Using panel survey and remote sensing data to explain yield gaps for maize in sub-Saharan Africa

Göran Djurfeldt, Ola Hall, Magnus Jirström, Maria Archila Bustos, Björn Holmquist and Sultana Nasrin

ABSTRACT
The aim of this paper is to combine remote sensing data with geo-coded household survey data in order to measure the impact of different socio-economic and biophysical factors on maize yields. We use multilevel linear regression to model village mean maize yield per year as a function of NDVI, commercialization, pluriactivity and distance to market. We draw on seven years of panel data on African smallholders, drawn from three rounds of data collection over a twelve-year period and 56 villages in six countries combined with a time-series analysis of NDVI data from the MODIS sensor. We show that, although there is much noise in yield forecasts as made with our methodology, socio-economic drivers substantially impact on yields, more, it seems, than do biophysical drivers. To reach more powerful explanations researchers need to incorporate socio-economic parameters in their models.

ARTICLE HISTORY
Received 17 March 2018
Accepted 1 August 2018

KEYWORDS
Smallholders; sub-Saharan Africa; yield gaps; panel data; transdisciplinary explanation

1. Introduction
Agriculture and its related industries are potential drivers of economic growth in sub-Saharan Africa (SSA) and are often seen as keys to poverty reduction and food security. Increased smallholder productivity is an important aspect of such agriculture-based growth and development because it is a source of income for the rural poor (World Bank, 2007). However, there is a productivity gap between achieved and potential yields in SSA. Studies show that there are major and growing discrepancies in productivity both when comparing African yields to those of the rest of the world and when comparing household level yields in SSA to locally achievable yields (Dzanku, Jirström, & Marstorp, 2015; Jayne, Mather, & Mghenyi, 2006; Van Dijk, Meijerink, Rau, & Shutes, 2012). Reducing these productivity gaps, known as yield gaps, is therefore a major goal in SSA.

An understanding of the major factors driving these observed yield gaps is important in helping to narrow productivity deficiencies and achieve the goals of increased agricultural productivity (Dzanku et al., 2015; Lobell, 2013; Tittonell & Giller, 2013). Likewise, a better understanding of the multitude of factors which contribute to the observed yield levels in SSA can add to a dialog about how these factors can be changed and improved in order to reduce yield gaps. However, ‘yield gaps are the end result of a myriad interacting biological, physical, and economic forces’ (Mueller & Binder, 2015, p. 45) which cannot easily be disentangled. For this reason, understanding yield gaps is not a simple task. Spatial and temporal variation in productivity are best regarded as the result of several interacting forces –
geographic, biophysical, socio-economic and political ones (Mueller & Binder, 2015). This is usually recognized in principle and theory but not always in practice and few studies examine the empirical relationship between these interacting forces.

In a review of 62 papers related to yield gaps, Snyder, Miththapala, Sommer, and Braslow (2016) found that most focused only on biophysical factors. In fact, only eight papers had models which included socio-economic factors. Similarly, in a meta-analysis of 50 peer-reviewed articles on yield gaps, Beza, Silva, Kooistra, and Reidsma (2017) found that only 16 of them considered the importance of farm characteristics such as labour and income in explaining yield gaps. An integrated approach to studying yields and yield gaps which incorporates not only biophysical but also farm management and socio-economic determinants is recognized as being important; however, in practice has not often been implemented empirically.

One reason for this lack of empirical research regarding the diverse factors that contribute to yield variations in SSA is a lack of good quality data. Official statistics often suffer from poor spatial and temporal coverage and low reliability which leads to deceptive conclusions and misleading, ahistorical policy recommendations (Calogero Carletto, Jolliffe, & Banerjee, 2013; Jerven, 2013, 2015). Rather than use official statistics, many researchers collect their own data. However, resource and time constraints often mean small samples, restricted locations and cross-sectional rather than longitudinal data. While more reliable on a local scale, such data still sacrifice spatial and temporal representativeness and have led to a situation where agrarian studies, from agronomic to anthropological ones, abound in number but lack in comparability. Good, reliable data across multiple time and spatial scales is hard to come by and this difficulty is magnified when attempting to collect detailed data across several disciplines. The consequence is that researchers are locked up in their different methodologies and corresponding world views.

