World Development 128 (2020) 104835

Contents lists available at ScienceDirect

World Development

journal homepage: www.elsevier.com/locate/worlddev

Assessing the downstream socioeconomic impacts of agroforestry in Kenya

Karl Hughes [a,*], Seth Morgan [b], Katherine Baylis [c], Judith Oduol [a], Emilie Smith-Dumont [a], Tor-Gunnar Vågen [a], Hilda Kegode [a]

[a] World Agroforestry (ICRAF), United Nations Avenue, PO Box 30677, Gigiri, Nairobi, Kenya
[b] Duke University, University Program in Environmental Policy, Box 90328, Durham, NC 27709-0328, USA
[c] University of Illinois, Agricultural and Consumer Economics, Mumford Hall, 1301 W Gregory Dr MC-710, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Accepted 7 December 2019
Available online 7 January 2020

Keywords:
Agroforestry
Impact
Kenya
Trees

ABSTRACT

Agroforestry is widely purported to improve the livelihoods of smallholder farmers, rehabilitate degraded landscapes, and enhance the provisioning of ecosystem services. Yet, evidence supporting these longer-term impacts is limited. Using a quasi-experimental impact evaluation design informed by a theory-based and mixed methods framework, we investigated selected intermediate and final outcomes of a nine-year effort led by Vi Agroforestry, a Swedish non-governmental organization (NGO), to promote agroforestry in large sections of Bungoma and Kakamega counties in western Kenya. We compared households belonging to 432 pre-existing farmer groups operating in 60 program villages and 61 matched comparison villages. To address potential self-selection bias, we used program targeting as an instrument for program participation, combined with the difference-in-differences approach to control for time-invariant differences between our treatment and comparison groups. We complemented the above with semi-structured interviews with a sub-sample of 40 purposively selected program participants. Despite evidence of variable program exposure and agroforestry uptake, we found modest, yet statistically significant, effects of Vi Agroforestry's program on intermediate outcomes, such as agroforestry product income, fuelwood access, and milk yields among dairy farmers. We also found that this program modestly increased asset holdings, particularly among households represented by female program participants.

1. Introduction

Agroforestry, or agriculture with trees (ICRAF, 2017), has long been touted as a triple-win for smallholder farmers, with the potential to mitigate environmental damage, increase income, and improve climate resilience. Recently, agroforestry and its suite of associated practices has increased in policy prominence (see, for example, Buttoud, Place, & Gauthier, 2013), and its promotion is one of the central pillars of Land Degradation Neutrality (LDN) investments globally, targeted to reach US$ 14.7 billion by 2021 (Maillard & Cheung, 2016). Despite this substantial interest and investment, little work has been undertaken to rigorously assess the longer-term impacts of agroforestry extension programs and integrated agroforestry systems (Brown, Miller, Ordonez, & Baylis, 2018).

While a number of studies have examined the effectiveness of specific agroforestry practices on intermediary outcomes, such as soil fertility and crop yields, results are mixed (see, for example, Akinnifesi, Ajayi, Sileshi, Chirwa, & Chianu, 2010; Odhiambo et al., 2001; Sjögren, Shepherd, & Karlsson, 2010). The impact of agroforestry on such outcomes is largely dependent on the specific practices introduced, the extent to which they are appropriately implemented, and their interaction with the biophysical and socioeconomic context in question. Fewer studies have examined the effects of agroforestry extension programs or integrated systems on more downstream outcomes, such as household income and food security. One exception is Place, Adato, Hebinck, and Omosa (2005) who estimate the effect of agroforestry-based soil fertility replenishment practices in western Kenya on food security and poverty using instrumental variables for adoption. However, the sample of households studied is small (n = 102), and the instruments used (e.g., whether any adult in the household previously

* Corresponding author.
E-mail addresses: k.hughes@cgia.org (K. Hughes), baylis@illinois.edu (K. Baylis), j.oduol@cgiar.org (J. Oduol), t.vagen@cgiar.org (T.-G. Vågen), h.kegode@cgiar.org (H. Kegode).

https://doi.org/10.1016/j.worlddev.2019.104835
0305-750X/© 2019 The Author(s). Published by Elsevier Ltd.

held a job) may violate the exclusion restriction[1], thereby rendering the results largely inconclusive.

A challenge in evaluating the impacts of agroforestry is that farmers tend to pursue several of its associated practices simultaneously, with the intention of deriving multiple benefits, e.g., improved soil and crop management and fodder, fuelwood, fruits, and timber for domestic use and sale. Many agroforestry practices are also expected to positively interact (Nair, 1993). Thus, while single practice efficacy and effectiveness studies are important, their use is limited in understanding the broader impacts of agroforestry.

Agroforestry also poses challenges for impact evaluation because of both the long duration in which outcomes are expected to manifest and the likely heterogeneity of such outcomes across social and agroecological settings. One possible approach could involve promoting contextually appropriate agroforestry practices in randomly assigned villages for a significant number of years and then comparing households within them with those in control villages against various intermediate and downstream outcome measures. Yet, executing such a study would be difficult given the time it takes for the synergistic interactions and expected impacts of the promoted agroforestry practices to fully manifest, exacerbated by the likelihood of significant contagion and spillover effects over this period.

Consequently, we took advantage of an agroforestry extension program that had been in operation in western Kenya for nearly a decade, implemented by Vi Agroforestry, a Swedish non-governmental organization (NGO). The next section describes this program and the specific projects we evaluated. Section 3—Methods—then follows by describing the approaches we used to mitigate both program placement and self-selection bias, while Section 4—Data—explains our survey administration exercise and the outcome variables we derived from the resulting data. Sections 5 and 6 present our results and discuss the corresponding policy implications, respectively.

The main contribution of this paper is that it assesses the longer-term impacts of an established and multifaceted agroforestry extension program, while testing intermediate mechanisms laid out in its theory of change. A novel measure of agroforestry adoption that captures its multidimensional nature is further put forward, and the proportion of change in economic outcomes that can be explained by agroforestry adoption is explicitly measured. Despite evidence of imperfect program exposure and agroforestry uptake, we identified modest, yet statistically significant, program effects on intermediate outcomes, such as agroforestry product income, fuelwood access, and milk yields among dairy farmers. There is further evidence that this program modestly increased asset holdings, particularly among households represented by female program participants.

2. Vi Agroforestry's program and implicit theory of change

2.1. Vi Agroforestry's program

2.1.1. General background

Vi Agroforestry is a Swedish NGO founded in 1983. It operates in four African countries—Kenya, Uganda, Tanzania, and Rwanda. It promotes the integration of woody perennials in smallholder farming systems to (a) directly produce agroforestry products, such as timber, fuelwood, fruits and livestock fodder, for both household use and sale; and (b) enhance the management of local natural resources by improving soil fertility, soil erosion control, and water infiltration. At the time of this study, Vi Agroforestry's program model focused on promoting the above practices among pre-existing smallholder farmer groups, coupled with other Sustainable Agricultural Land Management (SALM) practices, e.g., composting, crop rotation, and mulching. This was complemented with farmer group capacity strengthening through leadership training and the promotion of group savings and lending.

The primary agroforestry practices promoted by Vi Agroforestry are variants of three specific planting patterns: alley-cropping (intercropping trees with annual crops); boundary planting; and tree planting along soil erosion control structures. Boundary planting is common throughout both the targeted and non-targeted parts of the study area, but Vi Agroforestry encourages farmers to intensify this practice by integrating leguminous shrubs in the spaces in between long-term timber species, thereby creating multi-story boundary planting systems. In addition, farmers are trained to develop similar multi-story perennial systems along intra-plot erosion control structures, which include grass strips, trash lines consisting of crop residue, small contour bunds, trenches, and terraces.

In practice, farmers select specific tree species and agroforestry practices from the larger suite promoted by Vi Agroforestry, adapting them to their specific needs and circumstances. Its extension staff further tailor capacity development interventions to match the needs of each participating farmer group. Nevertheless, each group is expected to learn about the advantages of agroforestry, and Vi Agroforestry's activity calendar is coordinated around bi-annual tree seed distributions corresponding with the arrival of the two rainy seasons in its operational area. Tree seeds are distributed free of charge, including seeds for direct seeding and for raising in small-scale tree nurseries.

2.1.2. Specific projects implemented in the study area

Vi Agroforestry began promoting agroforestry in the study area in 2008 through the implementation of two projects: the Kenya Agricultural Carbon Project (KACP) and the Farmer Organizations and Agroforestry (FOA) project (Fig. 2). These two initiatives had their own field staff and funding structures but shared similar approaches for promoting agroforestry and other complementary SALM practices. The distinguishing feature of KACP is that it explicitly emphasised the carbon sequestration function of the promoted agroforestry and other SALM practices. Tree planting and management were incentivized by modest payments to farmer groups (equivalent to approximately US$ 3.00 per person per year) upon confirmation that trees had been planted and cared for on their farms. FOA, on the other hand, stressed the capacity development of farmer organizations, as a complement to the provision of tree seeds/seedlings and SALM training. It did not provide carbon payments, nor did it monitor tree planting with the same degree of rigor as KACP. Moreover, since FOA was focused on empowering farmer organizations, one of Vi Agroforestry's four supervision areas that made up the study area was handed over to partnering Savings and Credit Cooperatives (SACCOs) in 2014. Thus, Vi Agroforestry's training and other capacity development activities were implemented by SACCOs for the latter three years of the study period in this supervision area, i.e., from 2014 to 2016. Despite the above differences, the training and seed distribution regimen of the two projects were largely similar.

2.1.3. Socio-economic context

The study area comprises large sections of Bungoma and Kakamega counties. This area is dominated by Kenya's second largest ethnic group, the Luhya (KNBS, 2015). The Luhya trace their ancestry to the Bantu, Cushitic, and Nilotic peoples, completing their migration to the study area by 1850 (Jenkins, 2008). Here,

[1] The exclusion restriction states that the instrument in question should only affect the outcome variable indirectly by influencing treatment status (Angrist & Pischke, 2008).

they settled as farmers, and, with rising population density, most farm on less than one hectare of land (65% in our sample). Approximately 80% of the land is arable and comprises intensively managed crop and livestock systems. The area's main subsistence crops are maize, beans, finger millet, sweet potatoes, bananas, sorghum, potatoes, and assorted vegetables, while sugar cane, cotton, palm oil, coffee, tea, sunflower, and tobacco are grown as cash crops (County Government of Bungoma, 2018). However, given their small land holdings, many Luhya work in surrounding urban centers and the capital city. As such, many households in the study area have some reliance on remittances. Twenty-two percent of our sample, for example, reported having had received money from relatives working outside their respective communities in the last 12 months.

2.2. Program Theory of Change

To inform our impact evaluation, we constructed a basic Theory of Change (Fig. 1) for how Vi Agroforestry's program was expected to generate its intended longer-term outcomes. Its first precondition is that appropriate participation in this program took place among members of the targeted farmer groups, followed by their adoption of the agroforestry practices and tree/shrub germplasm it promoted. Adoption was then expected to result in multiple intermediary impacts. One such intermediary impact is improved soil health, which was expected to have, in turn, improved crop production, or—at the very least—reduced input costs, thereby increasing returns. And with the increased use and availability of tree/shrub fodder, increases in milk production and/or returns were also expected among dairy farmers. Moreover, increases in income from the sale other agroforestry products, such as timber, fuelwood, and fruit, was further expected, as well as more diversified income and food sources. Given their traditional role in collecting fuelwood, benefits specific for women were additionally expected, due to its increased availability on farm. The above intermediary outcomes were then expected to have interacted together, over approximately a five- to 10-year timeframe, to bolster household income, food and nutritional security, and resilience to shocks.

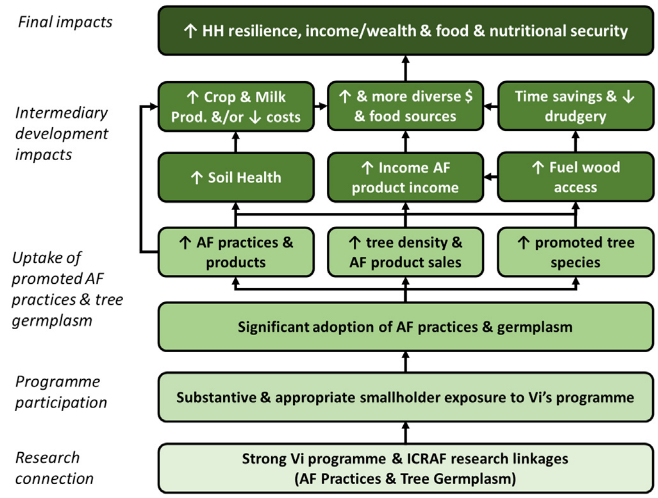


Fig. 1. Theory of Change framework for Vi Agroforestry's program.

3. Methods

3.1. Study area selection

Identifying the study area began with an exploration of locations where agroforestry had been substantially promoted in Kenya and where a credible quasi-experimental impact evaluation design could be pursued. While there have been several intensive, long-term efforts to promote agroforestry in this country, this was often done in combination with other interventions, thereby making it difficult to evaluate its specific impacts. The only organization found to have had a sustained and near exclusive focus on agroforestry promotion was Vi Agroforestry, hence the genesis of this study.

This culminated in a scoping mission in March 2016—conducted with Vi Agroforestry staff—to further narrow in on suitable intervention and comparison areas to serve as the study's focus. We found greatest potential in Bungoma and Kakamega counties. After visiting several sites and conducting informal interviews with farmers and other stakeholders, we concluded that this area had high potential for the impact evaluation, given the (reported) high rates of agroforestry adoption and agroecological comparability across program and potential comparison areas. In addition, neither Vi Agroforestry nor any other organization had substantively promoted agroforestry in the area prior to 2008.

