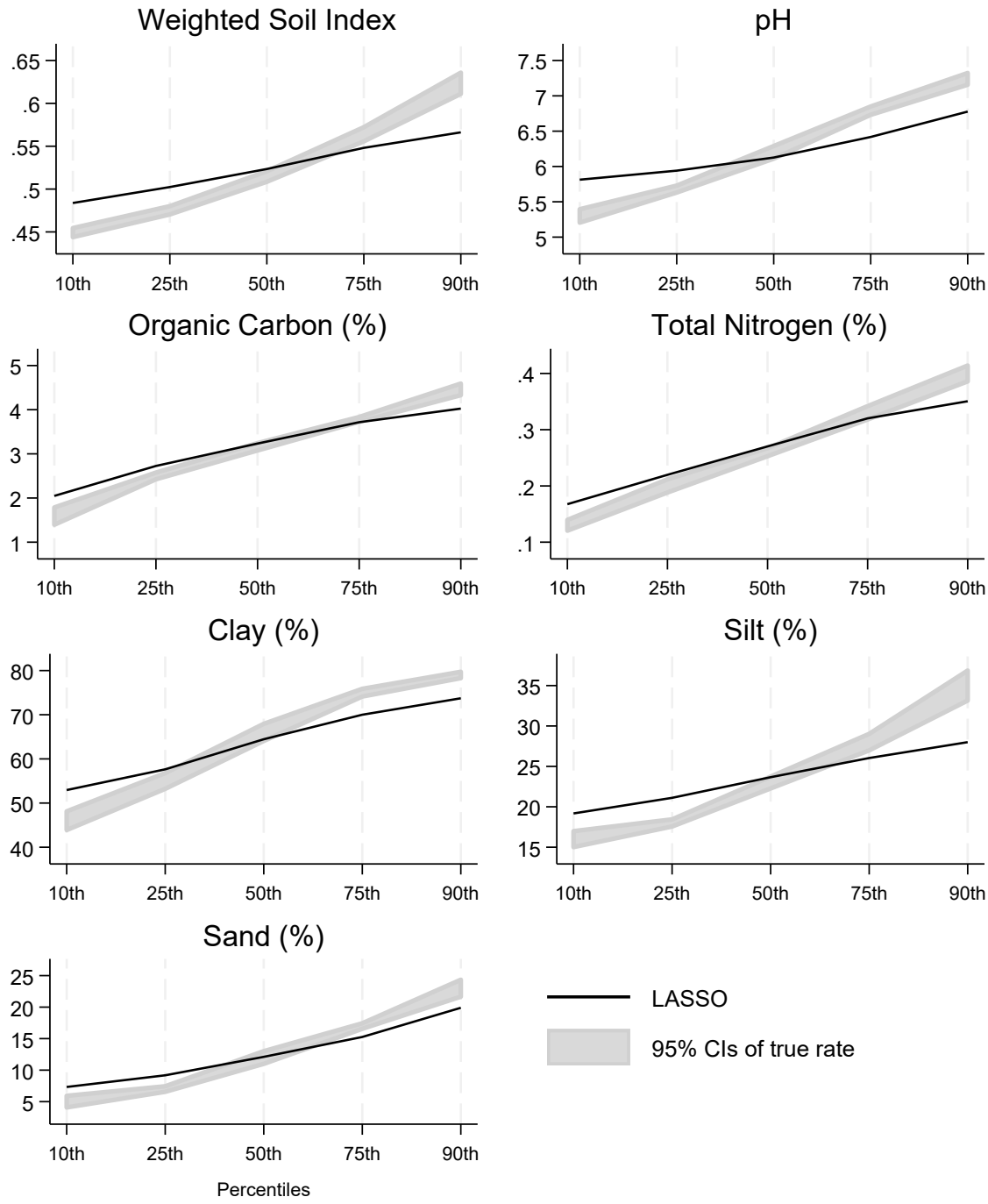
Note: Variables are standardized

Table D.10. Variable importance scores in Random Forest, Uganda

	WSI	pH	Carbon	Total Nitrogen	Clay	Silt	Sand
Head`s age	0.514	0.941	0.455	0.406	0.604	0.356	0.270
Head`s gender	0.335	0.686	0.294	0.261	0.398	0.290	0.218
Primary education	0.332	0.662	0.290	0.261	0.364	0.259	0.182
Secondary education	0.300	0.684	0.282	0.257	0.389	0.301	0.188
Vocational education	0.426	0.715	0.346	0.334	0.542	0.349	0.341
Higher education	0.154	0.523	0.109	0.090	0.130	0.328	0.113
Log of household size	0.468	0.821	0.414	0.371	0.540	0.331	0.259
Type of crop stand	0.362	0.691	0.307	0.290	0.487	0.393	0.239
Fertilizer (inorganic)	0.347	0.688	0.320	0.271	0.458	0.410	0.226
Fertilizer (organic)	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Pesticides	0.318	0.762	0.306	0.259	0.369	0.353	0.200
Cultivated in the previous season	0.562	0.760	0.503	0.443	0.463	0.341	0.226
Log of the parcel area	0.510	0.946	0.444	0.410	0.632	0.442	0.321
Problems with erosion	0.399	0.675	0.359	0.305	0.525	0.424	0.281
Hired labor	0.315	0.567	0.273	0.240	0.380	0.331	0.188

