Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 11496

Scanning for Soil Health

Assessing the Accuracy and Scalability of Innovations for In-Situ Soil Measurement

Karan S. Shakya
Leah E. M. Bevis
Sydney Gourlay



WORLD BANK GROUP
Development Economics
Development Data Group
November 2026



Reproducible Research Repository

A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

Policy Research Working Paper 11496

Abstract

Cost-effective and scalable soil testing tools can play an instrumental role in guiding agricultural decisions and policies, particularly in Sub-Saharan Africa, where declining soil fertility poses a serious threat to productivity. Laboratory tests of soil samples, which provide the most accurate measures of soil properties, remain costly and challenging to scale in the context of national surveys or programs. To test and validate alternative approaches to soil measurement, the World Bank launched the Uganda Climate, Land Area, and Soil Study project. This paper evaluates the accuracy and scalability of three soil testing tools (the AgroCares Scanner, the Palintest kit, and the Sonkir pH meter) as well as two publicly available soil maps (Innovative Solutions for Decision Agriculture and SoilGrids), against laboratory-based soil measurements conducted by the Center for International Forestry Research and the International Centre for Research in Agroforestry. While differences in soil parameter estimates may be expected to vary across testing methods, understanding the nature and magnitude of the divergence in estimates from those measured using established laboratory approaches is critical for practitioners and analysts seeking to leverage soil analyses to inform policy and program design. The paper finds that the soil tools—except the Sonkir pH meter, which performs poorly—provide only moderately adequate predictions of soil properties, and both soil maps fail to capture sample variation. These findings highlight key trade-offs in accuracy, cost, and scalability and offer practical guidance for scaling soil monitoring strategies in Uganda and similar contexts.

This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at sgourlay@worldbankgroup.org. A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.



The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Scanning for Soil Health: Assessing the Accuracy and Scalability of Innovations for In-Situ Soil Measurement

Karan S. Shakya*, Leah E. M. Bevis†, and Sydney Gourlay‡ §

Authorized for distribution by Talip Kilic, Senior Program Manager, Development Data Group, Development Economics, World Bank Group

Keywords: soil fertility, soil tests, spectroscopy, soil maps, Uganda

*Department of Environmental Studies and Economics, Kenyon College.

†Department of Agricultural, Environmental and Development Economics, The Ohio State University.

‡Living Standards Measurement Study (LSMS), Survey Unit, Development Data Group, World Bank.

§This paper was produced with financial support from the Korea Green Growth Trust Fund and the 50x2030 Initiative to Close the Agricultural Data Gap, a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030. For more information visit 50x2030.org. We thank additional World Bank colleagues involved in the design, implementation, and data processing related to the Uganda CLASS study, including Emanuele Clemente, John Ilukor, Siobhan Murray, Adriana Paolantonio, and Federica Petruccelli. We are also grateful to Hope Michelson, Armine Juergenliemk, and the participants of the 2024 International Conference of Agricultural Economists, the 2024 50x2030 Methods & Tools Expert Convening, and the World Bank's 2025 Better Data for Better Lives and Jobs conference for their insightful comments and suggestions on an earlier version of this work. Finally, we are grateful to the teams at CIFOR-ICRAF, AgroCares, and Jabba Engineering for their involvement in the training for the survey, support throughout fieldwork implementation, and subsequent feedback and discussions around this work.

1 Introduction

Accurate measurement of soil health plays a critical role in supporting livelihoods in Sub-Saharan Africa. Improved soil conditions correlate with lower malnutrition rates (Sanchez and Swaminathan 2005) and higher incomes (Bhargava, Vagen, and Gassner 2018; Heger, Zens, and Bangalore 2020), but much of the region experiences rapid land degradation (Barbier and Hochard 2018) and declining soil fertility (Sanchez 2002; Tully et al. 2015). Compounding these challenges, agricultural productivity in the region has stagnated relative to other parts of the world, in part due to persistently low fertilizer use (Suri and Udry 2022), heavily constraining the yield gains needed to raise household welfare. Additionally, substantial heterogeneity in soil properties across agricultural plots (Kihara et al. 2016; Hengl et al. 2021) complicates policy implementation and the dissemination of agricultural technologies. These challenges highlight the need for accurate, high-resolution soil measurements to effectively target interventions and sustain soil health in the region.

Because comprehensive laboratory testing for every agricultural plot is logistically and financially unfeasible, most efforts to measure soil health in Africa rely on either farmer self-assessments or publicly available spatial soil maps. Both approaches have notable limitations. Farmers' self-reported assessments often misrepresent actual soil conditions, as they typically reflect observable traits like crop yield or soil color rather than intrinsic soil properties (Berazneva et al. 2018; Gourlay et al. 2017; Marenya, Barrett, and Gulick 2008). Similarly, soil maps – often produced by modeling existing soil sample databases with machine-learning methods and satellite-derived covariates – can miss fine-scale variation, reducing their usefulness for plot- or village-level decision-making (Berazneva et al. 2018; Gourlay et al. 2017).

In this paper, we compare soil characteristics of samples collected across five districts in Uganda, through the World Bank's Uganda Climate, Land Area, and Soil Study (CLASS), as measured by four different soil testing tools and two spatial soil maps. The soil test tools include a mid-infrared spectroscopy-based test by CIFOR-ICRAF (our benchmark, validated with wet chemistry soil analysis); a rapid, in-situ near-infrared spectroscopy-based tool using the AgroCares Scanner; a portable soil test kit developed by Palintest and implemented by Jabba Engineering in Uganda; and the low-cost Sonkir pH meter designed specifically for measuring soil pH in the field. The two soil maps are iSDA and SoilGrids, both widely used to analyze soil properties in Sub-Saharan Africa. We also incorporate sample-specific survey data on self-reported soil quality to examine how well farmers' perceptions align with measured soil properties. Although estimates of soil properties are expected to differ to some degree across tools – as a function of the different methods, calibration libraries, and intensity of soil processing employed by

the tools – it is essential that practitioners and analysts seeking to utilize soil analyses to inform policies or programs have an understanding of the nature and magnitude of such differences. Ultimately, we aim to illuminate how methodological differences across soil testing tools can meaningfully influence agricultural diagnostics, advisory services, or interventions, particularly for survey practitioners and program implementers who lack specialized soil science knowledge.

We find that farmers' self-reported assessments poorly reflect both soil attributes and biophysical characteristics. Although most farmers cite observable indicators such as crop yields and soil color as key criteria for evaluating soil quality, these factors do not consistently inform their final assessments. These inconsistencies highlight the limitations of perception-based evaluations and the value of objective soil measurement tools.

We also find that the three soil test tools and the two soil maps generally perform poorly compared to the benchmark measurements from CIFOR-ICRAF, with variation in performance across soil parameters. We focus on five key soil parameters: pH, soil organic carbon, clay content, total nitrogen, and extractable potassium. The AgroCares Scanner measure of clay is fairly good, and its measures of soil organic carbon and total nitrogen are on the low margin of acceptable. The Palintest kit performs acceptably for pH only. Neither measures extractable potassium well, and the Sonkir pH meter performs very poorly in measuring pH, with essentially zero predictive power. Both soil maps provide nearly uniform predictions across all sample plots, failing to capture within-region variation, though reasonably estimating regional averages. Between the two, iSDA generally outperforms SoilGrids across most indicators (but neither map captures any variation in soil pH).

We also illustrate that assessing the similarity of measures using more aggregated, agronomically relevant categories (such as low, moderate, and high levels of a given property) does not alleviate all concerns about variation in parameter estimates. Even for soil tests that are the most correlated with ICRAF's measurements, soil properties are frequently misclassified across these categorical bins, with evidence of both systematic under- and over-estimation. For example, ICRAF categorized most soil samples (88%) as 'moderately acidic', but Palintest, the soil test with the strongest correlation for pH, correctly classified only 53% of these. Similarly, for soil organic carbon, although AgroCares had the strongest correlation with ICRAF, it misclassified approximately 47% of samples across the 'low', 'moderate', and 'high' categories.

Finally, we emphasize that predictive accuracy alone does not determine the usefulness of a soil test. Scalability is also critical, especially when the goal is to enhance soil monitoring at a national scale. CIFOR-ICRAF's laboratory-based MIR analysis, used as the benchmark in our comparisons, requires transporting samples to a lab, with costs de-

pending on sample volume and turnaround times depending on lab capacity. While more portable tools such as the AgroCares Scanner and Palintest kit offer greater flexibility, they also entail substantial fixed and operational costs, including licensing fees and the logistical challenges of training and deployment. These trade-offs must be considered when designing scalable soil assessment strategies for agricultural programs and policy.

This paper contributes to the growing literature on the effectiveness of portable soil test kits as tools to inform farmers about their soils and support decisions that increase agricultural productivity. In Sub-Saharan Africa, several studies have demonstrated the successful use of portable soil test kits to guide fertilizer recommendations (Harou et al. 2022) and incentivize farmers to engage more actively with soil health information (Berazneva et al. 2023). There is also increasing interest in spectroscopy as a cost-effective and scalable method for soil testing (Michelson 2025; Ahmadi et al. 2021), with recent projects using spectroscopy tools to advise farmers on soil health and fertility management (Nyondo et al. 2025). Ultimately, these tools can provide cost-effective soil measurement, with the potential to give farmers key recommendations for optimizing agricultural inputs (Murphy et al. 2020; Gars et al. 2025). Our paper compares two spectroscopy-based soil tests and two non-spectroscopy tests, evaluating their effectiveness and scalability for measuring key soil properties at a large scale.

This paper also contributes to the extensive literature on self-reported measures of soil fertility and the limitations of such assessments. Prior studies show that farmers commonly rely on observable indicators such as crop yields, soil color, texture, and the presence of weeds to evaluate soil quality (e.g., Dawoe et al. 2012; Desbiez et al. 2004). Even for essential agricultural inputs like fertilizers, farmers have been shown to rely on visual cues to assess their efficacy (Michelson et al. 2021). However, measures based on these perceptions often show weak correlations with actual soil fertility characteristics (Gourlay et al. 2017; Odendo, Obare, and Salasya 2010; Gray and Morant 2003), raising concerns about how misconceptions of soil quality can lead to inefficient agricultural decisions (Berazneva et al. 2018; Marenya, Barrett, and Gulick 2008). Our findings are consistent with this literature. Farmers in Uganda report using observable characteristics, such as soil color and yields, to assess soil quality. However, we find that these self-assessments correlate poorly with actual yields and soil color.

The paper proceeds as follows. Section 2 describes the study setting, survey data, and the soil tests and maps used for comparison. Section 3 presents the results, beginning with an assessment of how self-reported soil quality aligns with measured soil properties, followed by a comparison and discussion of each soil test and map across key soil parameters. Finally, Section 4 concludes.

2 Setting and Data

2.1 Uganda CLASS

Historically, household and agricultural surveys, including those in Uganda, have relied heavily on farmers' subjective assessments of soil health, or where georeferenced plot information is available, on integration with gridded geospatial soil products. However, there is growing demand for objective soil measures that capture actual, spatially varying soil fertility information. At the same time, technological innovation presents opportunities for objective soil analysis in a cost-efficient and scalable manner. To better understand this potential, the World Bank designed and launched the Uganda Climate, Land Area, and Soil Study (CLASS) to test and validate innovative soil measurement approaches that have the potential to transform soil data collection in agricultural surveys, while assessing trade-offs in accuracy, cost, and scalability.[5]

Soil tests and surveys took place across five districts in Uganda's Eastern and Northern sub-regions. Figure 1 maps these districts: Apac, Dokolo, Buyende, Kamuli, and Kaliro. The districts were selected based on multiple criteria, including the share of households growing maize, climatological variability, the existence of weather stations managed by the Uganda National Meteorological Authority, and network connectivity. A total of 75 enumeration areas were selected from those 5 districts and 12 maize-growing households were selected in each enumeration area.[6]

The survey was conducted between August 2023 and March 2024 with the help of EDI Global and consisted of four phases: listing, post-planting, crop-cutting, and post-harvest. During the listing phase, survey teams conducted a complete listing of households in each selected enumeration area, including information to determine eligibility for inclusion in the study - namely cultivation of maize in the current agricultural season. In the post-planting phase, a household survey was administered and soil sample collection and analysis were conducted on up to two randomly selected plots per household. The crop-cutting phase centered on the harvest of maize from crop-cutting subplots. Finally, during

[5]The Uganda CLASS was led by the World Bank's Living Standards Measurement Study (LSMS) team under the aegis of the 50x2030 Initiative, and in partnership with complementary financial support from the World Bank's Agriculture and Food Global Practice. Uganda CLASS had two additional objectives: validating the use of tablet-based tools for area measurement and validating the use of community-level weather stations vis-à-vis status quo methods. All three objectives fall within the purview of the 50x2030 Initiative, which was launched by the World Bank, FAO, and IFAD to support more informed decision-making in promoting agricultural productivity, agricultural sustainability, and rural development. The program is active in several countries, including Uganda, with the goal of transforming agricultural data systems through both new surveys and improved measurement methods.

[6]This design led to a target of 900 sampled households. Ten replacement households per enumeration area were also identified to account for non-response.

the post-harvest phase, a questionnaire was administered to collect information on self-reported yields, harvest labor, and other agricultural inputs used throughout the season. All survey questions were administered using Computer-Assisted Personal Interviewing via the World Bank's Survey Solutions software.

2.2 Soil Measurements

We assess plot-level soil properties using several methods. First, farmers self-report plot soil quality using three ordinal categories: 'poor', 'fair', and 'good'. Farmers also identify and rank the top three criteria they use to evaluate soil health. Second, we test plot-level soil samples using a laboratory-based mid-infrared spectroscopy approach, conducted by CIFOR-ICRAF, and three soil testing tools (the AgroCares Scanner, the Palintest kit implemented by Jabba Engineering in local laboratories, and a store-bought Sonkir pH meter). Third, we examine predictions from two soil maps (iSDA Soil and SoilGrids) to assess the soil properties they assign to these same plots based on GPS location.