3.2. Impact evaluation design

Given that Vi Agroforestry's program was not randomized, we faced two potential sources of selection bias. First, the communities targeted by Vi Agroforestry and potential comparison communities might be systematically different in both observable and non-observable ways that affect outcomes (i.e., *program placement bias*). Second, even within the targeted communities, households that participated in Vi Agroforestry's program may have been similarly systematically different when compared to those that did not (i.e., *self-selection bias*).

Our main strategy for countering **program placement bias** involved the following: First, we identified specific sub-locations (the smallest administrative unit above the village in Kenya) where Vi Agroforestry had operated since the baseline period within the two counties. We then worked with local informants to purposively match these sub-locations with potential comparison sub-locations based on their similarity in perceived wealth status and agroecological characteristics.[2] Next, a scoping survey was administered in all villages within the purposively matched program and comparison sub-locations to capture basic demographic information and geocoordinates of these villages, as well as to verify the existence of active farmer groups operational since the initial years of Vi Agroforestry's program. Village-specific data were then compiled on key geospatial and demographic variables from secondary data, including population density, baseline soil conditions and tree cover, elevation, rainfall, and distance from major road networks (as a proxy for market access).

Thereafter, propensity score matching (PSM) (Rosenbaum & Rubin, 1983) was used to match program villages to potential comparison villages within each of Vi Agroforestry's four main supervision areas, hereafter referred to as Village Sampling Zones (VSZs). In the end, the initial sample of 336 villages (194 program and 142 comparison villages) from all the purposively matched sub-locations was reduced to 121 (60 program and 61 comparison), with one additional village added, given that one of the matched comparison villages turned out to be two distinct villages. Fig. 2 presents the locations of the final set of matched villages, while Annex 1 describes the matching exercise in greater detail and the results of statistical balancing tests.

[2] We were informed by Vi Agroforestry staff that there was no particular reason why they had not expanded into these other areas. Lack of funding and human resource capacity were the key limiting factors.

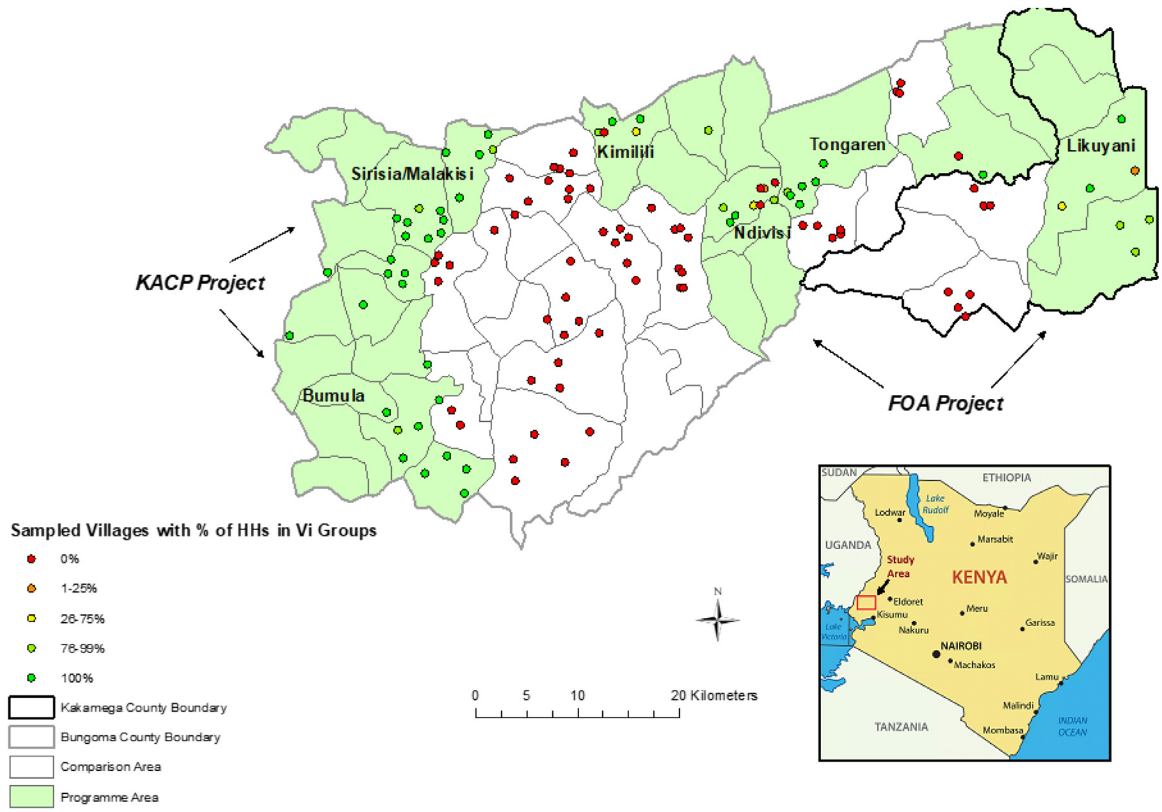


Fig. 2. Map of study area with matched villages.

The main strategy used to address **self-selection bias** involved mimicking Vi Agroforestry's targeting process when it engaged with the program villages during the baseline period. At this time, it specifically targeted pre-existing farmer groups, offering all—including their respective members—the opportunity to participate in its program. As such, we specifically ensured that the village sampling frames from which all the respondents (12 females and 12 males per village) were to be selected were currently active members of one or more existing farmer groups formed in 2008/09 or earlier. In the program villages, interviewing members of all pre-existing farmer groups, regardless of whether they engaged with Vi Agroforestry, ensured that we did not simply compare a unique set of farmer groups with a more general set in the comparison villages. Similarly, identifying and interviewing all groups and members that had been active in the early stages of the program period in the comparison villages enabled us to identify a set of farmer groups that would have been offered the opportunity to participate in Vi Agroforestry's program had the organization gone to these villages and followed the recruitment approach it used in the program area.

Assuming that (a) the above village matching exercise was successful in mitigating program placement bias and, by extension, (b) the matched program and comparison are as good as randomly assigned, this enabled us to generate both intention-to-treat (ITT) and local average treatment effect (LATE) estimates. We derived ITT estimates by simply comparing all sampled households in the villages targeted by Vi Agroforestry with those in the comparison villages, regardless of whether they belonged to a Vi Agroforestry affiliated group. Given that the sample of households from the program area included a significant number (~25%) that were not members of such groups and, hence, did not directly participate in Vi Agroforestry's program, the ITT estimates likely underestimate the effects of such participation.

We further assume that the opportunity provided by Vi Agroforestry to the pre-existing farmer groups of the program area to participate in its program made it more likely and never less likely for them to have participated, i.e., the monotonicity assumption (Heckman & Vytlacil, 2007). This, coupled with the above 'good as randomly assigned' assumption, enabled us to implement two-stage least squares regression (2SLS) to derive LATE estimates (Imbens, 2010), where program targeting was used as an instrument for participation. Provided that these two assumptions hold—and given that no households were found to have participated in Vi Agroforestry's program in the comparison villages—the LATE estimates approximate the average effects of Vi Agroforestry program participation.

Several other measures were taken to counter both program placement and self-selection bias. A key limitation of the study, for example, is that no suitable baseline survey was undertaken. This limited our ability to check and control for time variant baseline differences between the intervention and comparison groups, as well as the undertaking of difference-in-differences estimation. To address this limitation, recalled baseline data were collected from respondents on asset ownership, housing characteristics, livelihood pursuits, and tree planting and land management practices. We took advantage of a significant historical event which had taken place one year prior to the baseline period, i.e., Kenya's post-election violence, and used this as a historical marker.[3] As argued by White and Bamberger (2008), all survey questions

[3] Respondents were specifically requested to recall conditions prior to the nationwide events that took place in December 2007 to January 2008. The geographical area where the survey was carried out did not witness significant violence, as had taken place in other parts of Kenya, so we thought it was appropriate to take advantage of this specific historical marker.

are based on a degree of recall, and there are some items, e.g., major events and purchases, that can recalled with reasonable accuracy vis-à-vis an intervention's baseline period.

Because the village matching exercise was first implemented at the VSZ level and to minimize our results being influenced by VSZ specific trends, all our Ordinary Least Squares (OLS) and 2SLS models included VSZ dummies as fixed effects. Standard errors were further clustered at the farmer group level, given that Vi Agroforestry targeted *pre-existing* farmer groups (our pseudo unit of assignment). In addition, given that some measured differences were found between our intervention and comparison groups (see Section 4), all our outcome models included measures of such differences (i.e., covariates) significantly correlated with being in the program area ($p < 0.1$).

Further, a theory-based approach (White, 2009) was followed. Data were captured on various intermediary measures along the causal pathway towards the program's expected effects on consumption expenditure and asset accumulation, as per the theory of change. This enabled us to assess to the extent to which changes associated with this theory of change unfolded as expected. We complemented this with statistical mediation analysis (MacKinnon, 2008) using Stata's *sem* (structural equation modelling) command. Here, we assessed the extent to which the hypothesis that agroforestry adoption was responsible for our estimated program effects on asset accumulation is consistent with the data.

The study also included a qualitative component, the objectives of which were to (1) analyze variation in agroforestry adoption intensity between female and male Vi Agroforestry group members and across the two main project areas (KACP and FOA); and (2) explore the mechanisms through which different components of Vi Agroforestry's program may have contributed to livelihood improvements. This involved carrying out semi-structured interviews with a sub-sample of 40 purposively selected Vi Agroforestry group members, stratified by gender and location with 20 farmers in four villages under KACP (Bumula and Sirisia constituencies) and 20 others in four villages under the FOA project (Likuyani and Kimilili constituencies). We used a structured questionnaire combined with ranking exercises, record sheets for trees and products, farm sketches, and in depth interviews for formal local knowledge acquisition using the Agroecological Knowledge Toolkit (AKT5) (Dixon, Doores, Joshi, & Sinclair, 1999).

4. Data

4.1. Quantitative data collection

Based on (a) the first round of the abovementioned qualitative component; (b) the theory of change presented in Subsection 2.2; (c) a more in depth understanding of Vi Agroforestry's program; and (d) discussions within our team, a draft structured survey instrument was developed using the Open Data Kit (ODK) platform. This instrument was then reviewed with Vi Agroforestry technical staff in Bungoma County, followed by piloting it with four smallholder farmers residing in this same county but outside the study area. It was revised thereafter, and a team of 24 enumerators were trained to carry out its administration for a period of three days. This training included a practical exercise where each enumerator interviewed a farmer, followed by another extensive review of all the instrument's questions. Throughout the enumerator training program, this instrument was iteratively refined. The enumerators were recruited from a pool of 306 applicants, from which 58 were selected for interviews and 24 finally shortlisted for training. Even numbers of female and male enumerators were selected (given that 12 female and 12 male farmers were to be interviewed in each village). All spoke the Luhya dialect.[4]

The survey was carried out in the 121 matched program and comparison villages from August 4 to October 1, 2016. Following the village matching exercise described above, an advance team of enumerators was sent to each village ahead of the survey administration team to prepare lists of farmer group members. All respondents in both the program and comparison villages met the following criteria:

- Member of a farmer group formed in 2008/09 or earlier
- Currently an active member of that group since 2008/09 or earlier
- Household must have existed in 2007 or before
- Household must have been farming the same main parcel of land from 2007 to the present

These screening criteria were used, in part, to mimic Vi Agroforestry's selection process when it engaged with the villages in the program area during the early stages of its program (see above).

Male and female farmer group members belonging to 1450 and 1410 households in program and comparison villages, respectively, were interviewed, with the sex of each responded also selected at random. During these interviews—which typically lasted from one hour to one and a half hours—data were captured on both the respondent, e.g., their age, educational status, and farmer group participation, and their household, e.g., educational status of other household members, baseline livelihood pursuits, and asset ownership. The last part of the survey involved visiting the household's main farming parcel (*shamba*). While here, the respondent was asked questions about this main parcel, e.g., tenure arrangements and size. The enumerator then visited each plot within the parcel to (a) make observations (e.g., tree presence); (b) ask specific questions about the plot at present and during the baseline period (e.g., types of crops grown); and (c) take GPS coordinates. We believe that this systematic plot-by-plot observation and questioning process—coupled with periodically refreshing the respondent's mind about the baseline period using the abovementioned historical marker—significantly mitigated, but, of course, did not eliminate, the potential for recall bias.

During data collection, the data were downloaded periodically from a dedicated Internet site hosted by Ona and checked for survey administration errors. In the end, the collected data were imported into Stata for cleaning, variable construction, and analysis. For continuous measures, outliers were addressed by trimming them to the 1st and 99th percentiles, particularly when data entry errors were not clearly identifiable. During data cleaning, it became apparent that 63 interviewed farmer group members did not meet the above inclusion criteria, so were dropped from the dataset, reducing the total sample size to 2797. Finally, a pre-analysis plan was also prepared following the International Initiative for Impact Evaluation's (3ie) Registry for International Development Impact Evaluations (RIDIE) format, uploaded onto its site, and subsequently formally accepted.[5]

4.2. Outcome variable construction

We used the survey data to construct aggregated measures associated with the intermediary outcomes and impacts relevant to the theory of change presented in Section 2.

[4] It was not translated into Luhya because this language is not one of the primary written languages in Kenya and doing so phonetically in the English alphabet would have introduced additional complexity.

[5] See: http://ridie.org/index.php?r=search/detailView&id=504.

Fig. 3. Agroforestry Adoption Index for Vi's Program.