One data source which has grown in importance within yield gaps studies is remote sensing (RS) data. RS is routinely drawn on in crop models, combined with rainfall, temperature, soil nutrients, management and other input data. Such models are known for detailed and reliable forecasts of harvests. However, these models are much more common in Europe and the United States, where modelling is facilitated by large field sizes and mono-cropping, than they are in SSA, where poor quality and low resolution of input data combined with small field sizes and pervasive intercropping limit their use. RS data can also be used to estimate vegetation indices (VIs) which can describe vegetation density and health. These VIs have been shown to be correlated with characteristics like leaf area index, biomass and productivity (Forkel et al., 2013; Myneni, Hall, Sellers, & Marshak, 1995). Since RS data at different resolutions is available consistently over time and space, it can be matched to socio-economic, political and other data, as long as such data is georeferenced. This combination of survey and RS data has the potential to help to reduce some of the previously discussed data issues associated with yield gap studies including the lack of multidisciplinary perspectives.

The aim of this paper is to combine RS data with geo-coded household survey data in order to measure the impact of different socio-economic and biophysical factors on maize yields. We use empirical models to study yields in 56 villages in six SSA countries during a 12 year period from 2002 to 2013. We choose to study maize yields because maize is the most common crop harvested in our sample villages.

2. Materials and methods

In order to account for both the biophysical and socio-economic drivers of yield, this study employs a mixture of two datasets along with a triangulation of methods. The datasets used include panel household surveys from the Afrint database and RS data from the MODIS satellite. Each of these will be presented in more detail in the following section. The mixed methodologies used include time-series analysis of VIs in order to gain an understanding of vegetation change and development coupled with bivariate and multivariate modelling of maize yields. These will be presented in Section 2.2.
2.1 Materials

2.1.1 Household level data from the Afrint database

The socio-economic and yield data used for this study are taken from the Afrint database, a database generated as part of the African Agricultural Intensification (Afrint) project. This project has been running since the year 2002 and was initially aimed at assessing the possibilities for an Asian style Green Revolution in nine countries in SSA based on household level data for approximately 4000 smallholder farms. Since the first round of data collection in 2002, two additional rounds have been carried out, one in 2008 and one in 2013/2015.1 Because of resource constraints, only six countries from the original nine have continued throughout the entire project. The database consists of two panel rounds (2002–2008 and 2008–2013) and three cross-sections (2002, 2008, 2013). A detailed description of the database and sampling methodologies can be found in Andersson Djurfeldt, Dzanku, and Cuthbert Isinika (2018).

The sample used in this study consists of 2544 households from six countries, grouped by village (N = 56) and region (N = 15). The sample constitutes between 30 and 60 farmers from each village. Table 1 shows the breakdown of the sample by country and region.

This study draws on only a small subset from the dearth of data collected in the Afrint database, mainly area under maize, amount of maize harvested, amount of maize sold, other sources of income, and geolocation data. For the first two of these variables we have data regarding the panel rounds (2002, 2008 and 2013) as well as the two preceding years since farmers were asked to recall previous figures. However, the 2000 and 2001 data are not comparable to that from the other years because the questions regarding maize production and area under maize were not asked in the same way. Likewise, for the amount of maize sold we have data from each panel round as well as the two preceding years; however, only for the first and second rounds (2002, 2008). For the other sources of income we have only three time points of data, one for each of the panel rounds. The geolocation data consist of a point location (latitude and longitude coordinates) for the center of each of the 56 villages.

The area under maize and amount of maize harvested were used in order to estimate maize yields, the dependent variable in this study. The remaining variables were used to generate measures of commercialization, pluriactivity and distance to market which were used as the socio-economic indicators in this study. These indicators were chosen based on previous research drawing on Afrint data which indicate that commercialization, pluriactivity and distance to market are significant proximate causes of yield variability (Andersson, Djurfeldt, Holmquist, Jirström, & Nasrin, 2011; Andersson, Djurfeldt, Holmquist, & Jirström, 2007; Djurfeldt, Aryeetey, & Isinika, 2011; Djurfeldt, Larsson, Holmquist, Jirström, & Andersson, 2008; Jirström, Andersson, & Djurfeldt, 2011; Jirström, Archila Bustos, & Alobo Loison, 2018).