The four soil tests and two soil maps evaluate a range of key soil properties, with some, but not all, overlapping across methods (Table A1). This paper focuses on soil properties widely recognized as key indicators of soil fertility and agricultural productivity. These properties are covered, at least partially, by the selected soil tests and maps, allowing for cross-method comparison. The five indicators are: pH, clay content, soil organic carbon, total nitrogen, and extractable potassium.[7] While the different soil tests and maps aim to measure the same underlying property for the first four indicators, for potassium, the various tools differ in the pool of potassium they aim to measure (discussed further below). Given this, some disagreement across tools is expected, but we argue that understanding the differences in potassium estimates – whether extractable or exchangeable – is essential for practitioners not familiar with such nuances.

2.2.1 Soil Tests

We collected soil samples in the field following CIFOR-ICRAF's Land Degradation Surveillance Framework, with adaptations to fit the household survey sampling frame (Vagen and Winowiecki 2023). Under this framework, two samples, topsoil and subsoil, are collected from each agricultural plot. Both the topsoil and subsoil samples are composite samples collected from four points within the given agricultural plot, at 0-20cm and 20-50cm depths, respectively. This paper focuses on topsoil results, as the topsoil depth is

[7]Many types of soil tests attempt to measure plant-available soil mineral concentration rather than total soil mineral concentration. This is done by using an extraction solution that extracts the pool of the mineral that is in theory available to plants – though different solutions will extract slightly different pools of minerals, as we discuss later.

consistent across all measurement tools and soil maps. While soil samples were collected and sent to CIFOR-ICRAF and Jabba Engineering for analysis, soil characteristics were also measured in the field using the AgroCares Scanner and Sonkir pH meter. The four approaches for measuring soil properties in the field are described below.

The first approach is laboratory-based mid-infrared spectroscopy (MIR) implemented by CIFOR-ICRAF, which serves as our benchmark (henceforth referred to as *ICRAF MIR* for brevity). After collection, soil samples were transported to CIFOR-ICRAF's Soil and Plant Spectral Diagnostics Laboratory in Nairobi, where they were air-dried, passed through a 2mm sieve, fine-ground to <0.5mm, and analyzed using a Bruker Invenio-S Fourier-Transform Infrared spectrometer. Spectroscopy-based soil testing estimates soil properties by analyzing the light reflected from the sample, with specific spectral signatures corresponding to different soil characteristics. This method requires calibration, in which known soil properties from wet or dry laboratory tests are matched with their spectral profiles. ICRAF calibrates its MIR spectroscopy results against its global soil models, developed using wet chemistry data from across Sub-Saharan Africa and fitted using Bayesian Regularization for Feed-Forward Neural Networks and Random Forest algorithms. Additionally, 10% of our sample, 30 observations, underwent wet/dry chemistry analysis (the traditional forms of analysis for each soil characteristic), and the wet/dry chemistry values of 21 observations were used in the calibration of the ultimate spectroscopy model. Nine observations were used for validation for that local calibration process, and we do not know which observations were used for calibration vs. validation.

We use the ICRAF MIR predictions as the benchmark against which we compare alternative measurement tools and approaches — but admittedly, they are predictions themselves. Ideally, we would validate these predictions against the wet/dry chemistry results from the subsample of observations for which both were available; however, 21 of those 30 observations were used for the ICRAF MIR model calibration itself, so this comparison cannot serve as a true independent test. With that caveat, Figure 2 presents the comparison visually via scatterplots and by providing the correlations between these two measures, for each soil characteristic. The MIR estimates are highly correlated with the wet chemistry measurements, with Pearson's correlation coefficients above 0.95 in all cases but that of potassium ($\rho = 0.893$). Table A2 additionally provides means for each measure and demonstrates that a t-test cannot reject the equality of those means for any soil characteristic. If this sample-specific calibration played a significant role in shaping ICRAF's MIR model, it is unsurprising that the MIR estimates so closely mirror the wet/dry chemistry estimates, and this tight correlation would not hold outside of these 30 samples. If instead the ICRAF MIR predictions are largely shaped by ICRAF's global MIR model, with sample-specific calibration playing little role in altering the model, then in fact these correlations indicate that the MIR works very, very well to predict wet/dry

chemistry estimates. Unfortunately, we do not understand the model well enough to say which is the case.

Second is the AgroCares Scanner (Series-F model), which applies near-infrared (NIR) spectroscopy to estimate soil properties. While MIR measures fundamental molecular vibrations with well-defined absorption, NIR detects overtones and combinations of the fundamental bands detected by MIR, which can lead to weaker predictions (Shepherd and Walsh 2007; McCarty et al. 2002; Bellon-Maurel and McBratney 2011). Field operators break up the soil samples and place the scanner directly on the sample surface, all while in the agricultural plot. Like CIFOR-ICRAF's spectroscopy method, AgroCares' readings are calibrated against laboratory-measured reference data. The AgroCares global soil library contains roughly 17,500 soil samples from 36 countries, including samples from Uganda. Each reference sample was analyzed in the AgroCares Golden Standard Laboratory using standardized wet-chemistry methods, and its spectrum recorded from 1400-2500 nm, with resolution at 1% of wavelength.[8] These spectra-chemistry pairs were then used to train machine learning models, historically based on partial least squares regression and locally weighted learning, and more recently incorporating deep-learning architectures (Kok et al. 2024). It is important to note that AgroCares positions the scanner as a monitoring tool, acknowledging that it is not a replacement for traditional laboratory-based methods and recommends that adequacy classifications for soil parameters be used rather than point estimates.

The third method employs the SKW500 Palintest Complete Professional Soil Kit, implemented by Jabba Engineering Ltd., a Ugandan firm specializing in soil nutrient and water quality testing. The kit includes a range of equipment, but relevant to this paper are the multiparameter pocket sensor that utilizes an electrochemical approach to measure soil pH and a Bluetooth-enabled photometer that measures extractable potassium using a 0.1 molar magnesium acetate extractant. Each kit can be used to analyze 25 soil samples per day. Samples were collected directly from farmers' plots and sent to nearby regional laboratories, where trained Jabba technicians analyzed them using the Palintest kit. Because the kit does not measure soil texture, Jabba engineers measured soil texture separately using the Bouyoucos hydrometer sedimentation method.

The fourth method involves a low-cost pH meter designed specifically to measure soil pH: the Sonkir Soil pH Meter. The device features an integrated probe and an analog display and is inserted directly into soil to assess pH levels. The meter consists of two metallic probes: one functioning as a reference electrode, and the other as a measuring electrode. When inserted, it detects the electrical potential difference generated by hydrogen ion activity between the probes, then translates it to a corresponding pH value.

[8] More details can be found on their website.

The probe also measures soil moisture and flags when moisture levels are too low for a reliable pH reading; in such cases, enumerators moistened the soil with clean water before re-measuring. Nonetheless, as with most direct-insertion pH meters, readings can be affected by soil moisture, temperature, and salinity.

2.2.2 Soil Maps

We evaluate two publicly available and widely used soil maps to assess how closely their predicted soil properties align with benchmark ICRAF MIR values. The first map, iSDAsoil (referred to henceforth as iSDA), predicts soil properties using a stacked ensemble machine-learning framework that evaluates multiple learners, including Random Forest, gradient boosting, Cubist regression, neural networks, and generalized linear models with Lasso or Elastic Net regularization (Hengl et al. 2021). For each soil property, models are trained on datasets of laboratory-analyzed samples (wet chemistry and spectroscopy), compiled from multiple African soil databases. Although sample sizes vary by property, training datasets typically exceed 100,000 samples. Predictions are refined through a multi-scale modeling strategy: a coarse-scale model (250m covariates such as MODIS vegetation indices and climate data) is used to capture broad gradients, and a fine-scale model (30m covariates such as Sentinel-2 reflectance and high-resolution terrain derivatives) models the residual variation. The final 30m resolution maps thus integrate both large-scale environmental patterns and fine-scale spatial variability.

The second soil map, SoilGrids (specifically the SoilGrids 2.0 product), is developed and maintained by ISRIC - World Soil Information. Like iSDA, SoilGrids predicts soil properties from environmental covariates, but it differs in both its training data and modeling architecture (Poggio et al. 2021; Batjes, Calisto, and de Sousa 2024). SoilGrids draws from ISRIC's globally harmonized WoSIS database of approximately 230,000 soil samples, covering all continents. Predictions are produced with a single-scale stacked ensemble at 250m resolution, combining multiple base learners (e.g., Random Forest, gradient boosting, Cubist) within a quantile regression forest meta-model. Environmental predictors include a large set (≈400) of global covariates describing climate, vegetation, terrain, geology, and land cover. Unlike iSDA's explicitly multi-scale approach tuned for Africa, SoilGrids emphasizes globally consistent predictions at the coarser resolution, potentially at the expense of capturing local variation.

3 Results and Discussion

3.1 Self-Reported Soil Quality

Figure 3 illustrates average self-reported soil assessments across the five surveyed districts in Uganda. In Apac and Dokolo – the two northernmost districts – over 50% of farmers classify their soil as good, and fewer than 10% rate it as poor. In Buyende and Kamuli, responses are more evenly split between good and fair (each around 45%), with about 10% reporting poor soil. In Kaliro, the easternmost district, more than 65% of households classify their soil as fair, while roughly 20% describe it as good and 15% as poor. Farmers also identified the top three criteria they use to evaluate soil quality. Most cited crop yields and soil color, followed by soil type and the presence of weeds (Figure A1).

We begin by examining how closely farmers' self-reported soil quality aligns with both the soil characteristics they consider important and the underlying biophysical properties. We focus on crop yields and soil color, which are the two most frequently cited indicators, and use ICRAF MIR measurements for soil pH and soil organic carbon. Figure 4 illustrates the relationship between self-reported soil quality and these variables. There is no consistent association between pH and farmers' ratings. On average, farmers classify plots as poor, fair, or good regardless of the measured pH levels. A similar pattern emerges for soil organic carbon, where plots with low, moderate, or high levels appear across all three self-reported categories. These results demonstrate that, in this sample at least, farmers' evaluations do not reliably reflect underlying soil characteristics.[9]

Focusing on the more observable characteristics of soil color and crop yields, we find similar inconsistencies (Figure 4, bottom row). While farmers who describe their soil as good most frequently identify their soil color as black, many self-reportedly "good" soils are also described as red or some other color. This lack of a dominant color association is consistent across self-reported values of fair and poor soils. Similarly, there is no clear pattern between self-reported soil quality and observed crop yields. Farmers who rate their soil as good are found across plots with high, medium, and even low yields. The same holds for those reporting fair or poor soil quality.

These findings highlight the limitations of self-reported soil quality as a reliable indicator of actual soil health. They also reflect the inherent heterogeneity in how farmers classify their soil. Given such inconsistencies, it is especially important to use objective methods for measuring soil properties. Reliable diagnostics are essential for identifying soil constraints and designing appropriate interventions at the plot level.

[9]Similar analysis for other soil characteristics, not shown, demonstrates the same.

3.2 Correlations and Bias

Figure 5 presents a correlation matrix of ICRAF MIR results against results from the other three soil tests and two soil maps, across five key soil properties. The AgroCares Scanner measures clay content, soil organic carbon, and total nitrogen fairly well (correlation coefficients of 0.68, 0.45, and 0.43 respectively). However, it does not measure pH well (correlation coefficient of 0.09). In contrast, the Palintest kit measures pH well (correlation coefficient of 0.54), but Jabba's clay measure is uncorrelated with ICRAF MIR (correlation coefficient of -0.004). The Sonkir pH meter performs worst at capturing pH; its measurements are negatively correlated with ICRAF MIR (correlation coefficient of -0.10). The AgroCares measurements for exchangeable potassium have almost precisely zero correlation with the benchmark ICRAF MIR values of extractable potassium, which may be unsurprising since these two measures capture different pools of soil potassium. The Palintest measurement of extractable potassium correlates passably well with ICRAF MIR (correlation coefficient of 0.32).

Among the two soil maps, iSDA's measurements generally correlate more closely with ICRAF MIR-based measures than do those from SoilGrids. The iSDA measurements of soil organic carbon and total nitrogen are passably correlated with ICRAF MIR measurements (correlation coefficients of 0.37 and 0.35, respectively), and the iSDA measurement of clay content is weakly correlated with the ICRAF MIR measurement (correlation coefficient of 0.16). In comparison, SoilGrids measurements show consistently weak and negative correlations with ICRAF MIR measurements across all properties, indicating limited alignment with observed soil variation in the study areas. This makes sense, given the global nature of the SoilGrids models and its coarser spatial resolution. The iSDA measurements for both pH and extractable potassium are poor, correlating slightly negatively with the ICRAF MIR measurements.

This matrix indicates that different soil measurement tools vary in their effectiveness at capturing specific soil properties, and that iSDA generally outperforms SoilGrids in aligning with ICRAF MIR benchmark values. However, correlation measures only the strength and direction of association; it does not reveal systematic bias or provide distributional information. Therefore we next examine a series of scatterplots comparing soil measures from each tool to the ICRAF MIR benchmark. In each, the dashed 45-degree line represents perfect agreement – where ICRAF MIR and the tool provide identical values. Each figure also includes the mean values for each tool (center bottom subplot), along with their 95% confidence intervals.

Soil pH

Figure 6 demonstrates that only the Palintest kit's pH measurements correlate reasonably

with the benchmark ICRAF MIR pH values – though generally still under-estimating low values and over-estimating high values.[10] However, the mean soil pH recorded by the Palintest kit and the two soil maps is around 6, which is fairly close to the benchmark ICRAF MIR mean of 6.16. In fact, the soil maps essentially assign only this mean value to all samples – there is almost no variability in the range of pH values provided by these soil maps. The limited variation in pH values that the soil maps do provide is unrelated to the benchmark values. AgroCares pH values are also uncorrelated with ICRAF MIR values, but with more variation than the soil maps. The mean pH recorded by AgroCares (6.45) is also higher than the ICRAF MIR benchmark. Finally, the Sonkir pH meter substantially overestimates pH and shows rounding issues: most readings cluster at 7, 7.5, or 8, and its mean pH is more than one unit higher than the ICRAF MIR mean.