4.3. Uptake of promoted agroforestry practices and tree species

One complication about agroforestry is that it is not a singular practice. Consequently, we developed a composite Agroforestry Adoption Index (AF Index) to enable data associated with its various dimensions to be practically aggregated and analyzed. While we recognize that no 'one size fits all' when it comes to measuring agroforestry, we worked with Vi Agroforestry technical staff to devise 10 binary indicators to reveal the extent to which the specific agroforestry practices and tree and shrub species it promoted were taken up by the farmers it targeted. Following Alkire and Foster (2011), we grouped these indicators under three dimensions: Practice Uptake; Intensity of Practice; and Tree Species (Fig. 3). Each dimension was weighted equally, as well as each indicator under each dimension. The resulting index ranges from 0 to 1, with scores of 1 and 0 indicating that the household in question met all or did not meet any of the thresholds of the 10 binary indicators, respectively. Consequently, the more a household took up the practices and tree species associated with Vi Agroforestry's program, the higher its score on the index.

If a household had significantly taken up Vi Agroforestry's promoted agroforestry practices and tree germplasm, we would expect to both (a) see trees and shrubs integrated into its farming plots; and (b) hear about the harvesting of tree products, such as fuelwood, timber, fruits and/or fodder, from these plots. We would also expect to see tree-based natural resource management (NRM) techniques being applied, such as planted trees along contour lines interspersed with shrub species. We would also expect several other complementary agroforestry practices to be present on the farm, such as fruit orchards, woodlots, and/or fodder banks. This is the rational for the *Practice Dimension* and its associated indicators.

In addition, the uptake of these practices should be significant, hence the *Intensity of Practice* dimension. One would expect to see, for example, the presence of a relatively high density of trees on the adopting household's food and horticultural plots, coupled with other complementary agroforestry practices and significant income earned through the sale of the resulting products. Finally, one would expect to find 'signature' tree and shrub species promoted by Vi Agroforestry on the farm, ranging from leguminous shrubs through to more long-term exotic an d native species, hence the *Tree Species* dimension.

4.4. Tree product sales

Agroforestry may generate longer-term socio-economic impacts by directly generating income through the sale of agroforestry products, such as timber, fuelwood, and fruits. During our household survey, enumerators observed if there were any trees or shrubs within or along the boundaries of the interviewed households' fields and other land use areas. If so, the respondents were asked whether they had produced and, if so, sold any products over the last 12 months, such as fodder, timber, fuelwood, and/or fruits. While, again, we recognize the potential for recall bias, they were also asked the same questions about their plots with respect to the baseline period.

4.5. Fuelwood cash value and collection time

Another expected intermediary outcome associated with the increased uptake of agroforestry is increased access to fuelwood, given that it can be readily obtained from the household in question's farm. Given that nearly all households in the study area are dependent on fuelwood (>99%), coupled with the gender-based division of labor with respect to its collection, we assume that reducing the amount of time and effort spent collecting it would positively benefit women as well. To capture data on the amount of fuelwood accessed on farm and the time spent collecting it, the respondents were first asked whether their households had used any fuelwood for cooking, heating, or any other purpose during the previous month. They were then asked where they sourced it from, including the primary source, followed by (a) the number of times they collected it over the past month; (b) the approximate number of hours spent undertaking such collection on each occasion; and (c) how much of what was collected would have cost if it had been purchased from the local market. Through these data, variables pertaining to the estimated cash value of fuelwood collected on farm and time spent undertaking such collection were constructed.

4.6. Tree fodder and milk yields

Small-scale dairy production can be significantly profitable when expensive dairy concentrates designed for larger scale operations are substituted for high-protein leguminous shrub fodder (Franzel, Carsan, Lukuyu, Sinja, & Wambugu, 2014). Consequently, dairy producers sampled from the farmer groups were asked specific questions about milk yields and the fodder shrubs promoted by Vi Agroforestry. For the milk yield measures, we restricted our analysis to cows only, given that only seven respondents reported owning improved dairy goats in 2007. We also differentiated between local and improved cows, given that the cost of these animals and their milk yields differ substantially.[6]

4.7. Household wealth

Four primary sets of indicators were constructed as proxies for household income/wealth status: (1) daily household consumption expenditure per capita adjusted for Purchase Power Parity (PPP); (2) consumption expenditure weighted asset indices; (3) household asset indices derived through principal component analysis (PCA); and (4) unweighted asset indices. While recall data are plausibly reliable for assets, they are clearly problematic for consumption expenditure. To estimate baseline consumption expenditure, we followed O'Donnell, van Doorslaer, Wagstaff, and Lindelow (2008) and regressed the basket of assets reportedly owned in 2007 on our 2016 consumption expenditure data. Details on this procedure and how we constructed all four sets of the above measures are presented in Appendix 3.

5. Results

5.1. Baseline and time invariant respondent and household covariate balance

The objective of the village geospatial and secondary data matching exercise was to achieve an unbiased comparison between the households of the program and comparison areas. A comparison of these two groups against the full set of 46 covariates is presented Appendix 2, while Table 1 presents only those found to be statistically significant at a 95% level of confidence or greater net of VSZ. We controlled for these differences in subsequent analyses, as part of our bias mitigation strategy.

Farmer group members in the program area are slightly more likely to be female and about 6% less likely to head their respective households. Moreover, while they are more likely be technically skilled, they are also 8% less likely to own their respective household's main farming parcel outright. Program area households were also more likely to have reared livestock and have one or more members in formal employment in 2007. They are, furthermore, more likely to be elderly headed, reside further from tarmacked road networks, and have had soils richer in organic matter at baseline.

While by no means extreme, the variables associated with these differences are correlated with many of our study's outcome measures. Hence, we included all covariates correlated with our program area dummy ($p < 0.1$ and net of VSZ) in all our models used to estimate the effects of Vi Agroforestry's program.

Table 1
Covariate comparison of households in program and comparison areas.

Characteristic	Program Mean ($\hat{p}$)	Comparison Mean ($\hat{p}$)	Difference (net of VSZ)
Respondent is head of household	0.48	0.54	-0.0574^{***} (0.019)
Respondent is spouse of head	0.41	0.37	0.0383^{**} (0.019)
Respondent has specialized tech. skills	0.06	0.04	0.0190^{**} (0.008)
HH reared livestock in 2007	0.62	0.57	0.054^{***} (0.019)
HH member employed in 2007	0.16	0.13	0.0278^{**} (0.013)
Head is 60 or older	0.33	0.29	0.0424^{**} (0.018)
Respondent owned main parcel (07)	0.46	0.54	-0.083^{***} (0.019)
Estimated 2007 soil carbon (plot avg.)	24.38 g/kg	22.97 g/kg	1.36^{***} (0.222)
Observations	**1411**	**1386**	**2797**

Standard errors in parenthesis; VSZ = Village Sampling Zone; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; probit regression used for net of county and VSZ differences, so coefficients are not directly interpretable, only the *t/z*-statistics.

5.2. Vi Agroforestry program exposure

We employed three measures to gauge how intensively members of the targeted farmer groups had been exposed to Vi Agroforestry's program (Table 2). The first indicator is simply whether the respondent reported that their respective farmer group engages in tree planting and management. If Vi Agroforestry substantively engaged all members of the farmer groups it targeted, we would expect a high percentage to report tree planting and management as at least one of the key activities undertaken by their respective groups. Notably, only 60% of the Vi Agroforestry group respondents (representing 75% of the program area respondents) reported this to be the case. Nevertheless, the difference in favor of Vi Agroforestry affiliated respondents vis-à-vis comparison area respondents is 26%.

One of the ways Vi Agroforestry promoted agroforestry was through training farmers in tree planting and management. However, only 50% of Vi Agroforestry group respondents reported having had received such training in the last three years, compared to 23% of the respondents in the comparison area. This may be due to an expectation that certain trained group members (e.g., group leaders) would pass down the training they received to their fellow group members. However, either this did not happen for many respondents or it was not perceived as 'training'. The other possibility is that the respondents had been trained prior to the three-year recall period.[7]

Our qualitative findings complement the above story: While the most commonly cited form of training was related to tree planting and management, only 23 out of the 40 farmers (14 out of the 23 women and nine of the 17 men) reported having had received it. And like our quantitative findings, the highest numbers were in the Bumula and Sirisia sites.

In summary, exposure to agroforestry promotion was significantly greater in the program area in general and among Vi Agroforestry group members in particular. However, it is also clear that many members were not exposed, at least in the recent past. We will now examine the extent to which this trend is similar for the uptake of the agroforestry practices and tree and shrub germplasm promoted under Vi Agroforestry's program.

[6] Rather than taking the average of all the reported milk yields across improved and local cows by period and then finding the difference between these averages, we first differenced by cow type and then took the averages across these differences. If a household owned only one cow type in both time periods, then the average pertains to only this type. Moreover, if a household had only local cows in 2007 and then added one or more improved cows to its dairy portfolio, only the average changes in milk yield for the former were assessed. This was to ensure that the estimated changes in milk yield are not a reflection of a household having had, for example, upgraded from one or more local cows to one or more improved cows.

[7] While the three-year cut-off is arbitrary, Vi Agroforestry periodically trained the farmer groups it targeted, rather than training them as a 'one-off' at the beginning of its engagement. Hence, this cut-off was deemed reasonable to assess the extent to which respondents had been recently exposed to such training.

Table 2
Comparison of HHs in program and comparison areas, program exposure.

Program Mean ($\hat{p}$) (Vi & Non-Vi groups)	Vi Group Mean ($\hat{p}$) (subset of program group)	Comparison Mean ($\hat{p}$)	Program vs. comparison (net of VSZ)	Vi group vs. comparison (net of VSZ)
Respondent reported that tree planting and management is a key activity of their farmer group				
0.54	0.60	0.33	0.21***	0.26***
			(0.019)	(0.020)
Respondent reported receipt of tree planting & mgt. training in last 3 years with significant implementation				
0.46	0.50	0.23	0.24***	0.27***
			(0.017)	(0.019)
Reported receipt of any type of extension support in last 3 years				
0.49	0.53	0.40	0.098***	0.12***
			(0.019)	(0.020)
1411	**1094**	**1386**	**2797**	**2480**

Standard errors in parenthesis; VSZ = Village Sampling Zone.
*p < 0.1, **p < 0.05, ***p < 0.01; probit regression used for net VSZ differences (dprobit option).

5.3. Adoption of promoted agroforestry, related practices and tree germplasm

We compared the three groups vis-à-vis the AF index (Table 3). While the index scores are low for all groups in the 2007 period, they are slightly higher in the program area in general and among the Vi Agroforestry group members in particular. This is the case across all three dimensions[8], indicating that the practice of agroforestry was already slightly greater in the program area prior to the arrival of Vi Agroforestry. Nevertheless, a second observation is that the change from 2007 to 2016 was significantly greater for the program area and specifically among Vi Agroforestry affiliated households, both overall and across all three dimensions. There is evidence, therefore, that many households in the program area adopted what Vi Agroforestry promoted. The final observation, however, is that the average index scores for 2016 are not particularly high at 0.27 and 0.28 for the overall program area and Vi Agroforestry affiliated households, respectively, out of a maximum possible score of 1. This reveals that the uptake was not as significant as expected.

We further examined the extent to which the overall average results for the index and its specific dimensions differ by geographic area and specific sub-groups (hereafter referred to as sub-group analysis). The results do not differ significantly for households with female and male members, nor by sex of household head. However, we found considerable and statistically significant variation among the four VSZs. The spatial variation is clear in Fig. 4. Our qualitative findings corroborate this spatial variation in the uptake of agroforestry practices. For example, 23 out of the 40 purposively sampled farmers had planted rows of timber species along either contours or trash lines as promoted by Vi Agroforestry, but this was observed primarily in the Bumula and Sirisia sites.

5.4. Tree product sales and fuelwood access

We further compared program and comparison households in relation to their sales of agroforestry products (Table 4). There are several noteworthy observations. First, households in the program area reported higher income from the sale of agroforestry products, as well as increases in such income from the baseline period. However, the second observation is that most households in *both* the program and comparison areas reported no sales at all; the median is 0 for both groups. As is clear from Table 4, just over one-third of households in the program area reported sales over Ksh 1000 (≈US$ 10), as compared with about one-fourth in the comparison area.

5.5. Fuelwood

We found positive differences in favour of the program area for both the estimated cash value of fuelwood collected on farm and the number of hours per month spent collecting it (Table 5). On average, the cash value of fuelwood collected in the previous month was 14% higher and time spent collecting it about 1.5 h less.

5.6. Tree fodder use and milk yields

Twice as many dairy producers in the program area were found making use of shrub fodder, and this increased by 27% from the baseline period against 10% in the comparison area (Table 6).[9] For average milk yields and increases in milk yields, there are consistent and positive results. We see that reported milk yields increased in favor of the program area by over 0.25 L per day. A greater percentage of dairy producers in the program area also self-reported that their milk yields had increased from the baseline period—52% against 44%.

A relevant question, of course, is: To what extent was the relatively greater increase in milk yields among dairy farmers in the program area driven by their relatively greater use of tree/shrub fodder? While we recognize that comparing milk yields between tree/shrub fodder users and non-users would be inconclusive (i.e., there may be one or more 'omitted' variables correlated with such differential uptake that could account for the difference), failing to see such a relationship would provide strong grounds to reject this as a hypothesized mechanism.

In Fig. 5, four box plots for our differenced milk yield measure are presented (converted into a percentage change): the first is for the comparison area and the second for the overall program area, while the third and fourth are specific to tree/shrub fodder and non-fodder users residing in the latter, respectively. Changes in milk yields between the comparison area and non-tree/shrub fodder users in the program area are very similar. However, the box plot for the tree/shrub fodder users in the program area clearly standout, with three-quarters of the distribution reporting positive milk yield increases and a median increase of approximately 25%.