| Country    | Region          | Number of villages |
|------------|-----------------|--------------------|
| Ghana      | Eastern         | 4                  |
|            | Upper Eastern   | 4                  |
| Kenya      | Kakamega        | 5                  |
|            | Nyeri           | 5                  |
| Malawi     | Bwanje Valley   | 2                  |
|            | Ntchisi RDP     | 2                  |
|            | Shire Highlands | 2                  |
|            | Thiwi Lifuodzi | 2                  |
| Mozambique | Center          | 5                  |
|            | North           | 4                  |
|            | South           | 2                  |
| Tanzania   | Iringa          | 5                  |
|            | Morogoro        | 5                  |
| Zambia     | Mazabuka        | 4                  |
|            | Mkushi          | 5                  |
2.1.2 Modis-NDVI

VI's are used to indicate the level of greenness of vegetation and biomass. They are conditioned by geo- and biophysical factors like rainfall, temperature and soil fertility and are therefore a good indicator of biophysical conditions (Kawabata, Ichii, & Yamaguchi, 2001). They also reflect management conditions such as fertilizer application, irrigation and labour inputs because these can impact soil conditions and alter biomass production. In this study we therefore use the Normalized Difference Vegetation Index (NDVI) in order to develop a proxy for biomass production over the growing season, which reflects both biophysical factors and farm management factors that impact yields.

MODIS vegetation index products (specifically NDVI), MOD13 (Terra) and MYD13 (Aqua) with a 250 m spatial resolution and 16-day temporal resolution, were used. Because the vegetation indices from the Terra and Aqua satellites are a phased product, meaning that the 16-day composites are generated eight days apart when alternating between the two satellites, we were able to get an effective 8-day temporal resolution for NDVI.

In order to cover each of our 56 villages, seven different MODIS tiles were required. MOD13 data are available from the year 2000 until the present time and, in order to match our survey data, we used data from 2001 to 2014 for a total of 14 years and 322 images per tile. MYD13 data are only available from July 2002 until the present. As such only 288 images were used. This means that the temporal resolution for the first year and a half of data is lower (16-days) as compared to the remaining study period (8-days). In total, we studied 610 images in a time-series analysis.

2.2 Methods

We model yields first using a bivariate model and then proceeding to a multivariate model. In this section we describe each of these models. First, however, we explain the methodology used to calculate our independent and dependent variables, summarized in Table 2.

The smallest unit of analysis in our study is the village level. This is because the NDVI indicator is based on a geographical location and we did not collect locational information on individual farms. Instead, we use the village location. As such, our socio-economic indicators and yields are aggregated to the village level as well.

The time period for our analysis is between 2002 and 2013 because this is the time period for which the socio-economic and NDVI indicators coincide. Additionally, some study years contain more data than other years because, as previously discussed, some information was not collected for all study years.

2.2.1 Socio-economic indicators

Commercialization is the first socio-economic indicator that we derive. We define it as the proportion of the total village harvest of maize that is sold in a given year. Pluriactivity, also referred to as non-farm diversification in other studies, is the second socio-economic driver we use. It reflects the pursuit of

| Indicator group | Variable name | Time period | Input data |
|-----------------|---------------|-------------|------------|
| Dependent       | Maize yield   | 2002, 2006, 2007, 2008, 2011, 2012, 2013 | Amount of maize harvested, area under maize |
| Independent     | NDVI small integral | Yearly from 2002 to 2013 | MODIS NDVI |
| Socioeconomic   | Commercialization | 2002, 2006, 2007, 2008, 2013 | Amount of maize harvested and sold |
|                 | Pluriactivity  | 2002, 2008, 2013 | Sources of income |
|                 | Distance to market | Unchanged (reference year 2000) | Village coordinates, travel time to major cities (Nelson, 2008) |
work and income outside of the farm, whether inside the village (work for other farmers, for example),
or outside of it, in neighboring villages, in the nearest town or further away. We define pluriactivity as
the proportion of total village income that is from non-farm sources. Finally, the third socio-economic
indicator we define is distance to market. This is measured as the Euclidean distance to the nearest
town of at least 50,000 inhabitants in the year 2000 (Nelson, 2008).