It is not surprising that the Palintest predicts ICRAF MIR-based pH values well, as both the kit and the wet chemistry used to calibrate the ICRAF spectroscopy model mix soil with deionized water – a standard "water extraction" – and then use a pH electrode to measure hydrogen ion activity. The ICRAF solution uses a soil:water weight ratio of 1:2 while the Palintest uses a ratio between 1:3 and 1:5.[11] The AgroCares Scanner is calibrated on pH measured in a standard water extraction solution, but with a soil:water ratio of 1:5.[12] This lower soil:water ratio has lower ionic strength, which reduces exchange reaction and is therefore expected to result in higher measured pH levels than under a 1:2 soil:water ratio.

This may explain why the AgroCares pH values are higher, on average, than the ICRAF MIR (and Palintest) pH values, but not why the AgroCares pH values are essentially uncorrelated with the ICRAF MIR values. This lack of correlation may stem from differences in spectroscopy training data. It could also relate to differences in soil processing methodology; while CIFOR-ICRAF technicians dried and ground samples prior to scanning, the AgroCares Scanner is applied to unprocessed soil samples in the field with varying soil moisture levels and particle size.

While the manual for the Sonkir pH meter provides little information about its low-cost pH meter, two points of difference are clear. First, the meter is inserted directly into the soil, rather than into a soil-water slurry as with all other tests. Second, the sensor

[10] This regression to the mean style error is common for many types of prediction.

[11] The Palintest manual instructs users to "Take two level 2ml scoops of soil and place in the sample container" and then "Add deionised water to the 10ml mark", which would imply a ratio of 1:3. However, elsewhere in the same manual, it is indicated that Palintest methods for soil pH and conductivity are based on a 5:1 extraction using Deionised water only.

[12] The AgroCares Scanner actually calibrates on two pH measures, one measured in a water solution and another in a potassium chloride solution, both at a soil:solution ratio of 1:5. This is referred to as the NEN-ISO 10390 method. However, we work with results from the water solution as those are the only results provided to users of the Scanner.

measuring electric potential is metallic rather than the more conventional glass or gel electrodes used by the other tests. Both factors make the sensor much less reliable than conventional methods, as we observe in our data. Additionally, the observed rounding issues are likely, at least in part, a function of the analog interface and the interviewer's reading and recording of the pH value.

Finally, both the SoilGrids and iSDA soil maps exhibit systematic smoothing of pH values toward a likely regional average (≈ 6), offering no reflection of the small-scale variability found in our sample. This likely arises from their reliance on large-scale datasets and statistical models that prioritize minimizing mean squared error at a continental scale over capturing fine-scale heterogeneity. Their performance might improve if the regional variation in our dataset were greater.

Clay Content

Figure 7 presents scatterplots and mean values for clay content across the different soil measurement methods. The AgroCares measurements align closely with the benchmark ICRAF MIR measurements, as indicated by a trend line at almost precisely 45 degrees and a mean value of 46.3, which is nearly identical to the ICRAF MIR mean value of 45.8. In contrast, Jabba significantly underestimates clay content, frequently reporting values of zero even when the benchmark range varies from about 10 to 70 percent. The iSDA map performs slightly better than the SoilGrids map in terms of correlation with ICRAF MIR measures, but it substantially underestimates clay content, resulting in a mean value of only 15.2. The mean clay content provided by SoilGrids, 39.1, is closer to the ICRAF MIR mean, but the two measures are entirely uncorrelated. Both maps provide limited variation in predicted clay content.

The ICRAF spectroscopy model for clay content was calibrated using measurements obtained via laser diffraction particle size analysis (LDPSA) in wet mode.[13] The AgroCares spectroscopy model was also calibrated using clay measurements from laser diffraction, which may explain its close alignment with ICRAF MIR values.

The Palintest kit does not measure soil texture, but the Jabba team did so in their laboratory using the Bouyoucos hydrometer sedimentation method. This method disperses soil in water and infers particle size fractions from settling rates. However, it is more prone to procedural error because it omits the ultrasonication step in the ICRAF protocol that ensures complete dispersion of fine particles, and because it relies on indirect hydrometer density readings rather than direct gravimetric measurement. In the Jabba data, clay

[13] In this method, dried and sieved soil is dispersed in water using sodium hexametaphosphate (Calgon), then passed through a laser diffraction analyzer (Horiba LA-950). As laser light passes through the suspension, particles scatter light at angles related to their size – smaller particles (e.g., clay) scatter at wider angles, while larger particles (e.g., sand) scatter more narrowly. The resulting scattering pattern is used to estimate clay, silt, and sand content.

content is reported as zero for 19% of samples and shows no correlation with ICRAF MIR estimates. This pattern suggests systematic dispersion failure: if clay aggregates are not fully broken apart, they settle with sand and silt, leaving little in suspension at the two-hour reading and producing artificially low values. Such a failure would essentially erase true variation in measured clay content.

As with pH, the iSDA and SoilGrids maps provide a narrow range of clay content estimates, reflecting statistical models that prioritize minimizing mean squared error at a continental scale. It is not clear why the iSDA clay measures dramatically underestimate the benchmark clay content, despite performing slightly better than the SoilGrids measures in terms of benchmark correlation.

Soil Organic Carbon

Figure 8 presents scatterplots and mean values for soil organic carbon content as measured by the AgroCares Scanner and the two soil maps. As with clay content, the AgroCares measurements align fairly closely with the benchmark ICRAF MIR measurements, though with a weaker correlation (correlation coefficient of 0.45). Also as with clay content, the soil maps each provide narrow distributions for soil organic carbon content, with iSDA values weakly correlated with ICRAF MIR values (correlation coefficient of 0.37) and SoilGrids values essentially uncorrelated with ICRAF MIR values (correlation coefficient of -0.06). In this case, however, SoilGrids mean values are furthest from the benchmark ICRAF MIR values. ICRAF MIR values are centered around 1.21, while SoilGrids values are centered at double that value, 2.43. iSDA and AgroCares values are both centered at roughly 1.5, overestimating mean benchmark values only slightly.

The AgroCares near-infrared model was trained on soil organic carbon measured by dry combustion, in which acid-pretreated soils are burned at high temperature and CO_2 release is quantified.[14] This method is the modern reference standard in soil science. The ICRAF spectroscopy model was calibrated using soil organic carbon measured via dry combustion in a high-temperature elemental analyzer (Flash2000). In this method, carbonates are first removed by treating the soil with aqueous acid, and the remaining organic material is combusted in tin capsules at temperatures reaching 1800°C. The resulting elemental gases are separated in a chromatographic column and quantified by a thermal conductivity detector. This approach yields highly accurate estimates of total organic carbon and nitrogen. However, given historical constraints, earlier datasets may also include values derived from wet oxidation, which oxidizes organic carbon using potassium dichromate in sulfuric acid. Although widely used in African soil laboratories for its lower cost, wet oxidation typically recovers only 75-80% of total organic carbon unless corrected (Enang et al. 2018; McCarty et al. 2010). The moderate correlation

[14] Acid pre-treatment eliminates inorganic carbon in the soil.

between AgroCares' predictions and ICRAF's benchmark values may partly reflect this mixed-method calibration.[15]

The global scope of the SoilGrids model may contribute to its overestimation of soil organic carbon in our context. The benchmark mean SOC in our sample (1.21%) is low but typical of intensively cultivated smallholder soils in East Africa (often 1-2%; Vågen et al. 2016; Kiboi et al. 2017) and lies well below average values for temperate croplands (2-4%) or undisturbed grasslands and forests (>5%). A focus on minimizing global error may bias predictions upward in regions with genuinely low soil organic carbon content.

Soil Nitrogen

Figure 9 presents scatterplots and mean values for nitrogen, which closely mirror the patterns observed for soil organic carbon. This is expected, since the two soil characteristics are highly correlated. The AgroCares and ICRAF MIR measures of total nitrogen content are fairly correlated (correlation coefficient of 0.43), as expected since both models were calibrated on total nitrogen measured via combustion.[16] As with soil organic carbon, SoilGrids overestimates total nitrogen – averaging twice the ICRAF benchmark – and shows no correlation with ICRAF MIR values, likely reflecting its globally calibrated model. The iSDA and AgroCares mean values are slightly higher, at 1.20 and 1.26 respectively (compared to the benchmark ICRAF MIR mean of 1.06), but only AgroCares shows potential for capturing local variation.

Extractable/Exchangeable Potassium

Finally, Figure 10 presents scatterplots and mean values for soil potassium, measured by ICRAF MIR, AgroCares, the Palintest kit, and iSDA. The Palintest kit measures align most closely with the benchmark ICRAF MIR measures, though the correlation remains moderate (correlation coefficient of 0.32). The correlation is driven down by a set of observations with very high potassium levels (over 600 mg/kg) according to ICRAF MIR, which the kit fails to capture. AgroCares Scanner potassium measures are uncorrelated with ICRAF MIR measures (correlation coefficient of 0.01) and have a much lower mean value (102 mg/kg, whereas the ICRAF MIR and Jabba values are 252 and 225, respectively). The iSDA map values are slightly negatively correlated with the ICRAF MIR benchmark (correlation coefficient of -0.10), again with very limited variation.

The ICRAF MIR, AgroCares, and Palintest measurements differ in the soil potassium

[15] Differences in combustion methods, sample handling, and laboratory protocols can also introduce systematic variation that lowers inter-method correlation.

[16] The only notable difference in their wet chemistry protocols is the analyzer used to combust soil samples at high temperature and quantify the resulting nitrogen gas: AgroCares used an Elementar Vario MAX CN Cube and CIFOR-ICRAF a Thermo Scientific Flash 2000 CN. This difference in hardware should not affect their estimated total nitrogen or carbon content values.

pools they capture. The ICRAF mid-infrared model was calibrated on potassium measured via the Mehlich-3 extraction method, which estimates "extractable" K – encompassing exchangeable and water-soluble K, plus some non-exchangeable forms released by the acidic/chelating extractant. In contrast, the AgroCares near-infrared model was calibrated on potassium measured with hexamminecobalt trichloride, which quantifies only strictly exchangeable K held on the soil's cation exchange complex. The Palintest kit uses a 0.1 molar magnesium acetate extractant, which primarily measures exchangeable K but, because it is a milder aqueous extraction, also captures most of the water-soluble fraction. As a result, the Palintest measure represents a somewhat larger pool of K than AgroCares – though still narrower than the broader extractable K pool measured by ICRAF. This ordering is reflected in our data, as ICRAF reports the largest potassium pools and AgroCares the smallest. Notably, the iSDA map was also trained on Mehlich-3-extracted potassium (Hengl et al. 2021), yet its mean extractable potassium value is almost as low as the strictly exchangeable AgroCares value.

3.3 Classification Analysis

While differences in point estimates across soil measurement methods are expected to some degree, and are present as described in the sections above, what ultimately matters for policies, programs, and farmers is whether those differences cross decision-relevant thresholds that change recommended treatments. To assess whether the differences in means and distributions across measurement methods observed in the preceding section imply different interventions, we conduct a classification analysis, grouping measurements into agronomically meaningful categories (e.g., strongly/moderately/neutral/alkaline pH; low/moderate/high clay, soil organic carbon, total nitrogen, and extractable potassium), which are linked to intervention choices such as liming, organic amendments, and fertilizer type and application rate.

By comparing category assignments across methods, we gain insight into whether the divergence in property estimates translates into different actions on the ground. This is important, as misclassification can trigger over-treatment (e.g., unnecessary lime or organic input application) or under-treatment (e.g., persistent acidity or low soil carbon), leading to wasted resources, muted yield gains, and reduced trust in the efficacy and profitability of inputs. Table A3 lists the classification bin intervals for each soil property used in this paper. These classifications are broadly based on established thresholds from the literature and ICRAF's own critical values used in their soil measurements.

We compare soil pH classifications between ICRAF MIR and the Palintest kit, which

yielded the highest correlation among all methods (correlation coefficient of 0.54).[17] The top-left panel of Figure 11 presents the corresponding Sankey plot, revealing considerable misclassification despite this relatively strong correlation. The majority of soil samples were classified as moderately acidic by ICRAF MIR, of which only 53% were correctly categorized by the Palintest kit. Of the remainder, 26% were misclassified as strongly acidic and approximately 21% as neutral or alkaline. Comparable patterns of misclassification are observed across the other pH categories. Overall, approximately 54% of samples were classified in the same soil pH bin based on ICRAF MIR and Palintest kit results.

For clay content, we compare ICRAF MIR to AgroCares, which yielded the highest correlation among all methods (correlation coefficient of 0.68; Figure 11, top-right). The majority of samples were classified as moderate by ICRAF MIR, a categorization that AgroCares reproduced in 72% of cases, though approximately 9% were misclassified as low or very low and 19% as high. Among samples classified as high clay by ICRAF MIR, only 54% were correctly categorized by AgroCares, with the remainder misclassified as moderate. More notably, only 47% of samples with low clay content according to ICRAF MIR were correctly identified by AgroCares, while 25% were misclassified as moderate and 28% as very low. In total, 68% of samples were assigned the same clay content bin based on results from ICRAF MIR and AgroCares.

For soil organic carbon, we again compare ICRAF MIR to AgroCares, which showed closer agreement with ICRAF MIR than any of the soil maps (correlation coefficient of 0.44; Figure 11, bottom-left). Only 47% of samples classified as low by ICRAF MIR were correctly categorized by AgroCares, with the remaining 53% misclassified into higher categories. Similarly, among samples with moderate soil organic carbon according to ICRAF MIR, only 55% were correctly categorized, while 17% were misclassified as low and 28% as high.[18] In the sample overall, 53% of samples were classified in the same soil organic carbon bin based on ICRAF MIR and AgroCares results.

Finally, for extractable potassium, we compare ICRAF MIR to the Palintest kit, the method that yielded the strongest positive correlation (correlation coefficient of 0.32; Figure 11, bottom-right). The misclassification pattern is notably symmetric: approximately 62% of samples classified as low or moderate by ICRAF MIR were correctly categorized by the Palintest kit, while roughly 38% were misclassified into the adjacent category in each case. Most strikingly, none of the samples classified as high extractable

[17] We primarily focus on comparing classifications between ICRAF MIR and the soil test that yielded the highest correlation. Complete comparisons across all soil tests for pH (Figure A2), clay (Figure A3), soil organic carbon (Figure A4), total nitrogen (Figure A5), and extractable potassium (Figure A6) are provided in the appendix.