This is generally consistent with findings from previous research. Milk production has been shown to increase by 0.6–0.75 kg per kilogram of dried *Calliandra calothyrsus* under farmers'

[8] Each of the three dimensions was reweighted to fall on the same scale ranging from 0 to 1. This enables a comparison of each dimension with the overall index.

[9] Given that the dairy farmers are a sub-set of the overall sample, a specific set of covariates correlated with program area (at $p<0.1$) specific to this sub-sample was used.

Table 3
Agroforestry Adoption Index Scores: Intervention Group Comparison.

		Program Area Mean	Vi Group Mean	Comparison Mean	PA vs. non-PA (dif.)	Vi vs. non-PA (dif.)
Overall Index	2007	0.15	0.15	0.11	0.033*** (0.0079)	0.036*** (0.0084)
	2016	0.27	0.28	0.17	0.091*** (0.0097)	0.10*** (0.010)
	Change	0.12	0.13	0.06	0.058*** (0.0075)	0.067*** (0.0082)
Dimension 1: Practice Uptake	2007	0.20	0.20	0.15	0.041*** (0.0099)	0.045*** (0.011)
	2016	0.31	0.33	0.23	0.084*** (0.011)	0.093*** (0.012)
	Change	0.11	0.13	0.08	0.043*** (0.0098)	0.049*** (0.011)
Dimension 2: Intensity of Practice	2007	0.13	0.13	0.10	0.024*** (0.0087)	0.025*** (0.0092)
	2016	0.22	0.24	0.15	0.073*** (0.0094)	0.083*** (0.010)
	Change	0.10	0.11	0.05	0.049*** (0.0094)	0.058*** (0.010)
Dimension 3: Tree Species	2007	0.12	0.12	0.08	0.035*** (0.010)	0.039*** (0.011)
	2016	0.26	0.28	0.14	0.12*** (0.014)	0.13*** (0.015)
	Change	0.14	0.16	0.06	0.082*** (0.011)	0.094*** (0.012)
	Observations	**1411**	**1094**	**1386**	**2797**	**2480**

PA = Program Area; *p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated with program area (p = <0.1) used in all models estimated using OLS; VSZ dummies used for fixed effects.

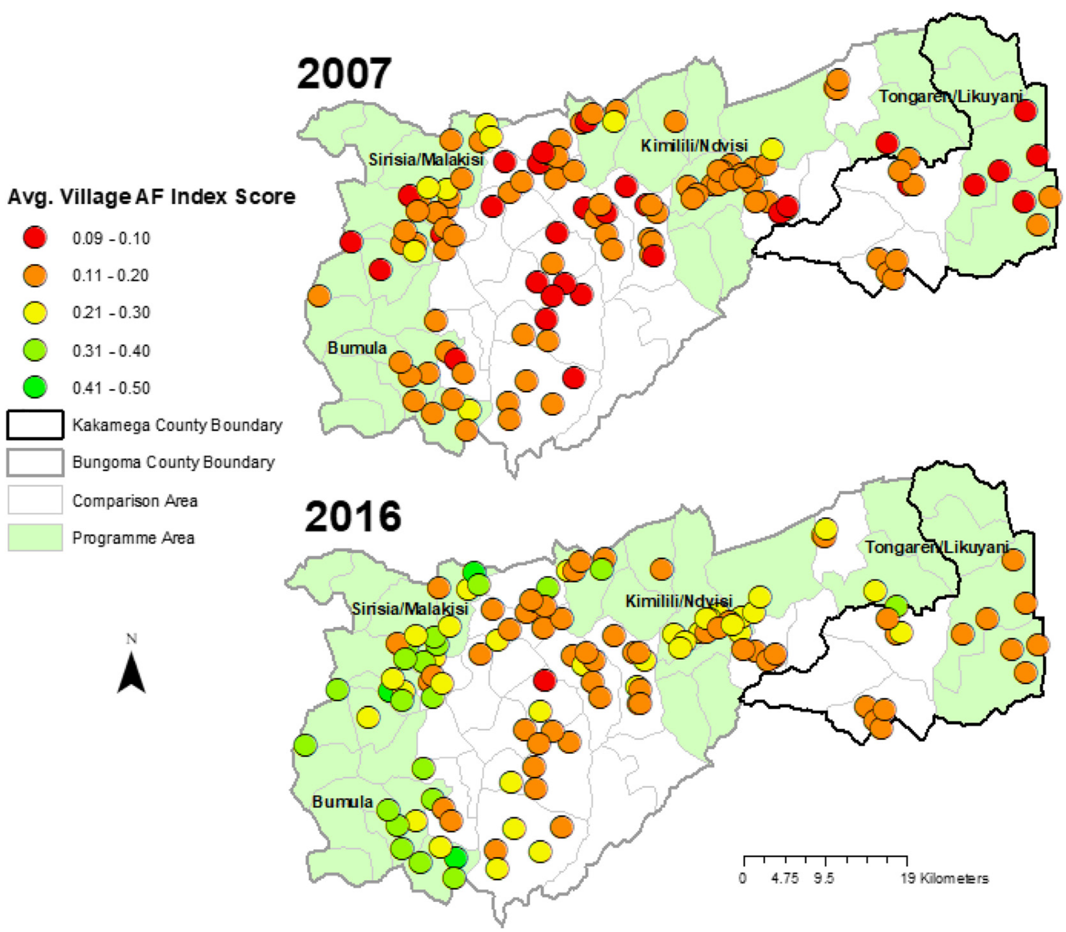


Fig. 4. Village level averages in Agroforestry Adoption Index 2007 to 2016.

Table 4
Agroforestry Product Sales: ITT and LATE Estimates.

	Sale of AF Products Ksh (2016)	p̂ over Ksh 1000 (2016)	Sale of AF Products Ksh (2016–2007)	p̂ over Ksh 1000 (2016–2007)
Raw results				
Program Mean	8647	0.36	6348	0.33
Comparison Mean	4785	0.25	3199	0.22
Unadjusted diff.	3863***	0.11***	3149***	0.10***
	(836.6)	(0.017)	(775.1)	(0.017)
Observations	**2797**	**2797**	**2797**	**2797**
OLS (ITT)				
Coefficient	3428***	0.11***	3385***	0.12***
	(886.8)	(0.020)	(858.2)	(0.020)
Observations	**2790**	**2790**	**2790**	**2790**
2SLS (LATE)				
Coefficient	4344***	0.14***	4290***	0.14***
	(1118.8)	(0.026)	(1082.0)	(0.025)
Observations	**2790**	**2790**	**2790**	**2790**

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated with program area (p = <0.1) used in all models; VSZ dummies used for fixed effects in all models; first two models control for recalled agroforestry product sales figures for 2007.

Table 5
Fuelwood Cash Value and Collection Time.

	Cash Value of Fuelwood from Farm	Hours in Past Month Collecting Fuelwood
Raw results		
Program Mean	1281	9.43
Comparison Mean	1123	10.84
Unadjusted dif.	158***	−1.41***
	(43.4)	(0.44)
Observations	**2797**	**2784**
OLS (ITT)		
Coefficient	144***	−1.47***
	(40.3)	(0.42)
Observations	**2790**	**2732**
2SLS (LATE)		
Coefficient	182***	−1.87***
	(50.8)	(0.52)
Observations	**2790**	**2732**

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated with program area (p = <0.1) used in all models; VSZ dummies used for fixed effects.

management (Franzel et al., 2014), the primary fodder shrub promoted by Vi Agroforestry. The profitability of such shrubs has been further demonstrated, with net benefits of $114 per cow per year when such shrubs are used as additional feed and $122 when substituted for commercial dairy meal entirely (Place et al., 2009).

5.7. Household consumption expenditure and asset wealth

Our results reveal that Vi Agroforestry's program had little, if any, impact on actual consumption expenditure, at least overall. The first column of Table 7 presents results for our 2016 consumption expenditure data. The models included an asset weighted index used to estimate 2007 consumption expenditure as a control (see Appendix 3). The single difference estimates for the 2016 consumption expenditure measure are in the right direction but are small, i.e., USD $0.067 and $0.085 per day per capita for the ITT and LATE estimates, respectively. They are also statistically insignificant.

The estimated effects presented in the latter two columns are larger but are still modest, with three out of the four being

Table 6
Tree fodder use and milk yields among dairy producers.

	Tree fodder use 2016 (p̂)	Tree fodder use difference (p̂)	Avg. 2016 milk ltr./day	Avg. difference ltr./day	Self-reported increase (p̂)
Raw results					
Program Mean	0.42	0.27	3.63	0.39	0.52
Comparison Mean	0.21	0.10	3.37	0.17	0.52
Unadjusted dif.	0.21***	0.17***	0.26*	0.22**	0.077**
	(0.030)	(0.027)	(0.14)	(0.10)	(0.033)
Observations	**932**	**932**	**932**	**932**	**928**
OLS (ITT)					
Coefficient	0.22***	0.20***	0.30***	0.27***	0.082**
	(0.11)	(0.11)	(0.10)	(0.10)	(0.086)
Observations	**929**	**929**	**929**	**929**	**925**
2SLS (LATE)					
Coefficient	0.27***	0.23***	0.36***	0.33***	0.10**
	(0.13)	(0.14)	(0.12)	(0.12)	(0.10)
Observations	**929**	**929**	**929**	**929**	**925**

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated with program area (p = <0.1) used in all models; VSZ dummies used as fixed effects.

significant at the 10% level. The OLS and LATE estimates for both measures are also identical. This is unsurprising because both measures were constructed by regressing the assets reportedly owned in 2016 on the 2016 consumption expenditure data. The results presented in the second to last column are derived from single difference models that control for the 2007 consumption expenditure weighted index, while the results presented in the last column are these same two indices but differenced. Both sets of estimates range from $0.10 (ITT estimate) to $0.13 (LATE estimate) per day per capita, equivalent to 2.1% and 2.9% of daily household consumption expenditure per capita, respectively.

We complemented our analysis of the 2016 consumption expenditure and consumption weighted asset measures with several other asset measures. The results are presented in Table 8. The program effect estimates are more robust and consistent for the asset gain index and overall raw asset score. One likely explanation for this is that both narrow in on gains in asset ownership over the two time periods and hence are more sensitive to picking up such gains. However, these effect sizes are still modest, e.g., d = 0.0037 for the differenced overall PCA measure. By examining the second half of the table, we see that the positive results in favor of the program area are not driven by any specific asset class.

Our subgroup analysis revealed a surprising finding: the effect estimates for women for all the variables presented in Table 8 are highly statistically significant (p < 0.001), with the opposite being the case for households with male participants. We implemented Wald tests to see if these effect estimates for households represented by female and male participants are statistically different from 0. While the difference is just below the 10% level for the asset gain measure (p = 0.116), the two single-difference PCA measures and the overall raw asset measure are below the 5% level. We implemented both robust and median quantile regression models to confirm the robustness of these results.[10] Fig. 6 visually illustrates the differences in the distributions associated with our asset gain measure. In general, male participant households are bet-

[10] Robust regression gives less weight to influential observations, while median quantile repression compares the median, rather than mean, values between the treatment groups. Using either is a useful robustness check to ensure that results are not being driven by influential observations and apply to bulk of the statistical distribution.

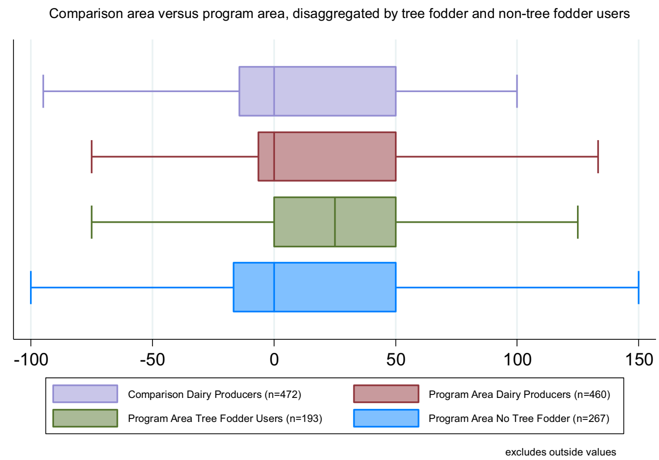


Fig. 5. Box Plots for % Change in Avg. Milk Yields from Baseline.

Table 7
Comparison of program and comparison areas vis-à-vis 2016 HH consumption expenditure and asset indices weighted by 2016 consumption expenditure data.

	2016 consumption expenditure per capita (USD; PPP)	2016 consumption expenditure per capita asset weighted index (USD; PPP)	Differenced 2016/ 2007 consumption expenditure per capita weighted index (USD; PPP)
Raw results			
Program Mean	4.61	5.01	1.00
Comparison Mean	4.55	4.83	0.91
Unadjusted dif.	0.057	0.18*	0.086
	(0.11)	(0.099)	(0.056)
Observations	**2797**	**2797**	**2797**
OLS (ITT)			
Coefficient	0.067	0.10*	0.10*
	(0.010)	(0.06)	(0.062)
Observations	**2790**	**2790**	**2790**
2SLS (LATE)			
Coefficient	0.085	0.13	0.13*
	(0.13)	(0.15)	(0.079)
Observations	**2790**	**2790**	**2790**

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated with program area (p = <0.1) used in all models; VSZ dummies used for fixed effects in all models; both single difference models also control for our 2007 predicted consumption expenditure measure described in Appendix 3.

ter off, but those represented by women in the program area gained more than their counterparts in the comparison area. The difference is most striking with respect to the median differences (represented by the middle line within the boxes) between the female represented households in the program and comparison areas.