2.2.2 Biophysical indicators derived from NDVI
Time-series analysis of NDVI data allows for the extraction of seasonality parameters which, in this
study, were used as proxies for biomass production and by extension biophysical and farm
management indicators of maize yield. TIMESAT software was used to perform the analysis
(Jönsson & Eklundh, 2002). TIMESAT extracts seasonality information by smoothing curves from
noisy satellite data in two steps. It first fits local functions to sets of data points and then builds
global functions to describe the VI variations in vegetation seasons based on those local functions
(Jönsson & Eklundh, 2002).

For each village, two separate locations, composed of one MODIS pixel each, were examined,
one representing intensively managed land close to the known village coordinates (the cropped
pixel) and a second representing land further away from the village which appeared to be
unmanaged (the reference pixel). The selection of these two locations was done manually based
on visual inspection of Google Earth imagery. Both locations were screened and moved if they
showed uncharacteristic seasonal patterns. Thus, the reference pixels do not show the clear
seasonal variations that the cropped pixels do. In order to avoid mistakes in this manual process,
Afrint-team members familiar with the villages were consulted.

The TIMESAT software was then used to fit a seasonal curve to the time series data for each of
the two pixels in each of the 56 villages, resulting in a total of 112 time-series. A Savitsky-Golay
filter was used to smooth the data. Additionally, a seasonality parameter was set based on
observation of the data. The value of this parameter determines whether there are one or two
growing seasons. Finally, the value for the start of season was set as the point at which the left
edge had risen from the left minimum to 50% of the seasonal amplitude. The end of season was
likewise set as the point at which the right edge had declined to 50% of the seasonal amplitude
from the right minimum, see Figure 1.

Figure 1. TIMESAT parameters, amplitude, start and end of season, base level, small integral

Two results were reported from each time series for each growing season (one growing season
in some villages and two growing seasons in others). Because the filtering algorithm can introduce
inaccuracies in the first and last growing seasons of the time series, the output from the 2001 and
2014 growing seasons was ignored. The first output result taken was the length of the growing
season. The second was the NDVI small integral, which is determined as the integral of the fitted
function from the season start to the season end, minus the area below the base level, as illustrated
in Figure 1. This is a reflection of the biomass produced over the entire growing season.

2.2.3 Yield estimation by household survey panel data
Production and area under cultivation data were used to generate yield estimates at the village
level by calculating truncated village geometric means. Studies have shown that yield calculations
are especially sensitive to bias in farmers’ estimates for small fields (Carletto, Jolliffe, & Banerjee,
2015; Fraval et al., forthcoming). Because our data includes self-reported estimates for area under
maize cultivation for fields that are usually of a small size we first pruned the input data in order to
attempt to reduce some of the inaccuracies introduced by farmer bias. Initially, cases with reported
land sizes of less than 0.1 hectares were excluded. Additionally, cases with household level yields
(calculated as the production divided by the cultivated land size) of over 10 tons per hectare were
omitted. Furthermore, we deleted cases that reported a more than six-fold increase in yields
between consecutive years.
As a village level characteristic of yield (production/cultivated land area) we calculated the geometric mean of the within-village farmers yields. This gives a measure less sensitive to the skew distributions in production as well as in cultivated land area (and by implication also in yield) among the sampled farmers in the village. In order to ‘robustify’ the village estimate further, we removed the top and bottom 10% of observations per village in both the production and area variables.

2.2.4 Statistical analysis and models for yield estimation

We begin by presenting descriptive statistics over the length of season and NDVI small integral for both the cropped and reference pixel as well as the maize yields for each of the three time periods and performing an ANOVA of the differences between the time periods and the two types of pixels. The null hypothesis is that there are no differences between these variables over time or for different levels of management. A rejection of this hypothesis could signify a changing significance of the length of season or biomass over time or for different levels of management.

Next, we introduce two statistical models of the variation in yields. The first is a bivariate model and the second is a multivariate one. The aim of the bivariate model is to establish the relationship between biomass production (for the cropped pixel) and maize yields, without taking into account the contribution of the socio-economic indicators. Here we expect a positive relationship, in line with previous research (Lewis, Rowland, & Nadeau, 1998; Vrieling, de Beurs, & Brown, 2008; Wang, Rich, Price, & Kettle, 2005). Thus, we regress the log geometric mean of maize yield on the log NDVI small integral. This model, however, does not serve to address our research question, which aims to identify the impact of both biophysical and socio-economic factors on yield variation.