[18] We omit a misclassification figure for total nitrogen, as the majority of samples fell within the high category according to ICRAF MIR, a classification that AgroCares largely reproduced.

potassium by ICRAF MIR were identified as such by the Palintest kit. Overall, approximately 60% of samples were classified in the same extractable potassium bin based on results from ICRAF MIR and the Palintest kit.

Taken together, these classification results demonstrate that even the best-performing soil testing tools exhibit substantial misclassification rates when measurements are grouped into agronomically relevant categories. If these tools were used to guide soil management decisions, the observed misclassification rates would lead to incorrect intervention recommendations for a considerable share of plots – whether through unnecessary amendments where soil conditions are adequate or insufficient treatment where deficiencies are present.

3.4 Implementation Considerations

Given our goal to test and validate soil testing methods for national survey use, evaluating data quality alone is not sufficient. Scalability, which can be considered to include both financial and implementation barriers, is a critical consideration. Even the most accurate soil testing method is unlikely to be adopted if it is not scalable within the target context.

Equipment requirements and costs vary by method. Table 1 summarizes the costs and additional considerations for all the soil testing tools. At CIFOR-ICRAF laboratories, spectral analysis costs were approximately USD 8 per sample. With additional costs incurred for calibration tests, costs were approximately USD 12 per sample, excluding costs associated with shipping of samples, permits for sample export, supplies, indirect costs, overhead, and staff time.[19] For analysis with the Palintest kit, Jabba Engineering conducted the testing using their own Palintest kits, with costs of about USD 27 per sample for the full sample set, plus an additional USD 22 per sample for extra parameters (S, Al, NH_4, Cu, Fe, Mn), which were measured on half of the samples, again excluding costs related to overhead and staff time. If one were to purchase the SWK500 Complete Soil Management Kit directly, it costs approximately USD 2,500, excluding reagents, consumables, and personnel costs for conducting and interpreting the tests. The AgroCares approach required the highest equipment investment, with each scanner costing approximately USD 7,000 and an annual license ranging from USD 1,800 to 2,600, depending on the selected features. For the Uganda CLASS sample, the average cost per sample was about USD 41. Most costs associated with the AgroCares method are fixed within a given license period, meaning the per-sample cost could decrease substantially with a larger sample size. In contrast, costs for the CIFOR-ICRAF and Jabba methods, which

[19] Calibration tests, which were conducted on a subset of soil samples in order to update the spectroscopy models, included wet chemistry analysis at USD 63 per sample (for 90 samples) and carbon and nitrogen testing at USD 35 per sample (for 45 samples). Note that unit cost estimates are based on the price list published by CIFOR-ICRAF (link).

offered lower per sample costs in the context of this study (estimates of which exclude personnel and overhead costs which are not applicable to the AgroCares model), are largely variable, offering limited opportunities to reduce per-sample costs. The Sonkir pH meter was purchased at approximately USD 16 per device, for a per sample cost of roughly USD 0.09.

Equipment costs also influence fieldwork design. Methods requiring less expensive equipment are more easily implemented by a larger number of enumerators. For example, if desired, all enumerators could be trained to collect soil samples for laboratory analysis, as the required tools (buckets, poles, a mixing trowel, and a soil auger – the most expensive item) are relatively low cost. Costlier tools such as the AgroCares scanner, on the other hand, limit the number of enumerators that can be equipped, requiring adjustments to fieldwork logistics. In the Uganda CLASS study, one enumerator per team was designated as the "soil enumerator" and given an AgroCares scanner. This person conducted all soil collection and scanning for the team, and therefore required additional mobility support, either through access to a vehicle or a transportation allowance. Alternative fieldwork structures could be explored, such as enumerators collecting and labeling soil samples and bringing them to a central point for scanning by a team member equipped with a scanner, though care would need to be taken to ensure the geographic coordinates of the relevant point were recorded in the AgroCares app before leaving the field.

Beyond fieldwork design choices driven by equipment costs, the complexity of sample collection and handling tied to each method needs to be carefully considered. For the low-cost pH probe, no soil sample collection was needed, which minimized effort, but the resulting data contained no valid information. For the other methods – ICRAF, AgroCares, and Jabba – soil sample collection followed the same procedure and thus had the same requirements. Where these methods differ is in sample treatment after collection. AgroCares scanned the collected topsoil in the field using its handheld scanner and mobile application. This process typically took five minutes or less and posed few technical challenges. Enumerators manually entered identifiers such as household and plot ID into the app to link results with survey data; GPS coordinates were also stored, providing a useful means of validating these linkages. In contrast, the ICRAF and Jabba methods required enumerators to bag and label soil samples in the field and transport them to the respective laboratories for drying, processing, and analysis. To ensure accurate linkage between samples and survey data, enumerators used barcode labels affixed to sample bags, which they also scanned into the CAPI-based questionnaire. In this study, all sample processing (e.g., drying and grinding) occurred at the laboratory. However, in previous operations, drying was done in the field before transport, which proved logistically complex and burdensome.

The timeline for receiving soil testing results is another important consideration. There

are clear trade-offs: AgroCares can provide results in near real-time (assuming the device has network connection), while lab-based methods require additional time for sample delivery and processing. In the Uganda CLASS study, soil collection concluded in early November 2023. Palintest results from Jabba were available by January 2024, whereas CIFOR-ICRAF results became available only in July 2024. This delay was partly due to the decision to store samples for CIFOR-ICRAF and deliver them to Nairobi in a single batch after all collection was completed, which was a simpler and more cost-effective approach. In contrast, because Jabba's laboratories were located in the study areas, samples could be delivered continuously, allowing soil testing to proceed in parallel with survey data collection.

Overall, the AgroCares, ICRAF, and Jabba approaches required the same level of effort in terms of sample collection, while sample handling logistics were more complex for the ICRAF and Jabba approaches given the need to bag, label, and deliver samples to the respective laboratories. AgroCares offered the fastest turnaround time for results, followed by the Jabba and then ICRAF approaches. From a cost perspective, CIFOR-ICRAF analysis offered the lowest per sample cost (excluding expenses for staff and overhead), followed by Palintest (implemented by Jabba) and then AgroCares, with the AgroCares approach largely made up of fixed costs allowing for significantly reduced per sample costs when larger sample sizes are needed. The low cost Sonkir pH meter required the lowest investment both in terms of cost and time, but given the limited and unreliable results, this is not considered a feasible option for adoption.

3.5 Policy Implications

There is ongoing debate about how best to collect accurate and cost-effective soil information in Sub-Saharan Africa, where livelihoods remain deeply tied to soil health. Because healthy soils are linked not only to higher yields and household income but also to greater resilience to shocks, growing emphasis has been placed on data-driven approaches to improving soil quality in the region (Abay et al. 2025). Accurate, high-resolution measures of soil properties are instrumental in implementing improved soil fertility management strategies (Hörner and Wollni 2021) and in enabling adequate responses to climate shocks (Ahmed 2025). Our evaluation of the accuracy and scalability of multiple portable soil tests and publicly available soil maps in Uganda carries several implications for survey and program design, extension and advisory services, and farmland information services intended to support more sustainable and successful agricultural productivity.

First, our findings reinforce that self-reported measures of soil quality remain subject to substantial bias. Even the observable characteristics that farmers commonly rely on to judge soil health, such as crop yields and soil color, correspond only weakly to

measured soil properties, introducing considerable noise into how farmers evaluate their soils. We therefore caution against relying on self-reported assessments as the sole basis for characterizing soils, particularly where they inform decisions about soil amendments or other agricultural interventions.

Because self-reported measures offer limited actionable insight, our analysis also speaks to the viability of alternative low-cost, portable soil tests and spatial soil maps. These tools are attractive because they can, in principle, measure soils in the field and at higher frequency than laboratory analysis allows. Yet our results, once again, call for caution. At present, these tools (at least those used in our study) correlate poorly with laboratory measurements, and their accuracy varies considerably by soil property. Even inexpensive devices designed to measure a single attribute, such as the Sonkir pH meter, fail to assess that attribute reliably. The same limitation applies to widely used soil maps, which lack the spatial resolution to capture the substantial variation in soil properties across plots and are therefore likely inappropriate tools to guide small-scale, farm-level decisions.

Accordingly, we advocate careful and deliberate use of these tools, paired wherever possible with laboratory-based measurement to validate their accuracy on an ongoing basis. Soil interventions are urgently needed across Sub-Saharan Africa, but realizing their benefits requires directing resources correctly and avoiding the costly, ineffective outcomes that misguided recommendations can produce. For soil health initiatives adopting these low-cost, increasingly market-available tools (see, for example, Nyondo et al. 2025, who uses a handheld reflectometer in Malawi, Kosmowski, Abebe, and Ozkan 2020, who utilizes a miniaturized NIR spectrometers, and Amuri et al. 2024, who uses a portable soil kit in Tanzania), we recommend validating a subset of samples against an accredited laboratory, so that implementers can appreciate the degree of bias or misclassification inherent in their chosen instruments and interpret results accordingly.

4 Conclusion

Drawing on soil samples from five districts in Uganda, we compare soil properties measured by three soil test tools and two spatial soil maps against benchmark measurements from CIFOR-ICRAF for five key indicators. We first find that farmers' self-reported assessments of soil quality align poorly with measured soil properties, corresponding only weakly even to the observable characteristics, such as crop yields and soil color, that farmers cite as the basis for those assessments. We find that the AgroCares Scanner – a portable near-infrared spectroscopy-based device – performs reasonably well for clay content, fairly well for soil organic carbon and total nitrogen, but poorly for pH and

potassium.[20] The Jabba Palintest kit performs reasonably for pH, though it also underestimates low values and overestimates high ones. The low-cost Sonkir pH meter performs poorly and cannot be recommended for reliable pH assessment. Many of these differences reflect the diverse methodologies and calibration protocols used by each tool, highlighting the importance of understanding underlying measurement processes when interpreting soil test results.

Turning to spatial soil maps, we find that both iSDA and SoilGrids offer little within-region variability and fail to capture the sample heterogeneity observed in our benchmark data. These findings echo results from prior studies: spatial soil maps often prioritize minimizing mean squared error over large regions, leading to smoother, less variable predictions that miss local differences. While iSDA performs slightly better than SoilGrids in most benchmark comparisons, both maps fall short in capturing variation at the scale needed for plot-level, farm-level, or even sub-county-level decision-making.

Beyond accuracy, scalability and logistics are essential considerations for national soil monitoring initiatives. Laboratory-based methods like those employed by CIFOR-ICRAF can require relatively high costs and/or longer turnaround time for results – linked to sample collection, transport, and processing requirements, that can limit their utility for timely decision-making, despite offering high quality data. Although the AgroCares Scanner enables rapid, in-field testing, its relatively high equipment costs and licensing fees pose scalability constraints. The Sonkir meter is inexpensive and easy to use, but its lack of reliability excludes it from serious consideration. These trade-offs indicate that no single method is ideal across all dimensions and for all purposes. Policymakers and practitioners must weigh accuracy, cost, and implementation feasibility when designing soil data strategies that are both robust in results and scalable.

[20] AgroCares measures exchangeable rather than extractable potassium, which complicates direct comparison.

Tables and Figures

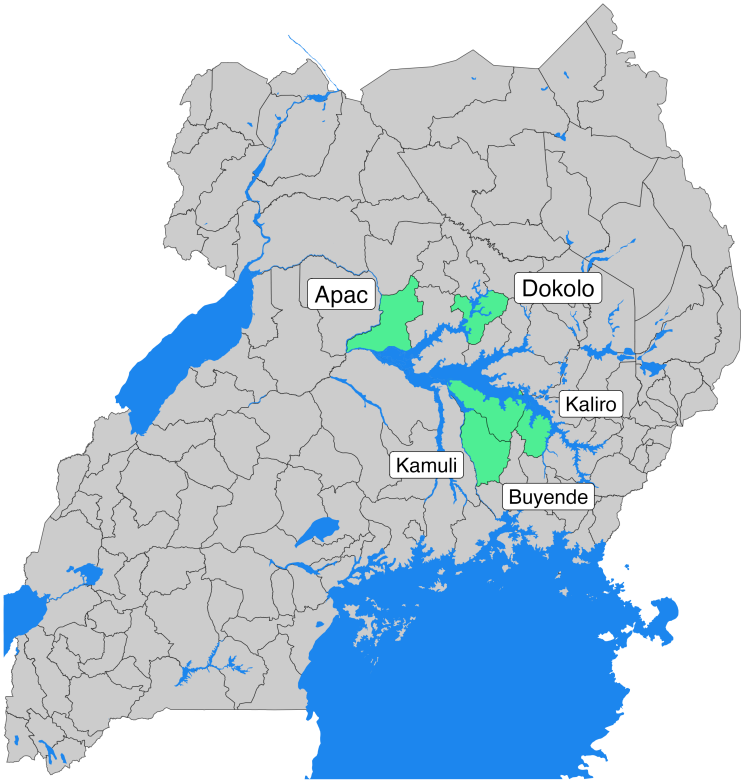


Figure 1: Sample of districts in Uganda. [back]