5.8. Is agroforestry uptake responsible?

Consistent with our theory of change for Vi Agroforestry's program, we have shown that relatively greater agroforestry adoption took place in the program area vis-à-vis the comparison area. We will now explore the extent to which this adoption may be responsible for our estimated gains in asset holdings brought about by Vi Agroforestry's program.

We used two analytical methods to explore this: statistical mediation analysis (MacKinnon, 2008) and the control function approach (Wooldridge, 2015). The former was implemented with Stata's *sem* (structural equation modelling [SEM]) command. Here, we explored the extent to which variation in the data are consistent with the hypothesis that agroforestry adoption was significantly responsible for our estimated asset accumulation effects. It is important to note that SEM, as a tool for undertaking mediation analysis, does not prove that the variables are causally related in the way they are specified in the model and certainly does not evidence the direction of the causal relationship. Nevertheless, significant mediation estimates can increase confidence in the veracity of a hypothesized mechanism.

Our results (Table 9) reveal that a significant, albeit far from complete, share of our asset accumulation effect estimates can be explained by the uptake of agroforestry. All the indirect effect estimates are statistically significant ($p < 0.01$), revealing that the variation in the data is consistent with the hypothesis that agroforestry adoption was at least partially responsible for our estimated asset accumulation effects.

That said, there is still significant variation—over 60–80% for the three measures—that is unaccounted for by the AF index. Measurement error may, of course, be partly to blame. The AF index, constructed from 10 binary variables, does not precisely measure agroforestry uptake. It is also true that Vi Agroforestry's program comprised other elements, most notably the promotion of other complementary SALM practices and microenterprise development. We measured the uptake of these aspects of Vi Agroforestry's program and found only negligible differences between households in the program and comparison areas (see Appendix 4). The uptake of SALM practices, therefore, is unlikely to account significantly for the unexplained variation. Nevertheless, our study evaluated the intermediate and longer-term effects of Vi Agroforestry's *program*, rather than *agroforestry adoption* itself. Hence, it is possible that unmeasured aspects of this program may have induced effects that we failed to account for.

We may be more convinced that agroforestry uptake was responsible for our estimated asset accumulation effects if these estimates only apply to those households that took it up—*all else being equal*. This brings us to our second method for assessing the extent to which agroforestry adoption was responsible: the control function approach. Table 10 compares high and low adopters in the program area, respectively, to all households in the comparison area.[11] The models control for all our covariates, as well as the SALM and microenterprise practice indices. As is clear, the coefficients for the high adopters are both statistically significant and significantly larger than those of the low adopters. They are also between 59% and 100% larger than the average OLS effect estimates presented in the above tables.

It is indeed possible that we omitted and, hence, did not adequately control for one or more variables correlated with both our AF index and asset accumulation measures. Thus, the above comparison between high and low adopters may be biased. Consequently, we attempted to control for this potential unobserved bias using the control function approach (also referred to as two-stage residual inclusion). The first stage involved predicting both high and low agroforestry adoption vis-à-vis our dataset's various covariates (see Appendix 2). We obviously could not predict either high or low agroforestry adoption perfectly; inevitably, there was much left unexplained, represented by the models' error terms. Hence, we constructed control functions using inverse Mills ratios for both the high adopters and low adopters separately to control for this unexplained variation. The second stage, then, involved including these control functions in the same OLS models

[11] To be coded as a high adopter, the respondent's household had to have gained at least 0.1 points on our 0–1 AF index and/or have gained an average increase of 10% tree cover (measured via remote sensing) on the plots of their main parcels.

Table 8
Household asset wealth and accumulation.

	PCA Asset Indices		Raw Asset Positive Gains				
	2016 index	Asset gain index	Overall raw asset score	House Char.	Home Durables	Productive Assets	Livestock
Raw results							
Program Mean	2.05	0.75	8.78	1.32	3.95	1.76	1.75
Comparison Mean	1.96	0.69	8.26	1.17	3.81	1.67	1.61
Unadjusted dif.	0.087*	0.063***	0.52***	0.15**	0.14	0.099	0.13**
	(0.046)	(0.022)	(0.19)	(0.067)	(0.095)	(0.061)	(0.059)
Observations	**2797**	**2797**	**2797**	**2797**	**2797**	**2797**	**2797**
OLS (ITT)							
Coefficient	0.050*	0.069***	0.59***	0.16**	0.17*	0.14**	0.12*
	(0.026)	(0.025)	(0.21)	(0.071)	(0.10)	(0.070)	(0.068)
Observations	**2790**	**2790**	**2790**	**2790**	**2790**	**2790**	**2790**
2SLS (LATE)							
Coefficient	0.064*	0.087***	0.75***	0.20**	0.22*	0.18**	0.15*
	(0.033)	(0.032)	(0.27)	(0.091)	(0.13)	(0.088)	(0.086)
Observations	**2790**	**2790**	**2790**	**2790**	**2790**	**2790**	**2790**

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group; covariates correlated with program area (p = <0.1) used in all models; VSZ dummies used for fixed effects in all models; 2007 PCA derived asset index controlled for in single difference models for the 2016.

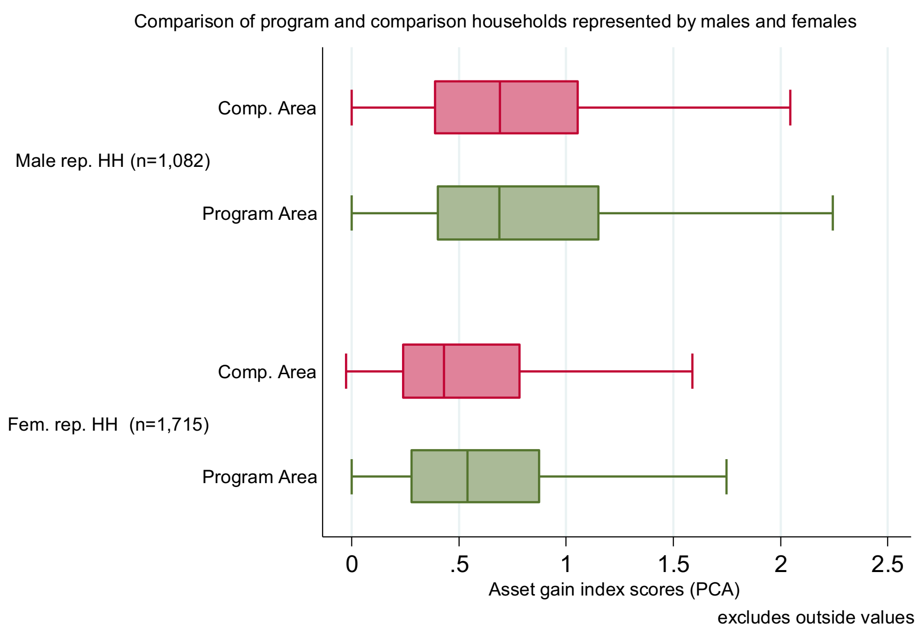


Fig. 6. Box Plots for Asset gain Index.

Table 9
Results of mediation analysis—asset effect estimates through AF Index and tree cover.

(1)	(2) Total Effect (equivalent to OLS estimates)	(3) Direct Effect (independent effect of program area)	(4) Indirect Effect (effect of program area + AF index)	(5) % mediated via AF index
Differenced Consump. Weighted Index	0.115*	0.073	0.043***	37%
	(0.059)	(0.059)	(0.010)	
PCA 2016 Asset Index	0.052**	0.035	0.017***	32%
	(0.026)	(0.026)	(0.005)	
Differenced Asset Gain Index	0.068***	0.055**	0.013***	19%
	(0.025)	(0.025)	(0.004)	

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses and clustered at farmer group level; covariates correlated (p < 0.1) with both our program area dummy and the AF index used in all models; models are all saturated, i.e., they were constructed to perfectly reproduce all variances, co-variances, and means.
Column 2—Total Effect—corresponds to the overall program effect estimated by the mediation models. These estimates are like those of the OLS models presented in Tables 7 and 8, with minor differences given the differing nature of the modelling approaches. Each total effect estimate is then subdivided into the Direct Effect and Indirect Effect. The direct effect estimates (column 3) correspond to the extent to which our program area dummy independently accounts for the total effect estimates. The indirect effect estimates (column 4), on the other hand, correspond to the extent to which the variation shared by *both* the program area dummy and the differenced AF index accounts for these estimates. The percentage of the total effect mediated (column 5), then, corresponds to the percentage of the total effect that is accounted for by the variation our program area dummy shares with our difference AF index measure (as a measure of the hypothesized mediator variable).

Table 10
Associations between AF Adoption and household asset accumulation variables.

	Consumption weighted differenced asset index	2016 PCA Asset Index	Asset Gain Index
OLS (High adopters)			
Coefficient	0.20***	0.090***	0.11***
	(0.068)	(0.029)	(0.028)
Observations	**2072**	**2072**	**2072**
OLS (Low adopters)			
Coefficient	−0.0097	−0.00067	0.019
	(0.072)	(0.029)	(0.028)
Observations	**2096**	**2096**	**2096**
OLS (High adopters with control functions)			
Coefficient	0.18***	0.087***	0.10***
	(0.068)	(0.029)	(0.028)
Observations	**2066**	**2066**	**2066**
OLS (Low adopters with control functions)			
Coefficient	−0.037	−0.0020	0.013
	(0.072)	(0.029)	(0.028)
Observations	**2096**	**2096**	**2096**

Standard errors in parentheses and clustered at farmer group level; *p < 0.1, **p < 0.05, ***p < 0.01.
All covariates in dataset used models, as well as SALM and microenterprise participation indices with VSZ dummies as fixed effects.

presented in the first section of Table 10. While the bias mitigation potential of this approach depends on a number of assumptions (Greene, 2003), this helped us to control for potential unobservable differences between the high and low adopters. As is clear in Table 10, while the coefficients for the high adopters are downgraded slightly with the introduction of the control functions, they remain statistically significant and much larger than those of the low adopters.

6. Discussion and conclusion

This study addresses a key gap in the literature on the longer-term effects of agroforestry promotion programs and integrated agroforestry systems. This evidence gap has likely persisted to date simply because generating rigorous impact evidence is particularly challenging when it comes to agroforestry: the timeframe within which such impacts are expected to manifest are typically long and non-linear; there is no one particular agroforestry system suitable for all agroecological, social, and economic contexts; agroforestry is generally taken up by farmers with varying levels of intensity, not as a binary 'technology'; and agroforestry promotion tends to be bundled with the promotion of other agricultural and NRM practices. Bearing these inherent challenges in mind, we took advantage of an effort led by Vi Agroforestry to promote agroforestry in two counties in western Kenya, Bungoma and Kakamega. Here, we implemented a quasi-experimental, mixed-methods framework to evaluate selected intermediate and longer-term impacts vis-à-vis Vi Agroforestry's theory of change.

We found that participation in Vi Agroforestry's program brought about modest gains in household asset accumulation, particularly among female represented households. While in no way conclusive, our supplementary mediation analysis indicates that the relatively greater uptake of agroforestry in the program area was likely at least partly responsible. In consumption expenditure per capita terms, our estimated average 'asset accumulation effect' on Vi Agroforestry affiliated households (LATE estimate) is USD $0.13. This is just under $50 per year per capita or about 3% of overall household consumption expenditure. This is not huge, transformative impact, but it should not be entirely dismissed either. Moreover, these average figures mask the variation experienced across households, as illustrated by the box plots presented in Fig 7.

Given that we captured data along several key dimensions of Vi Agroforestry's program theory of change, we offer possible insights why this downstream impact failed to materialize in a more significant way. Specifically, while we found that the program moved several intermediate outcomes, such as income from the sale of

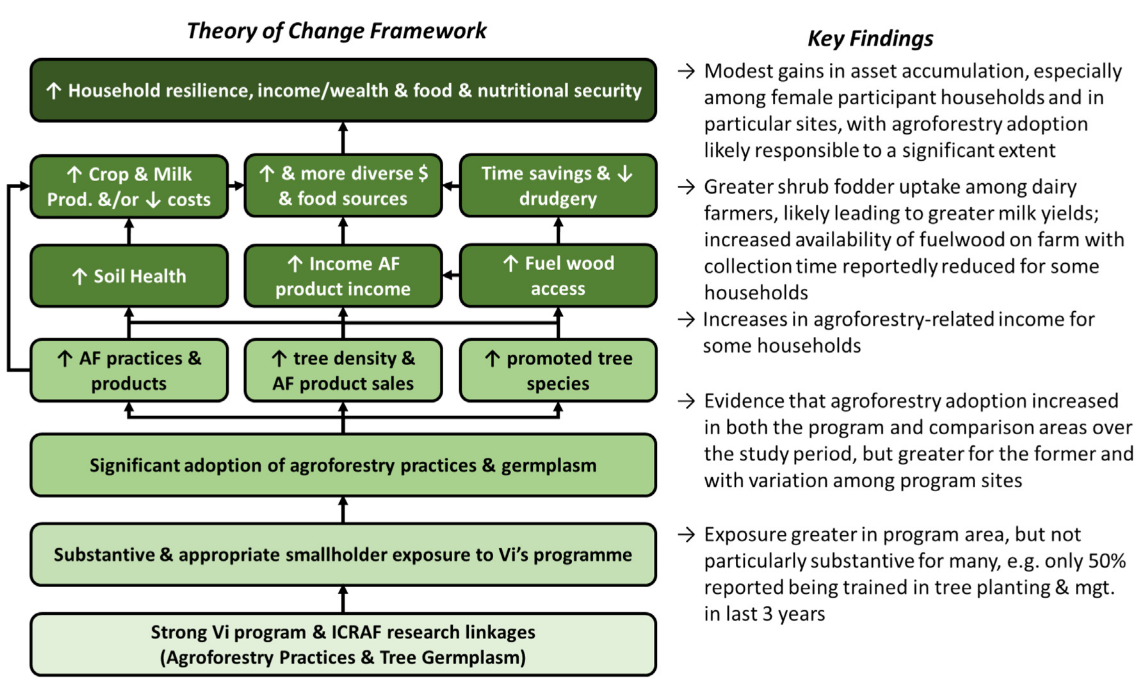


Fig. 7. Summary of key Findings along Theory of Change.

agroforestry products, fuelwood access, and milk yields, for many households, it certainly did not do so for all. For example, only just over a third of households in the program area reported agroforestry sales greater than $10 in the previous year. Moreover, while milk yields did increase to a greater extent for dairy producers in the program area, only about half self-reported experiencing such an increase. And when we move further down the causal chain to agroforestry adoption, as measured by our AF index, we see a similar trend. There was certainly an increase from the baseline period but nothing particularly dramatic; the gain for Vi Agroforestry affiliated households was 0.13 on the index that ranges from 0 to 1. Similarly, as we move even further down the causal chain to agroforestry extension exposure, the rates of such exposure were significantly higher in the program area. Yet, again, this does not appear to be the case for many households.