As such, we introduce our socio-economic indicators into a multivariate model. This model also examines the difference between the three survey periods, not with the primary aim of showing change between periods, but mainly to get three separate tests of the determinants of village yields. The null hypothesis is that these determinants do not change over time. If the

---

**Figure 1.** TIMESAT parameters, amplitude, start and end of season, base level, small integral.

Adapted from Jönsson and Eklundh (2002)
hypothesis is refuted it could, again, reflect changes in determinants of yields between periods. For this reason, we introduce two dummy variables, one for the second round of data (2006–2008) and one for the third round (2011–2013). The reference is the first round, which, as mentioned earlier, is effectively only the year 2002 because of limitations with our different data. We choose to simplify the time series into these three groups in order to simplify the model and because of limitations with data availability. For the yield and NDVI small integral, which were gathered over multiple years, the average for the corresponding panel period is taken. Here, as with the bivariate model, we use the NDVI small integral only for the cropped pixel. The socio-economic variables are taken from the 2008 period in order to assure that, at least in comparison to the more dynamic last period, they reflect impacts of the independent variables on the dependent one and not the other way around.

We employ a multilevel or hierarchical multivariate model with three levels, village, region, and country, mimicking the household survey sample design. The model is estimated by MLWin (Rasbash, Browne, Healy, Cameron, & Charlton, 2010), which is a software tailored to hierarchical models. Estimation is two-step, firstly by iterative generalized likelihood (IGLS) and, using the IGLS estimates as Bayesian priors, secondly by Markov-Chain Monte Carlo (MCMC) estimation.

Formally the model looks as below:

\[
\ln(y_{ijk}) = \beta_{0i} + \beta_1 V_{lijk} + \beta_2 \text{comm}_{ij} + \beta_3 \text{plur}_{ijk} + \beta_4 \text{dist}_{ijk} + \beta_5 \text{panel2}_{ijk} + \beta_6 \text{panel3}_{ijk}
\]

\[
\beta_{0ijk} = \beta_0 + \nu_{0k} + u_{0jk} + e_{0ijk}
\]

\[
\nu_{0k} = \sim N(0, \Omega_v) : \Omega_v = [\sigma^2_v]
\]

\[
u_{0k} = \sim N(0, \Omega_v) : \Omega_v = [\sigma^2_v]
\]

\[
e_{0ijk} = \sim N(0, \Omega_e) : \Omega_e = [\sigma^2_e]
\]

3. Results and discussion

Table 3 shows descriptive statistics for the length of season and NDVI small integral for both the cropped and reference pixel as well as the mean village maize yields by panel period.

An ANOVA of the differences between the three time period means shows that none of them are statistically significant. However, the difference between the total mean length of season observed in the cropped pixel and the one in the reference pixel is indeed statistically significant
The difference in the NDVI small integral between the two pixels is, however, not large enough to be significant. Thus, we do not see any significant differences over time in any of the variables studied. We do, however, see one difference when it comes to management. While the NDVI small integral is not changed, there is a shorter growing season for the managed land. An ad hoc explanation is that ploughing and sowing is somewhat delayed after the start of the rains, which is why the green season starts a little later in the cropped land.

The fact that the NDVI small integral is unchanged, despite a shorter growing season, can be interpreted as an indicator of the low productivity of the majority of the smallholders in our sample villages. Here we make an assumption that the cropped pixel should achieve higher biomass than the reference pixel. This is based on two further assumptions, drawn from our foundation that the NDVI small integral reflects both biophysical and farm management indicators of yield. First, we assume that good husbandry should lead to higher biomass production by improving the natural conditions of the land for growth of maize. Second, we assume that farmers knowingly select land with better biophysical preconditions for production. Therefore, we would expect to see that the cropped pixel would have a higher biomass production than the reference pixel; we, however, do not find such a trend. The farmers in our study have, therefore, not reached above the biological potential of the land which they are farming.