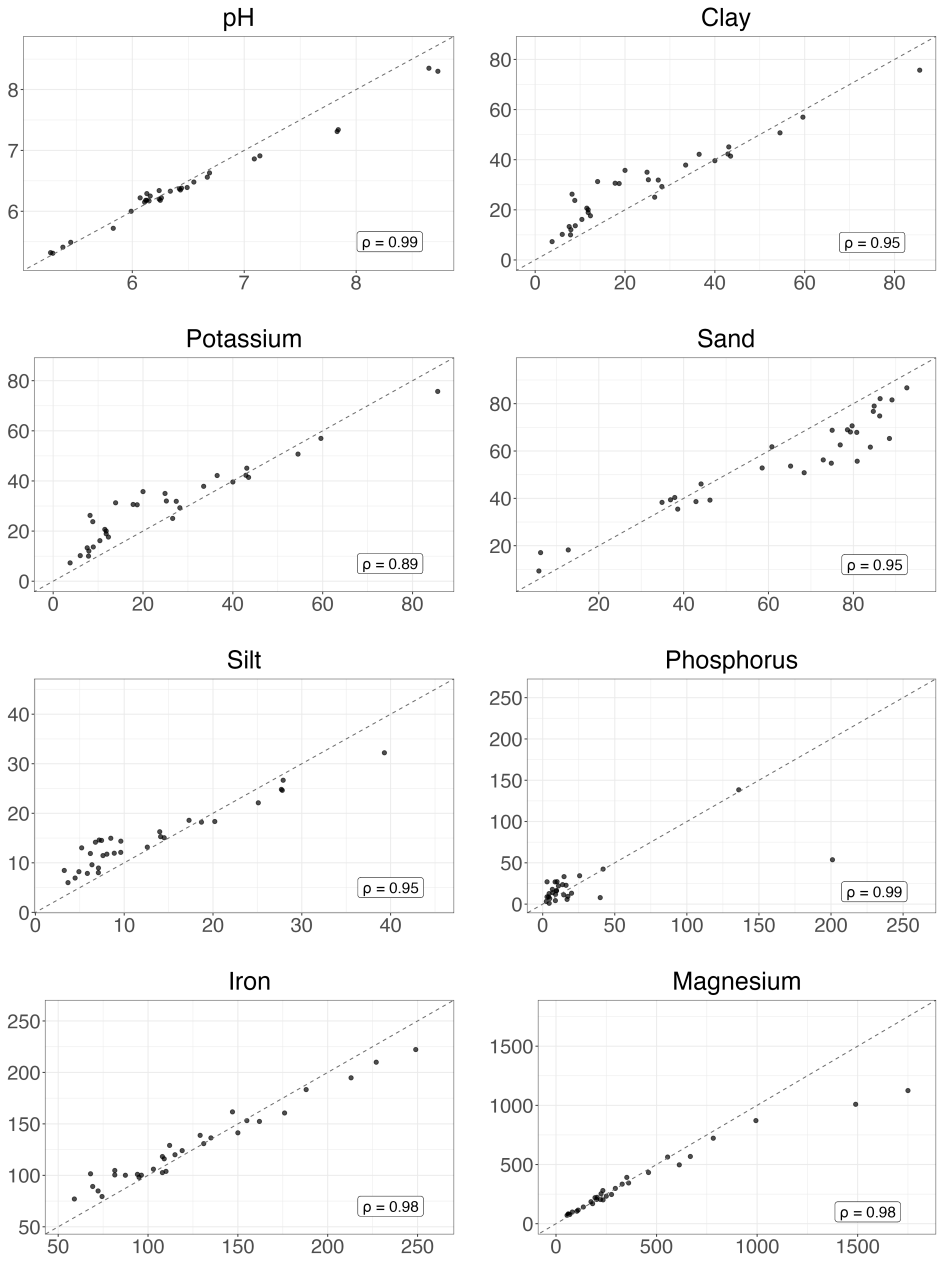


Figure 2: Scatterplot comparing values obtained using wet chemistry approaches (x-axis) and ICRAF MIR spectroscopy(y-axis). [back]

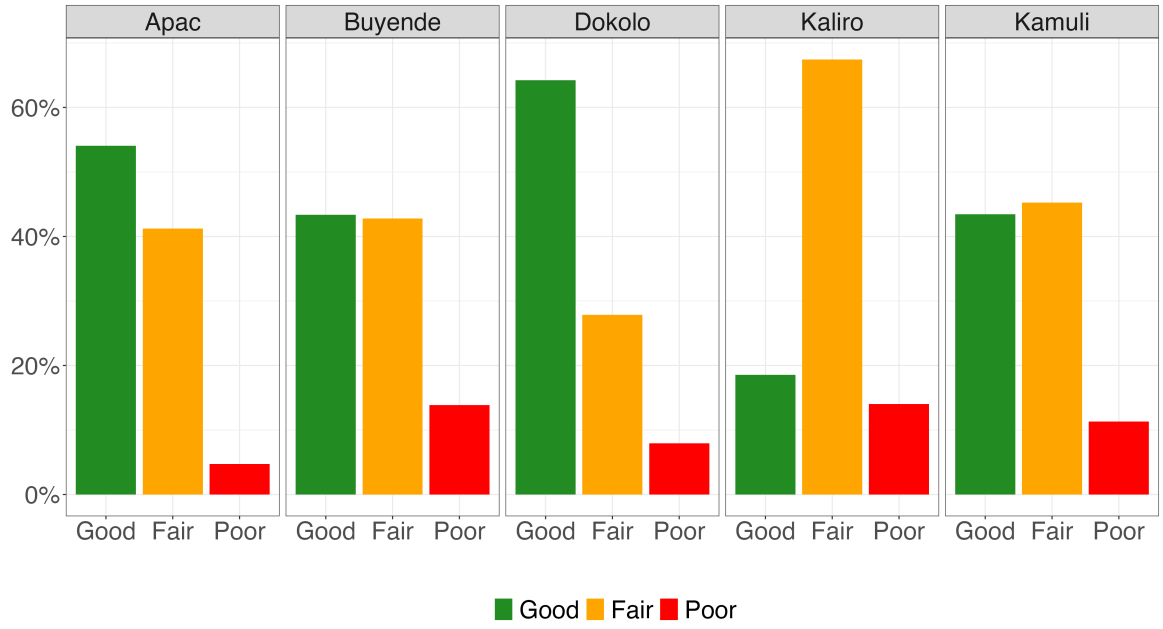
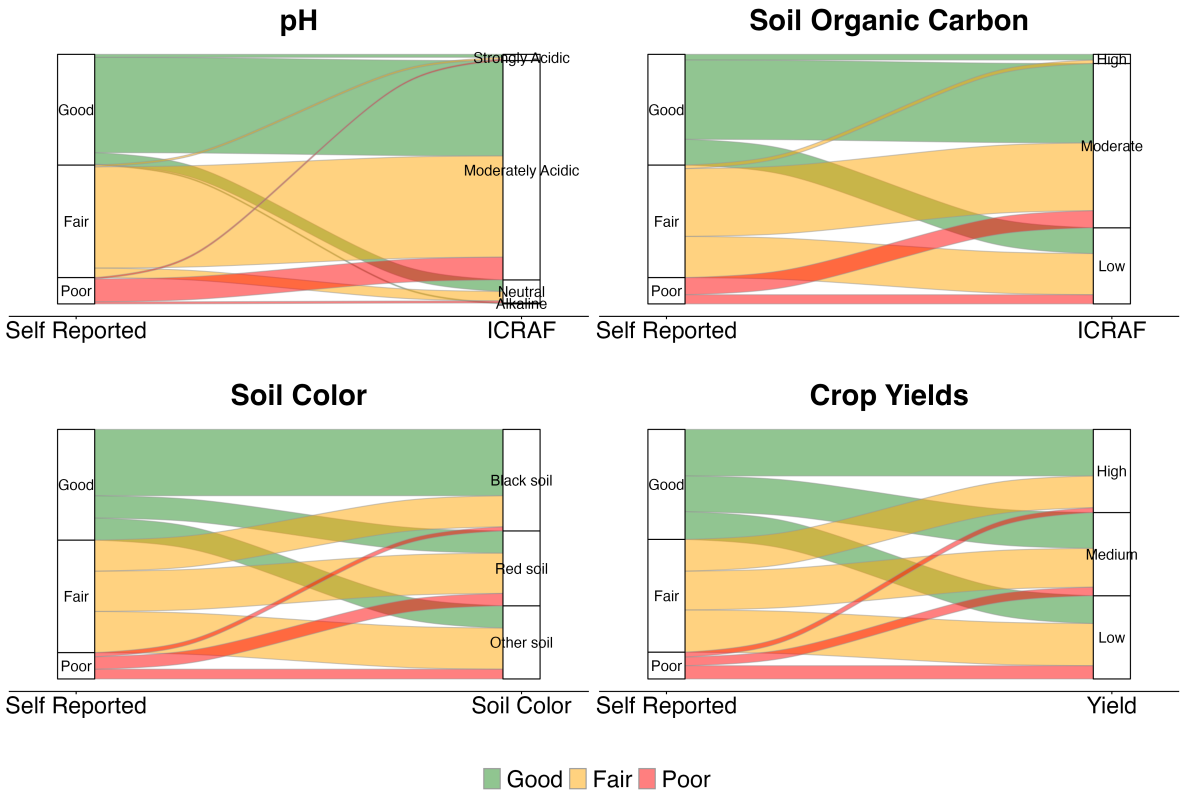
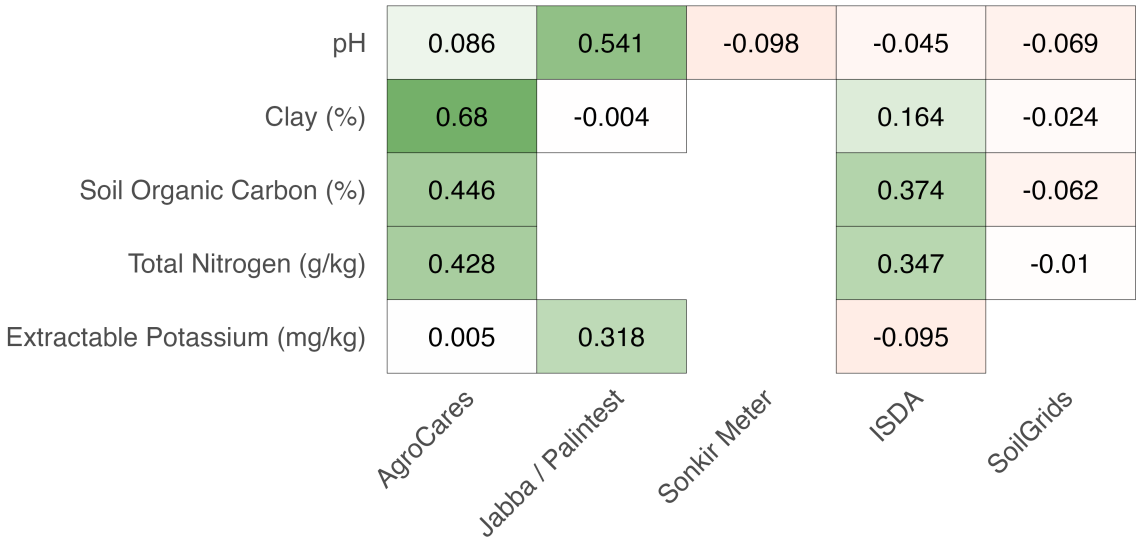


Figure 3: Proportion of self-reported soil quality across districts. [back]



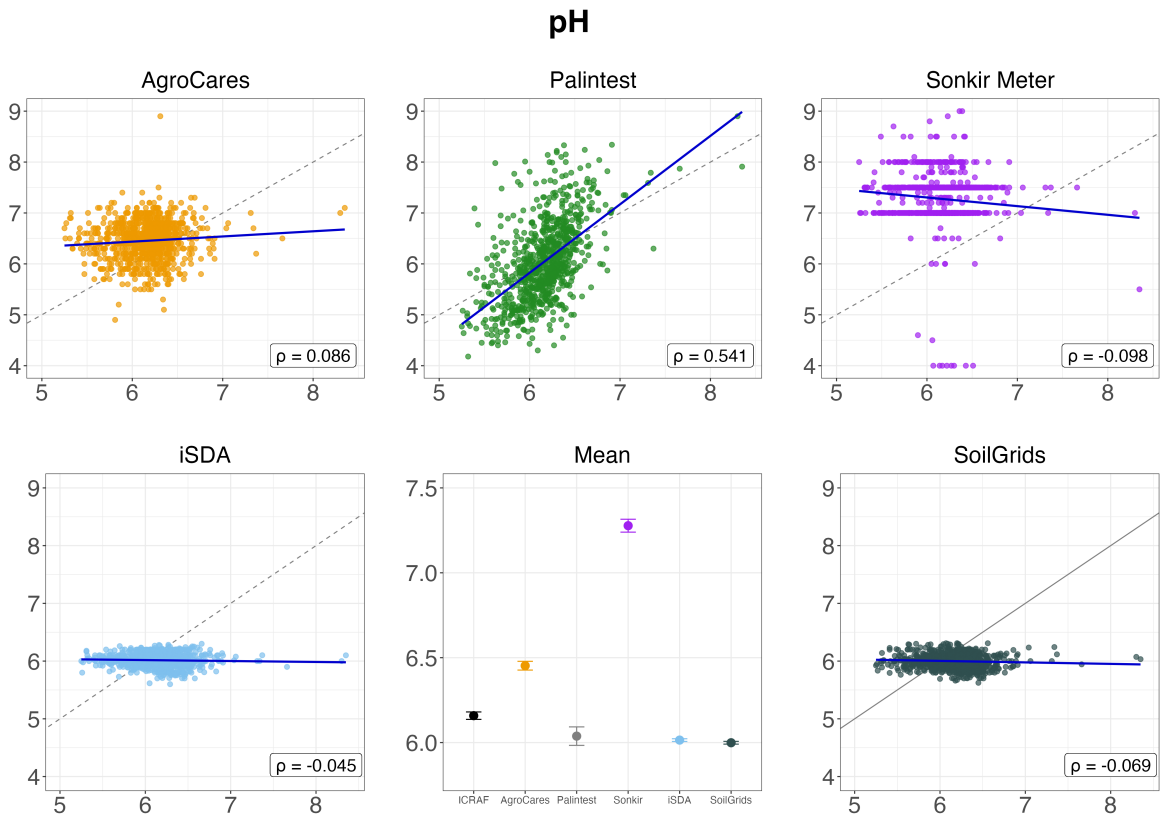
Note: The categories for pH are: pH <= 5.5 (Strongly Acidic), pH > 5.5 and pH <= 6.5 (Moderately Acidic), and pH >= 6.5 (Neutral / Alkaline). The categories for Soil Organic Carbon are: SOC <= 1.2 (Low), SOC > 1.2 and SOC <= 1.7 (Moderate), SOC > 1.7 (High). The categories for Crop Yields are based on terciles.