It is difficult to know with certainty whether the result would have been different if the uptake of agroforestry had been more substantive among the farmer groups targeted by Vi Agroforestry, particularly considering that such uptake also took place in the comparison area (albeit to a more limited extent). However, we have reasons to believe that it may very well have been: We found that asset accumulation took place to a much greater extent among 'high adopters' vis-à-vis 'low adopters' in the program area, even after undertaking efforts to control for potential sources of unobserved bias.

The mixed-results we found for agroforestry adoption is consistent with a broader finding in the literature: the uptake of research informed natural resource management (NRM) 'innovations' among smallholder farmers in low and middle income countries is generally low (Barrett, Place, Aboud, & Brown, 2002; Stevenson et al., 2019). Possible reasons include, of course, that farmers simply do not find such innovations particularly useful, cost-effective (particularly vis-à-vis the implicit discount rate), and/or well match to their idiosyncratic conditions (Coe, Sinclair, & Barrios, 2014). However, recent literature on the promotion of agricultural and NRM innovations points to other possible explanations. The 'pipeline model', where 'discovery' research leads to the development and testing of an innovation, followed by its 'piloting' and subsequent 'release' for large-scale 'scaling', is deemed too simplistic, even for 'simple' innovations, such as improved crop varieties (Wigboldus et al., 2016). Substantive modification to the innovation in question often does, and indeed often should, take place during the 'scaling phase'. This dissatisfaction has led to calls for more iterative and adaptive approaches for the scaling of research informed agricultural innovations (Cleaver, 2012).

Development and research-for-development organizations are under increasing pressure to 'deliver' more impact for larger numbers of 'beneficiaries' and with fewer resources. This lies in tension with these insights into the scaling process, particularly on the need for ongoing engagement, iteration, adaptation, and addressing issues in the wider system. It is noteworthy that Vi Agroforestry moved away from its more intensive extension approach in 2004 to the farmer group intervention model on which this study is based. This was, in part, for perceived cost-effectiveness reasons. While there are certainly merits associated with Vi Agroforestry's current implementation model (e.g., it potentially fosters less dependency), it may have moved too far away from its previous approach where there was more direct and ongoing engagement with farmers.

In an era where social and environmental impact is expected on a larger scale and at lower costs under the banner of 'value-for-money', this is certainly worthy of reflection. Many of the more complex and transformational changes we seek to bring about in the agriculture and NRM sectors are unlikely achievable through superficial training, extension visits, or input distributions. To ensure such impact at scale, there is need to re-think how we pursue the actual process of scaling.

Declaration of Competing Interest

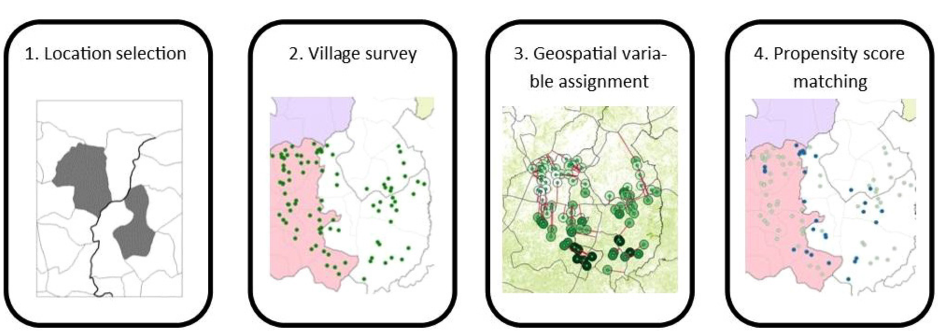
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our appreciation to the Vi Agroforestry Kenya field team for the great support provided, both in the formative work and data collection associated with this study. The Consultative Group for International Agricultural Research's (CGIAR's) Standing Panel on Impact Assessment (SPIA) was also instrumental in providing mythological and analytical guidance through various peer review processes. This research was supported by ISPC-SPIA under the grant "Strengthening Impact Assessment in the CGIAR System (SIAC)"

Appendix 1. Village matching using spatial and secondary data

The execution of the program area and comparison area village selection process followed a four-step procedure: (1) sub-location selection; (2) scoping survey administration; (3) geospatial variable assignment; and (4) propensity score matching.



Sub-location selection: Qualitative interviews were conducted to guide a purposive selection of sub-locations (the smallest administrative unit in Kenya). These sublocations were selected from Vi Agroforestry's program area and the comparison area based on their similarity in terms of relative wealth and agroecological characteristics. These interviews were carried out with Vi Agroforestry field staff, Kenya Ministry of Agriculture field officers, and farmer group leaders. Once these interviews were carried out, the sample was restricted to these purposively matched sublocations.

Scoping survey administration: A scoping survey was administered within the villages in the purposively matched sublocations. Local consultants were hired to visit the villages where they administered a short survey to key informants and captured the GPS coordinates of the center of the village. The scoping survey instrument was implemented using the Open Data Kit (ODK) platform. It captured: (a) the number of households in the village; (b) whether there were active farmer groups which have been active from the baseline period onwards; and (c) the activities of each group and their receipt of NGO or government support, if any.

Geospatial variable assignment. The village GPS coordinates captured during the scoping survey were used to determine the nine geospatial variables listed below. A 1 km buffer was generated around each village's central geocode and the Zonal Statistics tool in ArcGIS was used to calculate the average value across the raster grids containing each variable's values. The village average values were then assigned to the dataset of village names using the Extract Multi Values by Points tool in ArcGIS.

Propensity Score Matching (PSM). The geospatial variables presented below were used to assign a propensity score, and the villages were matched on this score using *psmatch2*'s one-to-one caliper matching algorithm in Stata. Applying this method to villages allowed us to compare households in villages within the program area to household in villages outside the program area, which are similar across the range of relevant covariates. Matching assumes that average treatment effects can be estimated by taking the average of the difference between the expected outcome of untreated observations—conditional on a vector of covariates—from the expected outcome of the treated observations conditional on the same covariates (Abadie & Imbens, 2016).

This methodology has been used to estimate the effects of protected areas (Andam et al., 2008). This literature includes examples of matching on observational units at multiple scales, including pixels in a raster grid (Andam et al., 2008; Robalino, Pfaff, & Villalobos, 2015), polygons corresponding to land management units (Honey-Rosés, Baylis, & Ramírez, 2011), and census tracts (Andam, Ferraro, Sims, Healy, & Holland, 2010). Our analysis focuses on a 1 km circular buffer drawn around each village considered for the study.

Matching designs in the literature use two primary methodologies for assessing observations' similarity across covariates: nearest neighbor and propensity score matching (Joppa & Pfaff, 2010). Nearest neighbor matching calculates the multi-dimensional distance between two observations given the vector of covariates. Propensity score matching condenses the covariates to a single score using a regression model to calculate each observation's conditional probability of receiving treatment given the covariate values (Rosenbaum & Rubin, 1983). Propensity score matching was identified as the most appropriate method, since the objective was to identify comparison villages with a high conditional probability of being included in the program area given the measured geospatial variables.

The propensity score is defined formally as the probability of treatment, conditional on a vector of covariates. The propensity score model can be expressed by the equation below:

$$e(X_i) = \Pr(T_i = 1|X_i)$$

where $e(X_i)$ is the probability of being included in the treatment group, and X_i is the vector of covariates listed above. The propensity score was generated using a probit model estimated within Stata by the psmatch2 command. The probit model takes the form:

$$z = X\beta + \varepsilon$$

where z is an unobserved variable, X is a vector of covariates with coefficients β, and y is the observed binary corresponding to treatment assignment such that:

$$y_i = \begin{cases} 1 if z \geq 0 \\ 0 if z \leq 0 \end{cases}$$

Due to the limited time before enumeration was scheduled to begin, the scoping process was completed in four phases. In particular, the study area was split into four zones, which we refer to as Village Sampling Zones (VSZs), and the matching process was applied to each one separately as the scoping data were collected. The targeted number of villages in each VSZ was set as 15 program area villages and 15 comparison area villages. Consequently, for each VSZ, the 30 best-matched villages were chosen by gradually reducing the caliper width using Stata's *psmatch2* command. Covariate balance was thereafter tested on the set of matching variables between the program and comparison villages for each VSZ. Then, as the team of enumerators began collecting data in the matched villages in the VSZ in question, the scoping team moved forward to the next VSZ and the PSM village matching exercise was subsequently undertaken for this particular VSZ.

Villages with farmer-led groups active since the beginning of the study period (2008–2016) were considered candidates for inclusion in the study. Using geocodes from the center of each villages, geospatial variables were assigned to each of these shortlisted villages for input into the matching model. The matching covariates included agroecological characteristics, as well as socio-economic indicators, such as population density and distance from major roads. We chose these particular covariates because we assume they are likely to affect agricultural production and market access, and, as such, they would likely be significant confounders of the measured treatment effect if the selected villages were unbalanced across them. The following variables, in particular, were used in the propensity score model:

- Number of Households
- Average Soil Sand Content (Vågen, Winowiecki, Tondoh, Desta, & Gumbricht, 2016)
- Average Soil pH (Vågen et al., 2016)
- Average Soil Organic Carbon in 2007 (Vågen et al., 2016)
- Average Tree Cover in 2005 (Sexton et al., 2013)
- Elevation (Jarvis et al., n.d.)
- Average Population Density in 2010 and 2015 (Stevens, Gaughan, Linard, & Tatem, 2015)
- Average Rainfall (Funk et al., 2015)
- Distance to Tarmac Road
- Binary for Villages 0.25 m from Tarmac Road ("on road")
- Binary for presence of microfinance activities

Elevation, tree cover, population density, and soil variables were measured as an average value calculated across a circle 1 km in radius extending from a central point in the village. Rainfall was measured as the value of the raster cell in which the village center was found. The cells for the Climate Hazards Infrared Precipitation with Stations (CHIRPS) rainfall dataset measure approximately 5.5 km across (Funk et al., 2015).

Household numbers were taken from the village-level scoping survey. The consultants requested the number of households from leaders of farmer groups, and, if they did not know this number, they requested it from a village elder. The tarmac road network was taken from OpenStreetMap data, and ground-truthed by travel in the region. The binary variable for the presence of microfinance activities was taken from Vi's records on their participating groups and from the scoping survey for the comparison villages. Activities listed as "table banking" or "merry-go-round" were counted as microfinance activities.

After each VSZ specific matching exercise, overall covariate balance was checked to confirm that the overall village sample of treatment and comparison villages are balanced across all selected covariates, as is presented in Table 1.

Table A1: Village-level Matching Balancing Statistics

	Sample Mean	Program Mean	Comparison Mean	Normalized Difference	Difference
Soil Sand Content	19.96	20.57	19.36	0.11	1.21 (1.42)
Soil pH	5.95	5.97	5.94	0.20	0.03 (0.02)
Tree Cover 2005	6.07	5.97	6.17	−0.06	−0.21 (0.45)
Elevation	1570.52	1575.63	1565.49	0.05	10.13 (26.18)
Population Density 2010	4.41	4.40	4.43	−0.02	−0.03 (0.22)
Soil Organic Carbon 2007	25.57	24.71	26.43	−0.19	−1.72 (1.16)
Rainfall	136.71	133.97	139.40	−0.23	−5.42 (2.90)
Distance to Road	0.03	0.03	0.03	0.07	0.00 (0.00)
On Road	0.02	0.02	0.03	−0.07	−0.02 (0.03)
Observations	**121**	**60**	**61**		

*p < 0.1, **p < 0.05, ***p < 0.01; standard errors in parentheses.