Descriptive statistics have shown no significant difference between the cropped and reference pixel and this we attribute to the fact that farmers are not improving the biological potential of their land. The cropped pixel managed by farmers who cultivate their land, while the reference pixel is not. For this reason, we move on to examine the cropped pixel specifically and explore how both biophysical and socio-economic factors influence the variance in yields. We start, as previously explained, by examining the biomass production only as a means of further examining the relationship between biophysical factors, farm management and yields.

Figure 2 presents a scattergram of the logged village mean yields and the logged NDVI small integral. There is a positive correlation, significant at the 10 per cent level; however, the coefficient of determination is very low ($R^2 = 0.013$), indicating that the log NDVI small integral accounts for a mere per cent of the variance in log yields. This is evident also by the wide scattering of cases around the line, as well as the breadth of the 90 per cent confidence interval. There are some

| Variable | 2002 | 2006–2008 | 2011–2013 | Total |
|----------|------|-----------|-----------|-------|
|          | No. of cases | Mean (SD) | No. of cases | Mean (SD) | No. of cases | Mean (SD) | No. of cases | Mean (SD) |
| 1. Length of season for cropped pixel, mean, days | 56 | 182.18 (6.02) | 138 | 174.48 (3.46) | 138 | 176.70 (3.83) | 332 | 176.70 (2.37) |
| 2. Length of season for reference pixel, mean, days | 56 | 184.75 (6.99) | 138 | 185.09 (3.90) | 138 | 187.56 (4.18) | 332 | 186.06 (2.64) |
| 3. Ratio of (2) to (1), mean | 56 | 1.07 (0.06) | 138 | 0.99 (0.03) | 138 | 1.02 (0.05) | 332 | 1.02 (0.03) |
| 4. Small integral ($VI$) for crop pixel, mean, 1000 units | 65 | 71.08 (2.86) | 168 | 71.78 (1.86) | 168 | 73.08 (1.84) | 401 | 72.21 (1.19) |
| 5. Small integral ($VI$) for reference pixel, mean, 1000 units | 65 | 66.67 (3.84) | 168 | 74.89 (2.46) | 168 | 73.58 (2.46) | 401 | 73.02 (1.59) |
| 6. Ratio of (5) to (4), mean | 65 | 0.985 (0.06) | 168 | 1.05 (0.03) | 168 | 1.002 (0.03) | 401 | 1.02 (0.02) |
| 7. Village maize yield, geometric mean, kg per hectare | 65 | 1104.94 (123.45) | 168 | 961.20 (34.84) | 168 | 1038.69 (38.37) | 401 | 1016.96 (29.53) |

Values in parentheses indicate standard error of mean.
outlying village-year means with low yields, all of them from 2008, which was a year with uncharacteristically low production in many villages due to flooding. Since they are located near the mean NDVI small integral (indicated by the vertical line), they contribute more to the variance than to the tilt of the regression line.

Figure 2. Regression of village ratio mean yield on the NDVI small integral with 90 per cent confidence interval.

While this bivariate exercise confirms the expected relationship between biomass and maize yields, in our SSA context it proves to be a weak relationship. For this reason, we expand the bivariate model in Figure 2 to a multivariate one, which allows us to look at the interaction between biophysical and socio-economic factors in driving yield variation. The results of this multivariate model are seen in Table 4.

The estimated model is statistically significant (Chi2 = 120.35, df = 10, p < 0.001). The variance between villages is highly significant while the within-village variance does not deviate far from normality, with no apparent bias in the estimation. The variances between countries and between regions are not statistically significant.

Let us comment on the results variable by variable, starting with the NDVI small integral. In the bivariate model presented in Figure 2, the log of the NDVI small integral was statistically significant, but only at 10 per cent level. This result holds in the multivariate setup and is, in fact, somewhat stronger (p < 0.05). However, the NDVI small integral still explains only a small share of the overall variance in village yields. A one per cent relative increase in the NDVI small integral is expected to result in a 0.18 per cent relative increase in the village yield. Thus, most of the variance in yields is explained, not by the biophysical and management factors, but by other factors, some of which are the socio-economic indicators which we now turn our focus to.