Figure 4: Self reported soil quality across different criteria. [back]



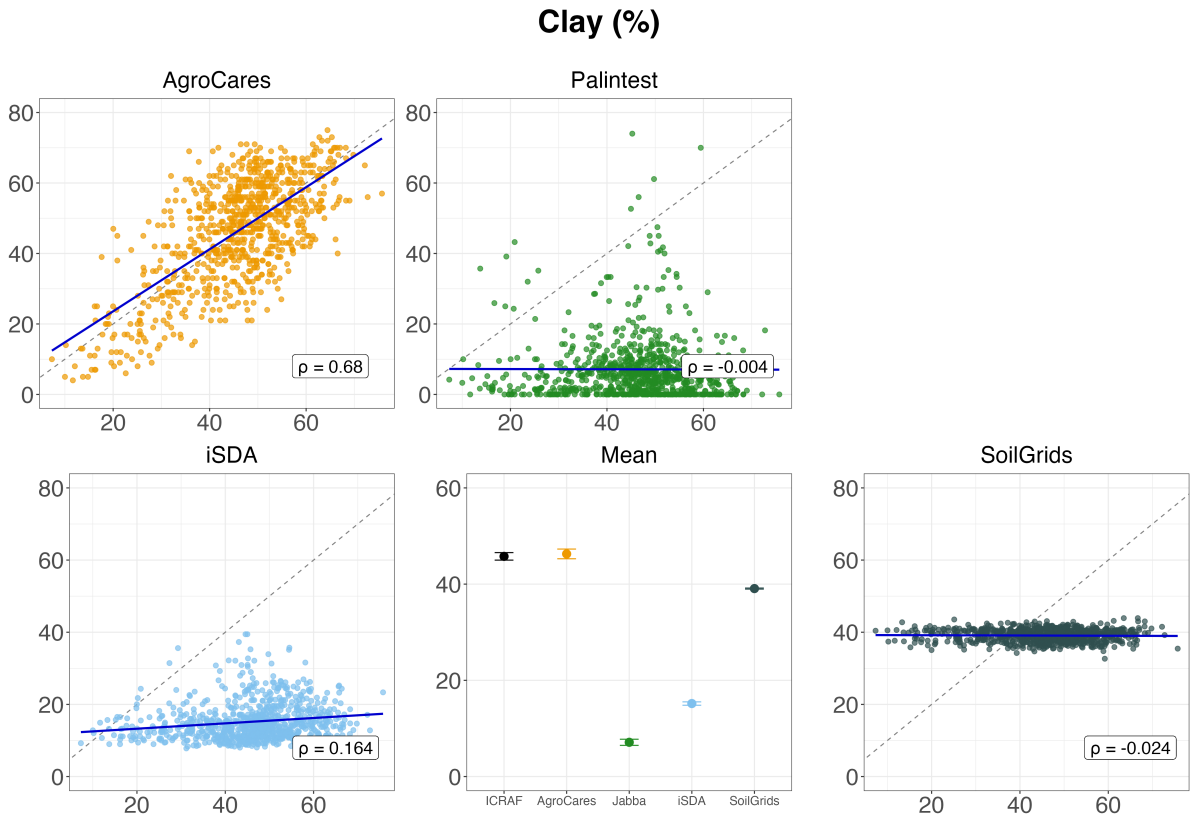
Note: Jabba measured clay content through their regional laboratories and measured pH and extractable potassium using the Palintest kit.

Figure 5: Correlation of soil tests and maps with ICRAF. [back]



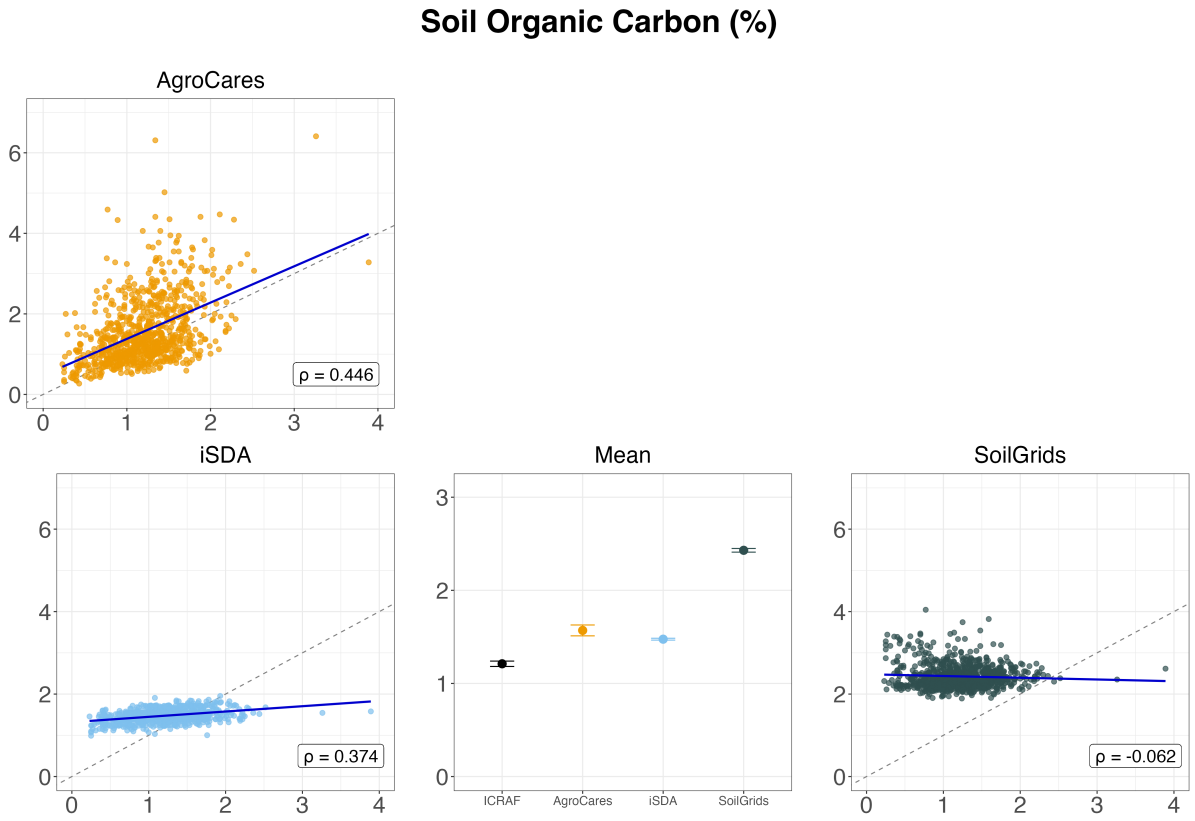
Note: In each scatter plot, the x-axis represents the ICRAF values, while the y-axis corresponds to the values from the respective tool.
The errorbars in the Mean plot represents the 95% confidence interval.

Figure 6: Correlations and mean values of various soil tests and maps compared to the ICRAF MIR pH values. [back]



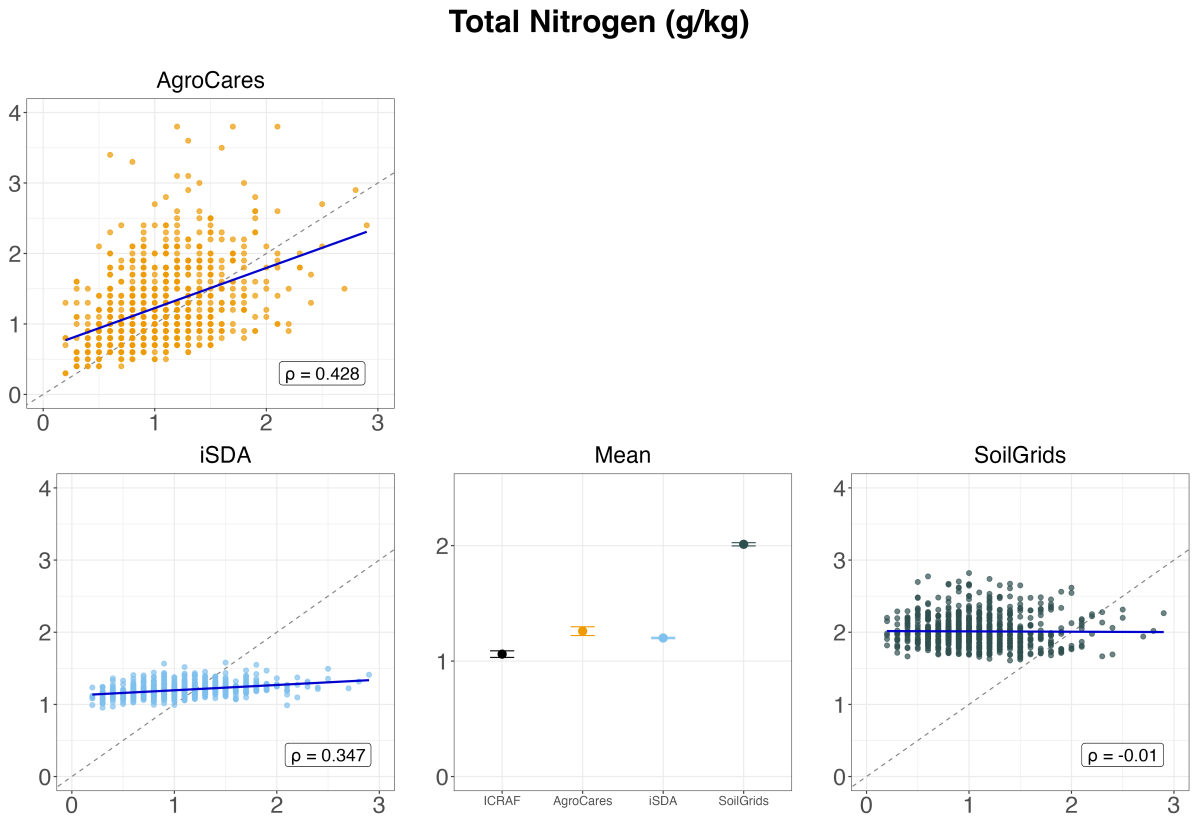
Note: In each scatter plot, the x-axis represents the ICRAF values, while the y-axis corresponds to the values from the respective tool.
The errorbars in the Mean plot represents the 95% confidence interval.

Figure 7: Correlations and mean values of various soil tests and maps compared to the ICRAF MIR clay values. [back]



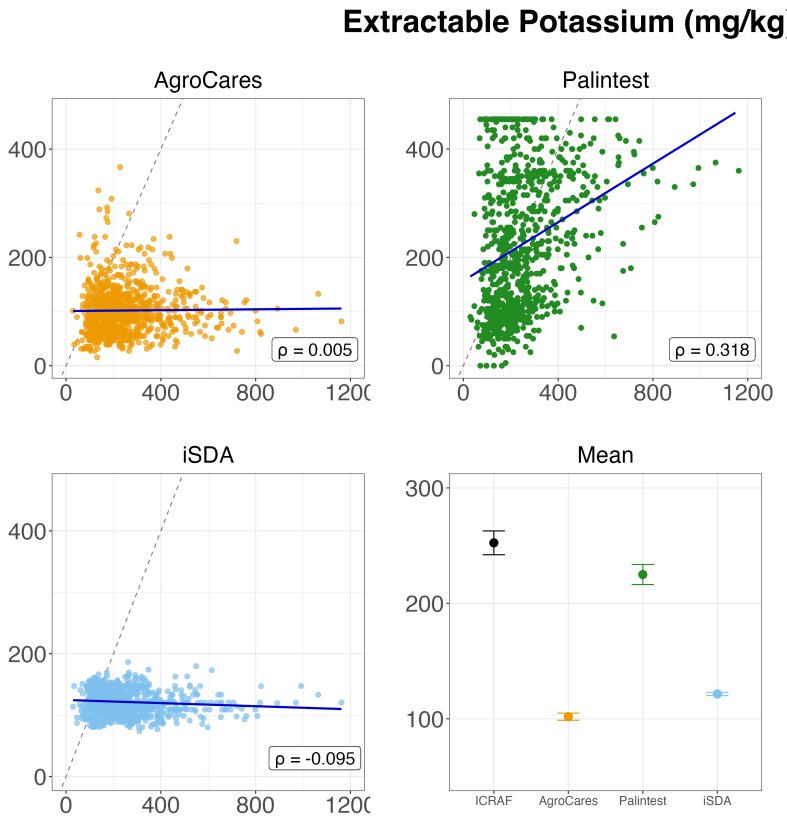
Note: In each scatter plot, the x-axis represents the ICRAF values, while the y-axis corresponds to the values from the respective tool.
The errorbars in the Mean plot represents the 95% confidence interval.

Figure 8: Correlations and mean values of various soil tests and maps compared to the ICRAF MIR soil organic carbon values. [back]



Note: In each scatter plot, the x-axis represents the ICRAF values, while the y-axis corresponds to the values from the respective tool. The errorbars in the Mean plot represents the 95% confidence interval.

Figure 9: Correlations and mean values of various soil tests and maps compared to the ICRAF MIR total nitrogen values. [back]



Note: In each scatter plot, the x-axis represents the ICRAF values, while the y-axis corresponds to the values from the respective tool. The errorbars in the Mean plot represents the 95% confidence interval.