Appendix 2 Baseline and time invariant characteristic comparisons

A2.1: Comparison of HHs in Program and Comparison Areas—Binary Characteristics

Characteristic	Program Mean ($\hat{p}$)	Comparison Mean ($\hat{p}$)	Difference (raw)	Difference (net of county)	Difference (net of VSZ)
Respondent Female	0.63	0.60	0.031* (1.70)	0.082* (1.69)	0.082* (1.69)
Respondent Christian	0.99	0.99	−0.0027 (−0.91)	−0.16 (−0.92)	−0.17 (−0.98)
Respondent Luhya	0.97	0.98	−0.0095 (−1.49)	−0.14 (−1.48)	−0.16 (−1.64)
Respondent Married	0.75	0.73	0.017 (1.01)	0.050 (0.98)	0.054 (1.06)
Res. Married by Spouse Elsewhere	0.08	0.08	0.0036 (0.35)	0.021 (0.31)	0.012 (0.18)
Respondent Widowed	0.14	0.16	−0.018 (−1.32)	−0.073 (−1.25)	−0.077 (−1.31)
Respondent is Head	0.48	0.54	−0.057*** (−3.01)	−0.14*** (−3.02)	−0.14*** (−3.03)
Respondent is Spouse of Head	0.41	0.37	0.038** (2.06)	0.099** (2.06)	0.100** (2.07)
Respondent Good Health	0.97	0.98	−0.0089 (−1.51)	−0.15 (−1.50)	−0.16 (−1.52)
Respondent Literate	0.89	0.89	0.00059 (0.049)	0.0039 (0.06)	0.0059 (0.09)
Respondent in School	0.02	0.03	−0.0033 (−0.55)	−0.057 (−0.57)	−0.070 (−0.69)
Respondent has Official Role	0.49	0.46	0.027 (1.40)	0.067 (1.40)	0.070 (1.48)
Respondent Farmer (main occupation)	0.87	0.86	0.017 (1.29)	0.078 (1.30)	0.079 (1.33)

(*continued*)

Characteristic	Program Mean ($\hat{p}$)	Comparison Mean ($\hat{p}$)	Difference (raw)	Difference (net of county)	Difference (net of VSZ)
Respondent Primarily Business	0.05	0.06	−0.0067 (−0.77)	−0.061 (−0.80)	−0.059 (−0.77)
Respondent Employed	0.05	0.06	−0.013 (−1.56)	−0.12 (−1.56)	−0.13 (−1.63)
Respondent has Tech. Skills	0.06	0.04	0.020** (2.30)	0.18** (2.31)	0.18** (2.29)
HH was farming in 2007	0.99	0.98	0.0046 (0.99)	0.13 (1.02)	0.12 (1.01)
HH reared livestock in 2007	0.62	0.57	0.055*** (2.97)	0.14*** (2.97)	0.14*** (2.92)
HH ran off-farm business in 2007	0.30	0.31	−0.00045 (−0.026)	−0.0015 (−0.03)	−0.0089 (−0.18)
HH member employed in 2007	0.16	0.13	0.030** (2.24)	0.13** (2.21)	0.12** (2.08)
HH used irrigation in 2007	0.05	0.04	0.0077 (0.96)	0.078 (0.96)	0.086 (1.04)
Head is productive	0.97	0.97	−0.0066 (−1.04)	−0.10 (−1.06)	−0.11 (−1.11)
all adults in HH over 59	0.06	0.05	0.0047 (0.54)	0.041 (0.53)	0.042 (0.54)
Head is 60 or older	0.33	0.29	0.043** (2.46)	0.12** (2.46)	0.12** (2.42)
female headed HH	0.22	0.22	−0.0040 (−0.25)	−0.011 (−0.20)	−0.015 (−0.28)
Literate adult in HH	0.98	0.97	0.0054 (0.93)	0.095 (0.93)	0.095 (0.93)
Female literate adult in HH	0.89	0.89	0.0026 (0.22)	0.013 (0.20)	0.011 (0.17)
Head is literate	0.90	0.90	−0.0039 (−0.34)	−0.022 (−0.34)	−0.020 (−0.30)
HH located on tarmac road	0.04	0.05	−0.010 (−1.33)	−0.11 (−1.32)	−0.11 (−1.24)
HH had formal title to main parcel (07)	0.35	0.32	0.030* (1.67)	0.082* (1.66)	0.078 (1.58)
Respondent owned main parcel (07)	0.46	0.54	−0.084*** (−4.44)	−0.21*** (−4.44)	−0.21*** (−4.40)
Observations	**1411**	**1386**	**2797**	**2797**	**2797**

z statistics in parenthesis; VSZ = Village Sampling Zone; *p < 0.1, **p < 0.05, ***p < 0.01.Probit regression used for net of county and VSZ differences, so coefficients are not directly interpretable, only the t-statistics.

A2.2: Comparison of HHs in Program and Comparison Areas—Continuous Characteristics

Characteristic	Program Mean	Comparison Mean	Difference (raw)	Difference (net of county)	Difference (net of VSZ)
Age of respondent	49.65	48.84	0.81 (1.60)	0.80 (1.60)	0.76 (1.50)
Years of education of respondent	8.55	8.61	−0.055 (−0.41)	−0.056 (−0.41)	−0.061 (−0.45)
Household size	6.53	6.65	−0.13 (−1.36)	−0.13 (−1.35)	−0.12 (−1.30)
Number of children in HH	3.56	3.71	−0.15** (−1.99)	−0.15** (−1.99)	−0.15* (−1.91)
Number of adults in HH	2.97	2.94	0.025 (0.46)	0.025 (0.46)	0.024 (0.44)
Years of education head	9.08	9.20	−0.12 (−0.86)	−0.12 (−0.86)	−0.12 (−0.88)
Highest years of educ. of any adult in HH	10.72	10.94	−0.23* (−1.86)	−0.23* (−1.87)	−0.23* (−1.92)
Number of productive adults in HH	3.52	3.49	0.022 (0.34)	0.022 (0.34)	0.020 (0.31)
Land Size at Baseline	2.30	2.36	−0.077 (−0.87)	−0.077 (−0.87)	−0.076 (−0.87)
2007 asset index (2016 expend. weighted) consumption expenditure data	13.07	12.96	0.11 (0.52)	0.11 (0.51)	0.077 (0.36)
2007 asset index (PCA weighted)	1.54	1.49	0.056 (1.38)	0.055 (1.39)	0.048 (1.21)
Estimated 2007 soil organic carbon (plot avg.)	24.38	22.97	1.42*** (4.63)	1.40*** (5.12)	1.36*** (6.12)
Estimated 207 soil erosion (plot avg.)	0.36	0.36	0.00055 (0.18)	0.00044 (0.15)	0.00021 (0.07)

(*continued on next page*)

(*continued*)

Characteristic	Program Mean	Comparison Mean	Difference (raw)	Difference (net of county)	Difference (net of VSZ)
Elevation (hh level)	1559.85	1529.87	29.6***	29.4***	−0.027
			(3.66)	(4.09)	(−0.61)
HH distance from tarmac road (km)	3.08	2.96	0.12*	0.12*	0.12*
			(1.67)	(1.68)	(1.77)
Observations	1411	1386	2797	2797	2797

z statistics in parenthesis; VSZ = Village Sampling Zone.*p < 0.1, **p < 0.05, ***p < 0.01.

Appendix 3. Asset measures used for both predicting consumption expenditure and constructing complement asset indices

Household consumption expenditure data were collected by incorporating several modules into the household survey. The respondents were asked, for instance, the types of food their households consumed over the previous seven-day period, as well as the specific quantities. These quantities were then converted into monetary value. This was done by asking the respondent how much was paid for each food item or, if the food item was sourced through the household's own production, how much it would cost if purchased from the local market. The respondents were also asked how much they spent on specific non-food items and services from a detailed list, such as soap, toothpaste and mini-bus fares, over the past four weeks (regular non-food expenditure). Finally, they were asked about particular 'big ticket' expenditures over the previous 12 months from another pre-defined list, such as school and hospital fees, clothes and home repairs (irregular non-food expenditure).

We then computed the basic per capita measure as follows for each household:

- The weekly cash value of each food item consumed during the past seven days were added together and divided by seven, thereby estimating the daily cash value of food consumed by the household.
- Household expenditure on items from both the regular monthly non-food expenditure list and annual non-food expenditure list were added together and divided by 30.42 and 365, respectively, thereby estimating the household's average daily expenditure on regular and irregular non-food items.
- The daily consumption expenditure estimated for food and the regular and irregular non-food items were then added together and converted into US dollars, while adjusting for PPP.[12]
- Finally, to derive each household's per capita consumption expenditure, its PPP adjusted dollar value was divided by the number of its members (household size), with another adjustment made for assumed lower consumption among children and economies of scale.[13]

A key limitation to using consumption expenditure data in our ex-post study is that reconstructing baseline data for difference-in-differences estimation was infeasible. We therefore sought to overcome this through the construction of two sets of asset indices where both current and recalled asset ownership data were collected through the household survey. The first approach involved weighting each asset item based on the extent to which it is correlated with consumption expenditure, with the resulting coefficient estimates for each asset used as weights. Such consumption regression approaches attempt to predict consumption when complementary consumption data are available (O'Donnell et al., 2008).

To implement this approach, we first took the binary set of 78 asset measures for 2016 (Table A3.1) and regressed them on the abovementioned 2016 daily household consumption data. We then removed those assets negatively correlated and/or only modestly correlated with consumption expenditure (i.e. $p > 0.2$). This enabled us to arrive at a list of binary assets measures for the 2016 period that are all positively and moderately correlated with 2016 consumption expenditure. This set explains over 40% of the variation of the consumption data ($R^2 = 0.4217$). We then took this same set of assets owned during the 2007 period and gave each the same weights. Finally, we divided each index by the adjusted household size measure described in footnote 13. This study's primary outcome measure, which is declared in our pre-analysis plan, takes the difference in these two consumption weighted indices between the baseline and endline periods.

We also constructed several other asset measures via PCA and what is dubbed the 'arbitrary' or 'naïve' approach (O'Donnell et al., 2008). The latter involves simply adding together the asset binary measures without differentially weighting them. We implemented the latter in part as a robustness check and in part to explore whether any differences between the program and comparison areas were specific to an asset class, e.g., housing characteristics or livestock ownership.

To construct the PCA measures, we first took the binary asset for the baseline and endline periods and assessed their inter-item correlation and removed those negatively correlated with the other assets. The resulting inter-item correlation that resulted was quite high (alpha = 0.9025 and 0.8917 for the 2016 and 2007 binary asset measures, respectively).[14] We then constructed tetrachoric matrices with them, and principal component factor analysis (PCA) was then run on these matrixes. Variables based on the first principal component were subsequently constructed. We did this for each period for the overall dataset.

Given that implementing PCA for each period separately would generate different time-specific sets of asset weights, we avoided

[12] Adjusting for PPP was done to consider Kenya's idiosyncratic purchasing power, i.e. the quantity of currency required to purchase a given basket of goods and services. The PPP conversion rates used were taken from the World Bank's website: http://data.worldbank.org/indicator/PA.NUS.PPPC.RF.

[13] While dividing the above by household size as the overall denominator is recommended in the literature, it is considered important to avoid underestimating expenditure for larger sized households relative to their smaller counterparts. A recommended formula for computing household size for this purpose is: HH size = $(A + \alpha K)^{\theta}$ where A is number of adults in the household; K is the number of children; α is the cost of a child relative to an adult; and θ controls the extent of economies of scale. For low income countries, is recommend that α be set at 0.25 or 0.33 and θ be set at 0.9 (Deaton & Zaidi, 2002).

[14] When items are used in a scale or index, they should all measure the same underlying latent construct (e.g. household wealth status). The items, therefore, must be significantly correlated with one another. Cronbach's alpha is a measure of this inter-item correlation. The more the variables are correlated, the greater is the sum of the common variation they share. If all items are perfectly correlated, alpha would be 1 and 0 if they all were all independent from one another. For comparing groups, an alpha of 0.7 or 0. 8 is considered satisfactory (Bland & Altman, 1997).

Table A3.1: List of assets and other wealth indicators used to construct asset indices

1. Stove
2. Pots
3. Plates
4. Cutlery
5. Cooking utensils
6. Bed
7. Mattress
8. Rug
9. Sofa
10. Table
11. Chair
12. TV cabinet
13. TV
14. Satellite dish
15. DVD player
16. Radio
17. Mobile phone
18. Computer or laptop
19. Internet
20. Solar electricity
21. Generator
22. Refrigerator
23. Iron
24. Lamp
25. Suitcase
26. Bicycle
27. Motorbike
28. Vehicle
29. Building for renting
30. Irrigation
31. Tractor
32. Plough
33. Cart
34. Wheelbarrow
35. Milling machine
36. Feed mixer
37. Crop cutter
38. Milking can
39. Hoe
40. Axe
41. Sickle
42. Shovel
43. Pick axe
44. Watering can
45. *Panga*
46. Slasher
47. Store
48. Livestock house
49. Improved cook fuel
50. Improved toilet
51. Improved floor
52. Improved walls
53. Improved roof
54. Private water source
55. Glass windows
56. Burglar bars on windows
57. Metal door
58. Metal gate
59. Metal fence
60. Water tank
61. Borehole
62. 3+ rooms
63. Local bull
64. Improved bull
65. Local oxen
66. Improved oxen
67. Local steer
68. Improved steer
69. Local heifer
70. Improved heifer
71. Goat
72. Sheep
73. Donkey
74. Local poultry
75. Improved poultry
76. Pig
77. Dairy cow
78. Improved dairy cow

simply differencing the two indices to obtain a differenced measure. Rather, we first identified whether there had been gains over the two periods for each asset indicator. Then we checked the inter-item correlation again for the resulting set, while iteratively removing negative values. We did this until we arrived at a low but still reasonable inter-item correlation coefficient (alpha) of 0.7343 for the overall dataset. After that, we constructed a tetrachoric matrix and ran PCA on it again, thereby creating an 'asset gain index'.

Appendix 4. Sustainable agricultural land management (SALM) and Micro-enterprise practice indices

Other Sustainable Agricultural Land Management Practices

As discussed in Subsection 2.1, Vi Agroforestry's program also involved promoting other SALM practices that fall outside of a strict definition of agroforestry. Consequently, if the uptake of these practices was significant and significant differences are also found between the program area and Vi Agroforestry groups on the one hand and the comparison area on the other on our study's socioeconomic measures, it may prove difficult to disentangle the relative contribution each may have played.