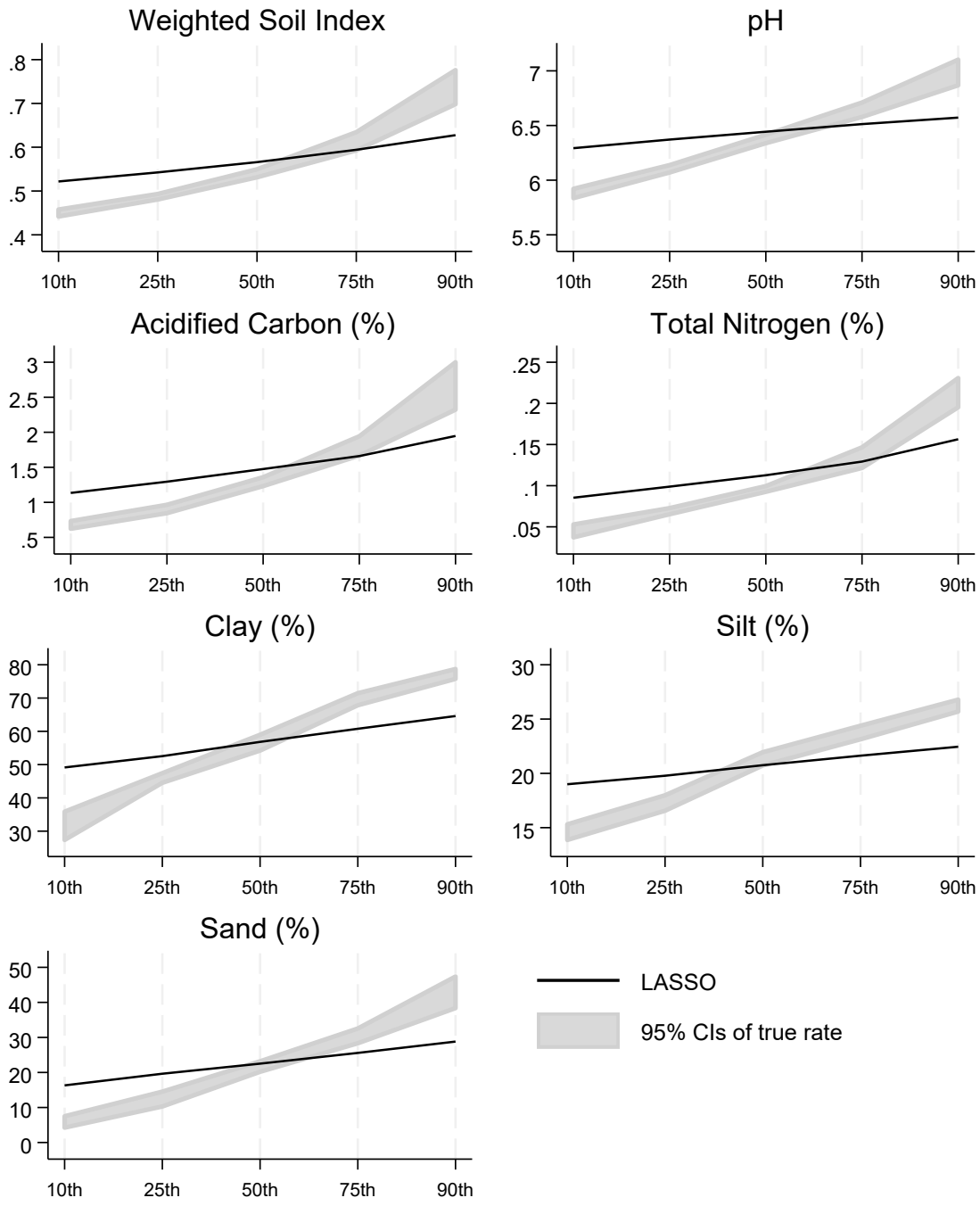
Note: The values are scaled proportional to the largest value in the set.

Figure D.1. LASSO Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Sample (Base Survey), Top Soil Analysis, Ethiopia (2013/14)



Note: The estimation sample is generated by splitting LASER data into two random samples: 50% in-sample, 50% out-of-sample. Estimates are obtained using 50 iterations using "out-of-sample" data as the target sample. Simultaneous quintile regression with bootstrapping SEs was used to get multiple imputed quintiles in the target sample. Total "in-sample" size is 837 plots, the "out-of-sample" size is 837 plots.

Figure D.2. LASSO Estimates of Soil Quality Index and its Components for Different Percentiles of the Benchmark Sample (Base Survey), Top Soil Analysis, Uganda (2015/16)



Note: The estimation sample is generated by splitting MAPS1 data into two random samples: 50% in-sample, 50% out-of-sample. Estimates are obtained using 50 iterations using "out-of-sample" data as the target sample. Simultaneous quintile regression with bootstrapping SEs was used to get multiple imputed quintiles in the target sample. Total "in-sample" size is 438 plots, the "out-of-sample" size is 438 plots.