Figure 10: Correlations and mean values of various soil tests and maps compared to the ICRAF MIR extractable potassium values. [back]

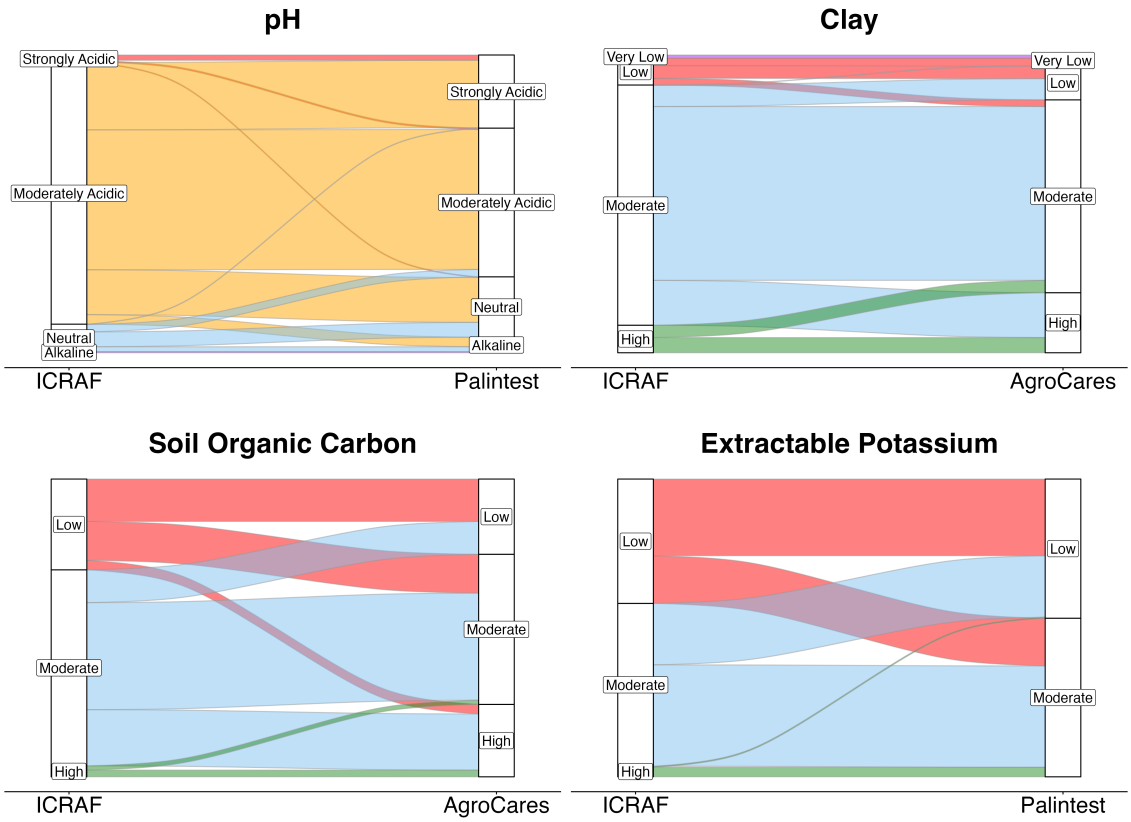


Figure 11: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for pH (top left), clay (top right), soil organic carbon (bottom left), and extractable potassium (bottom right). [back]

Table 1: Comparison of costs and implementation requirements across the four soil testing tools. [back]

	Costs in USD ($)			
	Equipment Cost	**Annual License**	**Cost Structure**	**Estimated Per Sample Cost**
CIFOR-ICRAF Spectroscopy	–	–	Variable	12[a]
AgroCares Scanner	7,000	1,800 – 2,600[b]	Fixed	41[c]
Palintest Kit	2,500[d]	–	Primarily Variable	27[e]
Sonkir pH Meter	16	–	Fixed	0.09
	Soil Sampling & Processing Needs			
	Soil Sample Collection Required?	**Soil Sample Transportation Required?**	**Soil Processing Requirements**	**Turnaround Time for Results**
CIFOR-ICRAF Spectroscopy	Yes	Yes	High	Variable[f]
AgroCares Scanner	Yes	No	Low	Real time
Palintest Kit	Yes	Yes[g]	High	Variable[f]
Sonkir pH Meter	No	No	None	Real time

Overhead, staff, and soil transport costs are not included in the table.
[a] Includes calibration costs. Omitting this, per sample cost would have been $8.
[b] Depends on chosen features.
[c] Given the fixed cost structure, per sample cost could be significantly reduced with a larger sample size than employed in the Uganda CLASS study.
[d] Excludes additional regent and consumable costs.
[e] Instead of purchasing the Palintest Kit, which would follow a more fixed cost structure, we had Jabba Engineering conduct the test using their existing equipment, following a variable cost structure. Extra soil parameters (S, Al, NH_4, Cu, Fe, Mn) required an additional cost of $22 on top of the $27 indicated above.
[f] Timeline for lab-based methods depends on survey design, transportation needs, and laboratory set-up and backlog.
[g] Assuming the Palintest Kit is used in a laboratory setting or other centralized location.

References

Abay, K.A., J. Chamberlin, P. Chivenge, and D.J. Spielman. 2025. "Fertilizer, soil health, and economic shocks: A synthesis of recent evidence." *Food Policy* 133:102892.

Ahmadi, A., M. Emami, A. Daccache, and L. He. 2021. "Soil Properties Prediction for Precision Agriculture Using Visible and Near-Infrared Spectroscopy: A Systematic Review and Meta-Analysis." *Agronomy* 11:433, Number: 3.

Ahmed, M.H. 2025. "Early growing season temperature variation and fertilizer use among smallholder farmers." *Food Policy* 133:102793.

Amuri, N.A., J.M.R. Semoka, R. Weil, M. Mzimbiri, L. Gatere, C. Palm, and P. Sanchez. 2024. "Validating SoilDoc kit for site-specific fertilizer recommendations for maize production in Tanzania." *Agronomy Journal* 116:2641–2656.

Barbier, E.B., and J.P. Hochard. 2018. "Land degradation and poverty." *Nature Sustainability* 1:623–631.

Batjes, N.H., L. Calisto, and L.M. de Sousa. 2024. "Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023)." *Earth System Science Data* 16:4735–4765.

Bellon-Maurel, V., and A. McBratney. 2011. "Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives." *Soil Biology and Biochemistry* 43:1398–1410.

Berazneva, J., A. Maertens, W. Mhango, and H. Michelson. 2023. "Paying for agricultural information in Malawi: The role of soil heterogeneity." *Journal of Development Economics* 165:103144.

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

Bhargava, A.K., T. Vagen, and A. Gassner. 2018. "Breaking Ground: Unearthing the Potential of High-resolution, Remote-sensing Soil Data in Understanding Agricultural Profits and Technology Use in Sub-Saharan Africa." *World Development* 105:352–366.

Dawoe, E.K., J. Quashie-Sam, M.E. Isaac, and S.K. Oppong. 2012. "Exploring farmers' local knowledge and perceptions of soil fertility and management in the Ashanti Region of Ghana." *Geoderma* 179-180:96–103.

Desbiez, A., R. Matthews, B. Tripathi, and J. Ellis-Jones. 2004. "Perceptions and assessment of soil fertility by farmers in the mid-hills of Nepal." *Agriculture, Ecosystems & Environment* 103:191–206.

Enang, R.K., B.P.K. Yerima, G.K. Kome, and E. Van Ranst. 2018. "Assessing the Effectiveness of the Walkley-Black Method for Soil Organic Carbon Determination in Tephra Soils of Cameroon." *Communications in Soil Science and Plant Analysis* 49:2379–2386, _eprint: https://doi.org/10.1080/00103624.2018.1510948.

Gars, J., R. Fishman, A. Kishore, Y. Rothler, and P.S. Ward. 2025. "Confidence and information usage: Evidence from soil testing in India." *American Journal of Agricultural Economics* 107:1406–1437, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12513.

Gourlay, S., E. Aynekulu, K.D. Shepherd, and C. Carletto. 2017. "Collecting the dirt on soils : advancements in plot-level soil testing and implications for agricultural statistics."

Gray, L.C., and P. Morant. 2003. "Reconciling indigenous knowledge with scientific assessment of soil fertility changes in southwestern Burkina Faso." *Geoderma* 111:425–437.

Harou, A.P., M. Madajewicz, H. Michelson, C.A. Palm, N. Amuri, C. Magomba, J.M. Semoka, K. Tschirhart, and R. Weil. 2022. "The joint effects of information and financing constraints on technology adoption: Evidence from a field experiment in rural Tanzania." *Journal of Development Economics* 155:102707.

Heger, M.P., G. Zens, and M. Bangalore. 2020. "Land and poverty: the role of soil fertility and vegetation quality in poverty reduction." *Environment and Development Economics* 25:315–333, Num Pages: 315-333 Place: Cambridge, United Kingdom.

Hengl, T., M.A.E. Miller, J. Križan, K.D. Shepherd, A. Sila, M. Kilibarda, O. Antonijević, L. Glušica, A. Dobermann, S.M. Haefele, S.P. McGrath, G.E. Acquah, J. Collinson, L. Parente, M. Sheykhmousa, K. Saito, J.M. Johnson, J. Chamberlin, F.B.T. Silatsa, M. Yemefack, J. Wendt, R.A. MacMillan, I. Wheeler, and J. Crouch. 2021. "African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning." *Scientific Reports* 11:6130.

Hörner, D., and M. Wollni. 2021. "Integrated soil fertility management and household welfare in Ethiopia." *Food Policy* 100:102022.

Kiboi, M.N., K.F. Ngetich, J. Diels, M. Mucheru-Muna, J. Mugwe, and D.N. Mugendi. 2017. "Minimum tillage, tied ridging and mulching for better maize yield and yield stability in the Central Highlands of Kenya." *Soil and Tillage Research* 170:157–166.

Kihara, J., G. Nziguheba, S. Zingore, A. Coulibaly, A. Esilaba, V. Kabambe, S. Njoroge, C. Palm, and J. Huising. 2016. "Understanding variability in crop response to fertilizer and amendments in sub-Saharan Africa." *Agriculture, Ecosystems & Environment* 229:1–12.

Kok, M., S. Sarjant, S. Verweij, S.F.C. Vaessen, and G.H. Ros. 2024. "On-site soil analysis: A novel approach combining NIR spectroscopy, remote sensing and deep learning." *Geoderma* 446:116903.

Kosmowski, F., A. Abebe, and D. Ozkan. 2020. "Challenges and lessons for measuring soil metrics in household surveys." *Geoderma* 375:114500.

Marenya, P., C.B. Barrett, and T. Gulick. 2008. "Farmers' Perceptions of Soil Fertility and Fertilizer Yield Response in Kenya." *SSRN Scholarly Paper ID 1845546*, pp. .

McCarty, G.W., J.B.R. Iii, R. Yost, P.C. Doraiswamy, and M. Doumbia. 2010. "Evaluation of methods for measuring soil organic carbon in West African soils." *African Journal of Agricultural Research* 5:2169–2177.

McCarty, G.W., J.B. Reeves, V.B. Reeves, R.F. Follett, and J.M. Kimble. 2002. "Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement." *Soil Science Society of America Journal* 66:640–646, _eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2136/sssaj2002.6400a.

Michelson, H. 2025. "Navigating the Measurement Frontier: New Insights Into Small Farm Realities." *Agricultural Economics* 56:526–542, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/agec.70013.

Michelson, H., A. Fairbairn, B. Ellison, A. Maertens, and V. Manyong. 2021. "Misperceived quality: Fertilizer in Tanzania." *Journal of Development Economics* 148:102579.

Murphy, D.M., D. Roobroeck, D.R. Lee, and J. Thies. 2020. "Underground Knowledge: Estimating the Impacts of Soil Information Transfers Through Experimental Auctions†." *American Journal of Agricultural Economics* 102:1468–1493, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12101.

Nyondo, C.J., J. Minofu, J. Goeb, W.J. Burke, D. Kambewa, R. Chikowo, and S. Snapp. 2025. "The impact of site-specific soil-test-based extension advice on farm management in Malawi." *Food Policy* 133:102850.

Odendo, M., G. Obare, and B. Salasya. 2010. "Farmers' perceptions and knowledge of soil fertility degradation in two contrasting sites in western Kenya." *Land Degradation & Development* 21:557–564, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ldr.996.

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

Sanchez, P.A. 2002. "Soil Fertility and Hunger in Africa." *Science* 295:2019–2020.

Sanchez, P.A., and M.S. Swaminathan. 2005. "Hunger in Africa: the link between unhealthy people and unhealthy soils." *Lancet (London, England)* 365:442–444.

Shepherd, K.D., and M.G. Walsh. 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.

Suri, T., and C. Udry. 2022. "Agricultural Technology in Africa." *Journal of Economic Perspectives* 36(1):33–56.

Tully, K., C. Sullivan, R. Weil, and P. Sanchez. 2015. "The State of Soil Degradation in Sub-Saharan Africa: Baselines, Trajectories, and Solutions." *Sustainability* 7:6523–6552, Number: 6.

Vagen, T.G., and L.A. Winowiecki. 2023. "The Land Degradation Surveillance Framework Field Manual." Working paper, World Agroforestry.

Vågen, T.G., L.A. Winowiecki, J.E. Tondoh, L.T. Desta, and T. Gumbricht. 2016. "Mapping of soil properties and land degradation risk in Africa using MODIS reflectance." *Geoderma* 263:216–225.

Appendix: Additional Results

Table A1: List of soil properties measured by ICRAF MIR, AgroCares, and Jabba / Palintest. [back]

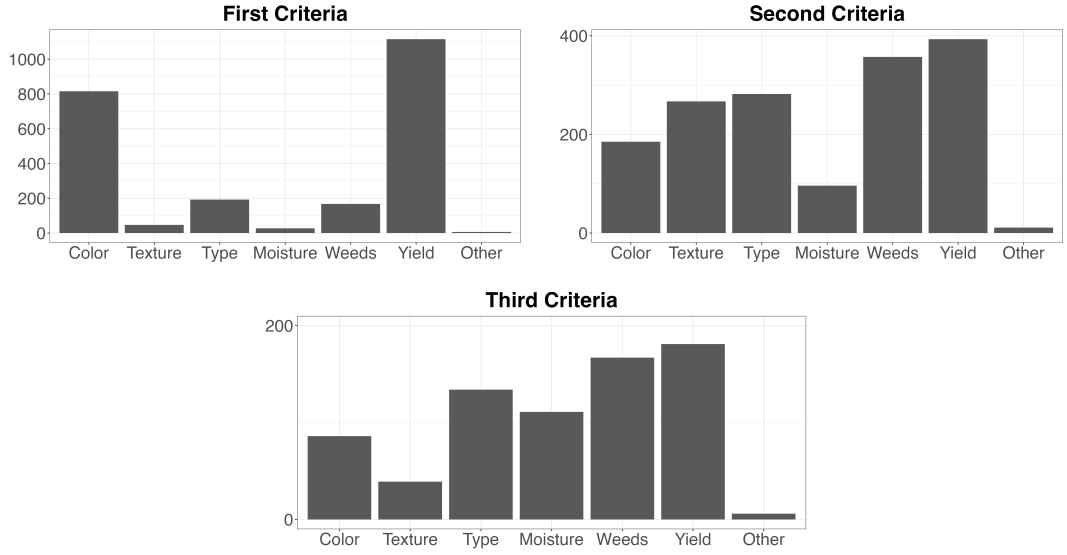
	ICRAF	AgroCares	Jabba / Palintest	ISDA	SoilGrid
pH	*	*	*	*	*
Organic Carbon	*	*		*	*
Organic Matter		*			
Total Nitrogen	*	*		*	*
Nitrate Nitrogen			*		
Ammonia Nitrogen			*		
Potentially Mineralizable Nitrogen		*			
Clay	*	*	*	*	*
Silt	*		*	*	*
Sand	*		*	*	*
Soil Moisture		*			
Cation Exchange Capacity (CEC)	*	*		*	*
Potassium (extr.)	*			*	*
Potassium (exch.)		*			
Calcium (extr.)	*		*	*	
Calcium (exch.)		*			
Magnesium (extr.)	*		*	*	
Magnesium (exch.)		*			
Phosphorous	*	*			
Phosphate phosphorous			*		
Phosphorous Sorption Index (PSI)	*				
Total Aluminum	*	*	*		
Total Iron	*	*	*		
Electrical Conductivity	*		*		
Sulphur	*		*	*	
Manganese	*		*		
Copper	*		*		
Boron	*				
Zinc	*			*	
Fertilizer recommendations		*			
Aluminum (extr.)				*	
Bulk Density				*	*
Total Carbon				*	
Iron (extr.)				*	
Stone				*	
Texture Class				*	
Coarse Fragments					*
Organic Carbon Density					*
Organic Carbon Stock					*

Abbreviations: exch. (exchangeable), extr. (extractable).