Plot specific data were therefore also captured on a total of 15 non-agroforestry SALM practices covered in Vi Agroforestry's SALM manual. As was the case for the AF index, to compare the program and comparison areas in an integrated way, we grouped them into the three categories or dimensions presented in Fig A4.1 and created an 'other SALM practice index'. Like the AF index, each practice is weighted equally under each dimension. The results comparing the three groups are presented in Table A4.1.

For the overall index, the groups were at about the same level in the baseline period. However, we found modest, albeit statistically significant differences, for the 2016 index. The differences for differenced 2007–2016 index are statistically insignificant when the program and comparison areas are compared. This is also the case for the crop management and soil and water conservation dimensions, but the difference in favor of the program area for the fertility management dimension is statistically significant. Moreover, the overall difference is significant when Vi Agroforestry households are compared with all households in the comparison area.

While the uptake of agroforestry practices promoted by Vi Agroforestry in the program area in general and among Vi Agroforestry affiliated households in particular was not particularly dramatic, at least overall, it was significantly greater in comparison with the comparison area. It was also considerably more substantial than the other SALM practices that Vi Agroforestry promoted, as is clear from Table A4.1. Consequently, the asset accumulation effects estimated in favor of the program area are unlikely to have been induced by the uptake of the other non-agroforestry related SALM practices promoted by Vi Agroforestry, given that the differential uptake of these practices between the program and comparison areas was minimal.

Micro-enterprise practice

We further constructed a micro-enterprise practice index in a similar way to both the AF index and other SALM practice index. It comprises of seven indicators weighted equally under the following three dimensions: (1) household has off-farm business as livelihood; (2) respondent is active in at least one micro-finance group, i.e. respondent is a member, participates in decision-making to at least a medium extent, and attended a meeting at least four times in the previous 12 months; and (3) respondent has been trained in a micro-finance/business topic and reports having had implemented this training at least to a medium extent. These indices were constructed for both the 2016 and 2007 periods and differenced. We found no difference between households in both the program and comparison areas vis-à-vis the differenced version of this index,

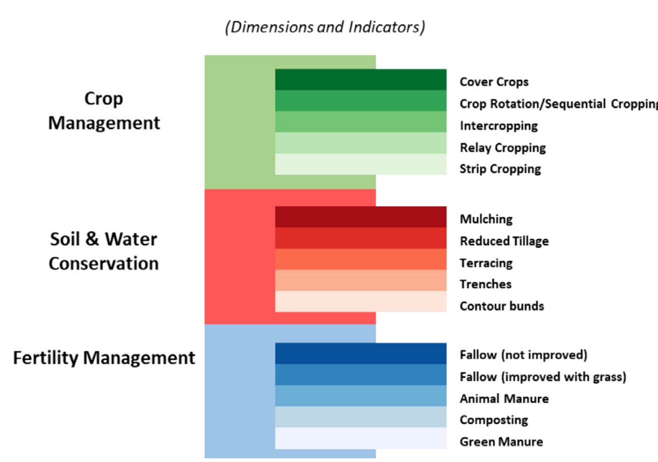


Fig. A4. 1 Index for Other Sustainable Land Management Practices Promoted by Vi Agroforestry.

Table A4.1: Comparison of HHs—SALM practice index and dimensions

Binary Indicators		PA Mean	Vi Group Mean	Non-PA Mean	PA vs. non-PA (dif.)	Vi vs. non-PA (dif.)
Overall Index	2007	0.26	0.26	0.25	0.0047 (1.07)	0.0049 (1.05)
	2016	0.29	0.30	0.28	0.010** (2.37)	0.016*** (3.31)
	Change	0.03	0.04	0.03	0.0058 (1.62)	0.011*** (2.81)
Dimension 1: Crop Management	2007	0.50	0.50	0.52	−0.019** (−2.17)	−0.023** (−2.39)
	2016	0.54	0.54	0.55	−0.016* (−1.82)	−0.013 (−1.45)
	Change	0.04	0.04	0.03	0.0037 (0.51)	0.0095 (1.20)
Dimension 2: Soil & Water Conservation	2007	0.09	0.09	0.07	0.014*** (3.17)	0.018*** (3.69)
	2016	0.11	0.12	0.10	0.014*** (2.69)	0.022*** (3.82)
	Change	0.03	0.03	0.03	−0.00031 (−0.07)	0.0037 (0.76)
Dimension 3: Fertility Management	2007	0.18	0.18	0.16	0.019*** (3.10)	0.019*** (2.98)
	2016	0.22	0.23	0.19	0.033*** (5.07)	0.039*** (5.60)
	Change	0.04	0.05	0.03	0.014** (2.52)	0.020*** (3.33)
	Observations	1411	1094	1386	2797	2480

t statistics in parenthesis; PA = Program Area; *p < 0.1, **p < 0.05, ***p < 0.01.

thereby also making it an unlikely to have contributed to our estimated program effects on asset accumulation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.worlddev.2019.104835.

References

Abadie, B. Y. A., & Imbens, G. W. (2016). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*(1), 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Akinnifesi, F. K., Ajayi, O. C., Sileshi, G., Chirwa, P. W., & Chianu, J. (2010). Fertiliser trees for sustainable food security in the maize-based production systems of East and Southern Africa. A review. *Agronomy for Sustainable Development, 30* (3), 615–629. https://doi.org/10.1051/agro/2009058.

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics, 95*(7–8), 476–487. https://doi.org/10.1016/j.jpubeco.2010.11.006.

Andam, K. S., Ferraro, P. J., Pfaff, A., Sanchez-Azofeifa, G. A., & Robalino, J. A. (2008). Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(42), 16089–16094. https://doi.org/10.1073/pnas.0800437105.

Andam, K. S., Ferraro, P. J., Sims, K. R. E., Healy, A., & Holland, M. B. (2010). Protected areas reduced poverty in Costa Rica and Thailand. *Proceedings of the National Academy of Sciences of the United States of America, 107*(22), 9996–10001. https://doi.org/10.1073/pnas.0914177107.

Angrist, J., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Barrett, C. B., Place, F., Aboud, A., & Brown, D. (2002). The challenge of stimulating adoption of improved natural resource management practices in African agriculture. In Christopher B. Barrett, F. Place, & A. A. Aboud (Eds.), *Natural resource management in Africa: Understanding and improving current practices*. CABI Publishing.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ, 314* (7080), 572. https://doi.org/10.1136/bmj.314.7080.572.

Brown, S. E., Miller, D. C., Ordonez, P. J., & Baylis, K. (2018). Evidence for the impacts of agroforestry on agricultural productivity, ecosystem services, and human well-being in high-income countries: A systematic map protocol. *Environmental Evidence, 7*(1), 24. https://doi.org/10.1186/s13750-018-0136-0.

Bungoma, C. G. (2018). Bungoma country integrated development plan. Retrieved from https://cog.go.ke/media-multimedia/reportss/category/106-county-integrated-development-plans-2018-2022?download=303:bungoma-county-integrated-development-plan-2018-2027.

Buttoud, G., Place, F., & Gauthier, M. (2013). Advancing agroforestry on the policy agenda. Agroforestry Working Paper No.1. Rome. https://doi.org/10.1080/14728028.2013.806162.

Cleaver, K. (2012). Scaling up in agriculture, rural development, and nutrition: Investing in agriculture to reduce poverty and hunger (2020 vision briefs No. 19 (2)). Washington, D.C.

Coe, R., Sinclair, F., & Barrios, E. (2014). Scaling up agroforestry requires research " in " rather than " for " development. *Current Opinion in Environmental Sustainability, 6*, 73–77. https://doi.org/10.1016/j.cosust.2013.10.013.

Deaton, A., & Zaidi, S. (2002). Guidelines for constructing consumption aggregates for welfare analysis (LSMS Working Paper No. 135). Retrieved from https://openknowledge.worldbank.org/handle/10986/14101.

Dixon, H., Doores, J., Joshi, L., & Sinclair, F. (1999). Agroforestry knowledge toolkit for windows: Methodological guidelines, computer software and manual for WinAKT. Bangor.

Franzel, S., Carsan, S., Lukuyu, B., Sinja, J., & Wambugu, C. (2014). Fodder trees for improving livestock productivity and smallholder livelihoods in Africa. *Current Opinion in Environmental Sustainability, 6*, 98–103. https://doi.org/10.1016/j.cosust.2013.11.008.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... Michaelsen, J. (2015). The climate hazards infrared precipitation with stations: A new environmental record for monitoring extremes. *Scientific Data, 2*. https://doi.org/10.1038/sdata.2015.66.

Greene, W. H. (2003). Econometric analysis. (P. Education, Ed.), Journal of the American Statistical Association (Vol. 97). Prentice Hall. https://doi.org/10.1198/jasa.2002.s458.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics, 6*, 4779–4874. https://doi.org/10.1016/S1573-4412(07)06070-9.

Honey-Rosés, J., Baylis, K., & Ramírez, M. I. (2011). A spatially explicit estimate of avoided forest loss. *Conservation Biology, 25*(5), 1032–1043. https://doi.org/10.1111/j.1523-1739.2011.01729.x.

ICRAF. (2017). Corporate strategy 2017-2026. Nairobi. Retrieved from http://www.worldagroforestry.org/publication/corporate-strategy-2017-2026.

Imbens, G. W. (2010). Better LATE than nothing – Comments. *Journal of Economic Literature, 48*, 399–423. https://doi.org/10.1257/jel.48.2.399.

Jenkins, O. B. (2008). The Luhya of Kenya. Retrieved September 7, 2019, from http://orvillejenkins.com/profiles/luhya.html.

Joppa, L., & Pfaff, A. (2010). Reassessing the forest impacts of protection: The challenge of nonrandom location and a corrective method. *Annals of the New York Academy of Sciences, 1185*, 135–149. https://doi.org/10.1111/j.1749-6632.2009.05162.x.

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, National Council for Population and Development/Kenya, I. I. (2015). Kenya demographic and health survey 2014. Rockville, MD, USA. Retrieved from http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates.

Maillard, S., & Cheung, R. (2016). Unlocking the market for land degradation neutrality. Comines. Retrieved from https://www.innpact.com/uploads/news/files/MIROVASTUDYUnlocking-the-Market-for-LDNEN.pdf.

Nair, P. K. R. (1993). An introduction to agroforestry. Forest Ecology and Management (Vol. 73). Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1016/0378-1127(95)90008-X.

O'Donnell, O., van Doorslaer, E., Wagstaff, A., & Lindelow, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. Washington, DC: World Bank. Retrieved from https://openknowledge.worldbank.org/handle/10986/6896 License: CC BY 3.0 IGO.

Odhiambo, H. O., Ong, C. K., Deans, J. D., Wilson, J., Khan, A. A. H., & Sprent, J. I. (2001). Roots, soil water and crop yield: Tree crop interactions in a semi-arid agroforestry system in Kenya. *Plant and Soil, 235*(2), 221–233. https://doi.org/10.1023/A:1011959805622.

Place, F., Adato, M., Hebinck, P., & Omosa, M. (2005). The impact of agroforestry-based soil fertility replenishment practices on the poor in western Kenya. Research report of the international food policy research institute. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-33947281535&partnerID=40&md5=96562b0ae8a5fd22cb2280a04495a23a.

Place, F., Roothaert, R., Maina, L., Franzel, S., Sinja, J., & Wanjiku, J. (2009). The impact of fodder trees on milk production and income among smallholder dairy farmers in East Africa and the role of research (World Agroforestry Centre Occasional Paper Series). Nairobi.

Robalino, J., Pfaff, A., & Villalobos, L. (2015). Deforestation spillovers from Costa Rican protected areas (Universidad de Costa Rica Working Paper No. 15–2).

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1083), 41–55. https://doi.org/10.2307/2335942.

Sexton, J. O., Song, X. P., Feng, M., Noojipady, P., Anand, A., Huang, C., . . . Townshend, J. R. G. (2013). Landsat tree cover continuous fields. https://doi.org/10.1080/17538947.2013.786146.

Sjögren, H., Shepherd, K. D., & Karlsson, A. (2010). Effects of improved fallow with Sesbania sesban on maize productivity and Striga hermonthica infestation in Western Kenya. *Journal of Forestry Research, 21*(3), 379–386. https://doi.org/10.1007/s11676-010-0085-0.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. *PLoS ONE, 10*(2), 1–22. https://doi.org/10.1371/journal.pone.0107042.

Stevenson, J., Vanlauwe, B., Macours, K., Johnson, N., Krishnan, L., Place, F., ... Vlek, P. (2019). Farmer adoption of plot- and farm-level natural resource management practices: Between rhetoric and reality. *Global Food Security, 20*, 101–104. https://doi.org/10.1016/j.gfs.2019.01.003.

Vågen, T.-G., Winowiecki, L. A., Tondoh, J. E., Desta, L. T., & Gumbricht, T. (2016). Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma, 263*, 216–225. https://doi.org/10.1016/j.geoderma.2015.06.023.

White, H. (2009). Theory-based impact evaluation: Principles and practice. *Journal of Development Effectiveness, 1*(3), 271–284. https://doi.org/10.1080/19439340903114628.

White, H., & Bamberger, M. (2008). Introduction: Impact evaluation in official development agencies. In H. White (Ed.), *Impact evaluation: The experiences of official agencies (Volume 39)*. IDS Bullentin.

Wigboldus, S., Klerkx, L., Leeuwis, C., Schut, M., Muilerman, S., & Jochemsen, H. (2016). Systemic perspectives on scaling agricultural innovations. A review. *Agronomy for Sustainable Development, 36*(3), 46. https://doi.org/10.1007/s13593-016-0380-z.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources, 50*(2), 420–445. https://doi.org/10.3368/jhr.50.2.420.