Table A2: Difference between ICRAF MIR and ICRAF wet chemistry results. [back]

	Spectroscopy Mean	Wet Chemistry Mean	Mean Difference	Difference P-Value	Correlation
pH	6.393	6.464	-0.071	0.721	0.991
Clay	29.779	24.479	5.300	0.231	0.951
Potassium	236.613	277.058	-40.445	0.585	0.892
Sand	55.571	63.050	-7.479	0.202	0.953
Silt	14.650	12.482	2.168	0.273	0.954
Phosphorus	70.276	72.598	-2.322	0.958	0.986
Iron	127.144	123.374	3.770	0.732	0.981
Magnesium	333.855	378.032	-44.177	0.616	0.979

Statistics are based on 30 observations.



Note: The y-axis is relative is each sub-plot.

Figure A1: Top three criteria chosen by farmers when assessing their soil quality. [back]

Table A3: Classification bins for soil properties used in cross-method comparison. [back]

Soil Property	**Categories**	**Range**
pH	Strongly Acidic	< 5.5
	Moderately Acidic	$5.5 - 6.5$
	Neutral	$6.5 - 7.5$
	Alkaline	> 7.5
Clay (%)	Very Low	< 15
	Low	$15 - 30$
	Moderate	$30 - 60$
	High	> 60
Soil Organic Carbon (%)	Low	< 1
	Moderate	$1 - 2$
	High	> 2
Total Nitrogen (g/kg)	Low	< 0.1
	Moderate	$0.1 - 0.3$
	High	> 0.3
Extractable Potassium (mg/kg)	Low	< 190
	Moderate	$190 - 600$
	High	> 600

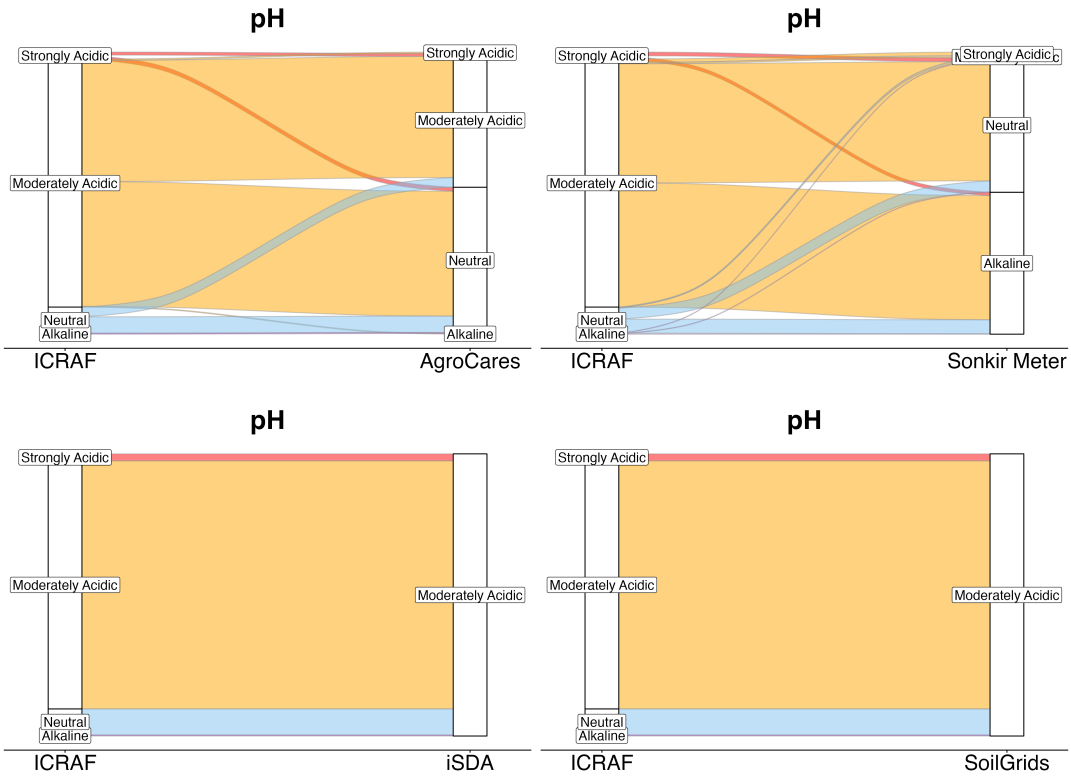


Figure A2: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for pH. [back]

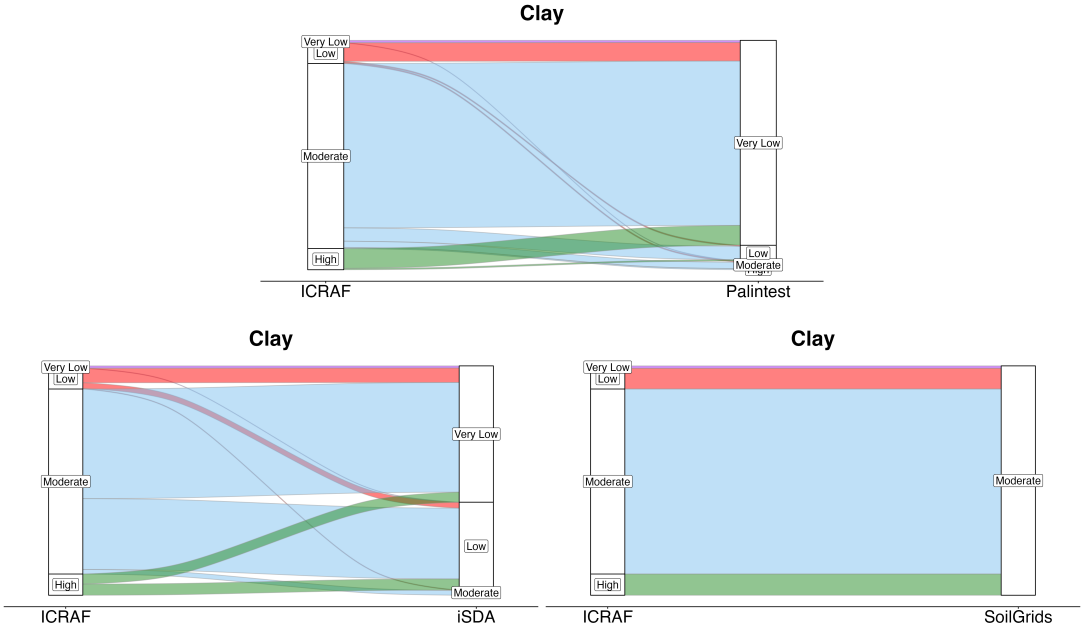


Figure A3: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for clay. [back]

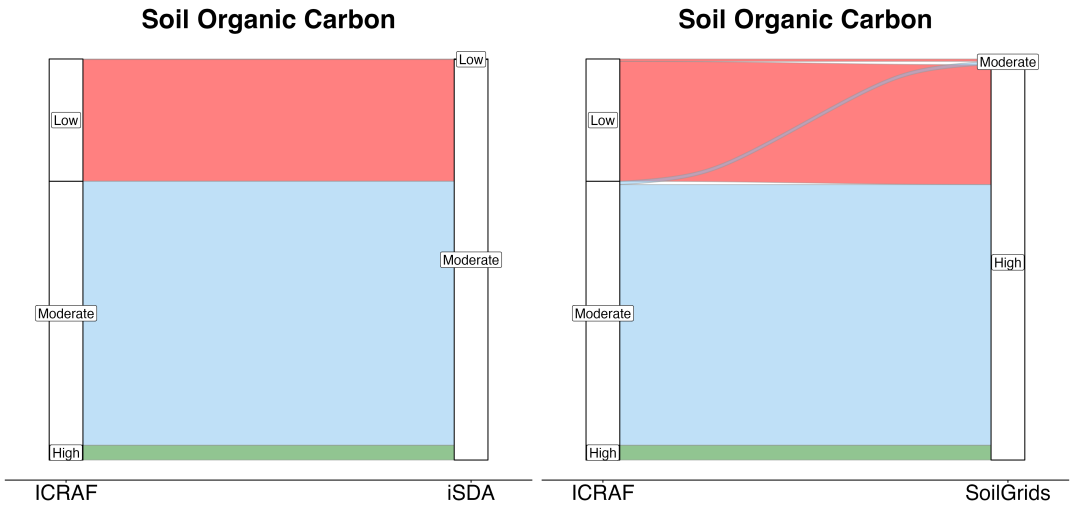


Figure A4: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for soil organic carbon. [back]

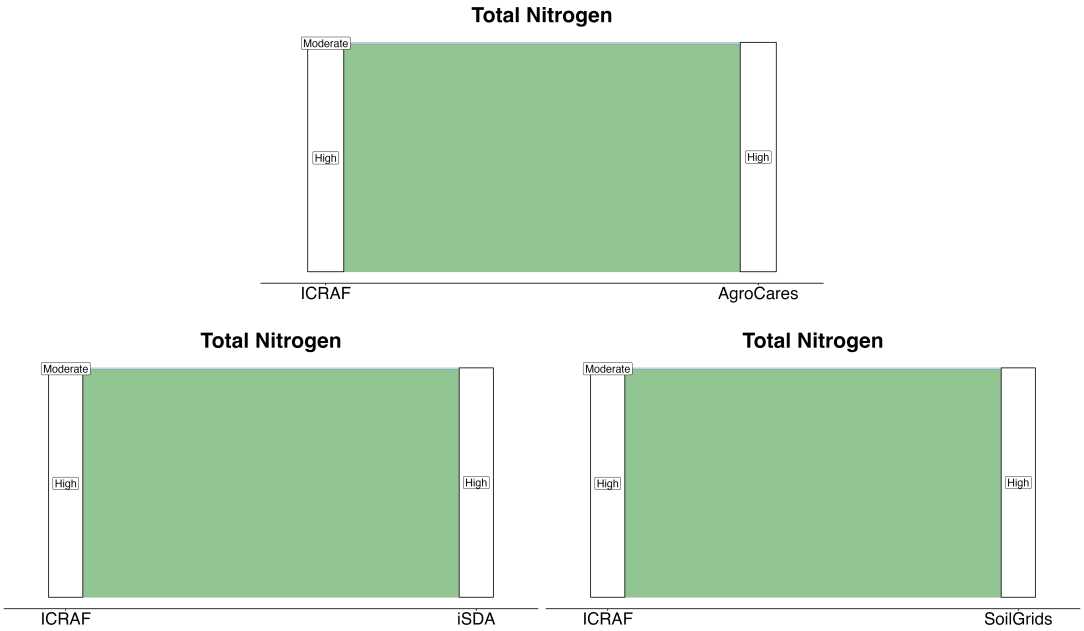


Figure A5: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for total nitrogen. [back]

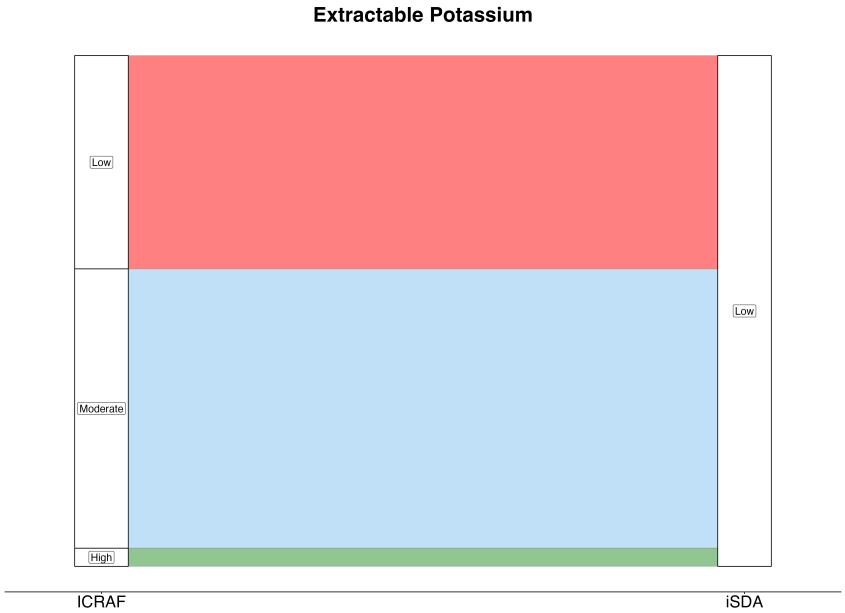


Figure A6: Categorical misclassification between ICRAF MIR and alternative soil measurement methods for extractable potassium. [back]