It is first, however, important to recognize one of the main limitations of this study. We have found that NDVI small integral is not a very strong indicator of yields under the observed conditions in SSA and among smallholders. It should be stressed, however, that this is partly a scale effect, since we are studying MODIS pixels of the size of 250 × 250 meters with no indicator of within-pixel variance. We do
this, however, in order to be able to get a detailed time series. Other RS datasets (such as Landsat) have higher spatial resolutions, but this comes at the expense of temporal resolution. Because our aim was to integrate survey data, which we have over a 12 year period, with RS data using the NDVI small integral, which requires detailed time series data, we feel that this trade-off is justified.

Looking now at our socio-economic indicators, both commercialization and pluriactivity are statistically significant (p < 0.05 and p < 0.001 respectively); however, the distance indicator is not significant. A unit increase in commercialization is expected to result in an almost 60 per cent increase in yields, controlling for other factors. However, since commercialization is defined as a ratio, this implies an unrealistic change between zero and unity. To clarify, increasing commercialization from, for example, 0.5 to 0.6 (one tenth of a unit) is expected to result in a relative change in village yield of 5.9 per cent. This result reflects the differences in land productivity between highly commercialized villages and those more oriented to subsistence production. Unlike the relationship with commercialization, the relationship between pluriactivity and yields is a negative one. A unit increase in the pluriactivity rate of a village is expected to be associated with approximately 30 per cent lower yields than otherwise. Similar to commercialization, pluriactivity is a ratio variable so, again, looking more specifically at increasing pluriactivity from 0.5 to 0.6 (one tenth of a unit), for instance, would give a relative negative change in village yield of 6.6 per cent. Here we see that villages with a high share of households drawing on non-farm sources of income may be more oriented to subsistence production than villages focused on maize production where farmers may dedicate more time and resources to trying to increase maize yields since this is their main source of income.

Like commercialization and pluriactivity, travel time to the nearest major market was expected to be significantly associated with yield levels; however, in the current model we find that travel time is not statistically significant, even when the other socio-economic variables are removed from the model. One possible explanation for the discrepancy between the observed findings and those expected based on previous studies may be that the lag time is too long, since the travel time is estimated under conditions prevailing in the year 2000. Much can have changed since then, both in terms of urbanization, in the size of towns neighbouring our villages, and in the infrastructure connecting the villages to major markets. Another possibility is that the distance indicator is collinear with the rates of commercialization and pluriactivity. This, however, is less likely, since running separate models for each panel period shows substantially the same results.

In terms of change over time, the period dummies in the multivariate model are both statistically significant. The regression coefficients should be interpreted in comparison to the reference

Table 4. Multilevel linear regression of village mean log yield, truncated sample, on the log NDVI small integral, panel period and socio-economic indicators.

| Fixed effects:                  | β     | SE   | Sig. |
|---------------------------------|-------|------|------|
| Intercept                       | 4.90  | 0.91 | ***  |
| NDVI small integral, logged     | 0.18  | 0.08 | *    |
| Panel 2, 2006–2008, dummy       | 0.12  | 0.07 |      |
| Panel 3, 2011–2013, dummy       | 0.20  | 0.07 | **   |
| Commercialization: Share of village maize harvest sold, 2008 | 0.58  | 0.25 | *    |
| Pluriactivity: Share of village households that had non-farm income, 2008 | −0.68 | 0.21 | ***  |
| Travel time to nearest town with 50,000+ inhabitants, 2000 | −0.01 | 0.02 |      |

Random effects:

| Country                        | 0.15  | 0.26 |      |
| Region                         | 0.15  | 0.10 |      |
| Village                        | 0.21  | 0.02 | ***  |
| Number of cases                | 388   |      |      |
| Missing, per cent              | 3     |      |      |
| −2*loglikelihood, null model   | 673.33|      |      |
| Ditto, full model              | 552.98|      |      |
| Chi², df, sig.                 | 120.35| 10   | ***  |

Note: All regression coefficients have been tested to establish if they are significantly different from zero. Significance levels are denoted as *, **, *** for 5, 1 and 0.1 per cent significance respectively.
category, which is the year 2002. Compared to the reference period and controlling for the other factors, maize yields in 2006–2008 are expected to be approximately 12 per cent higher (p < 0.1), while those in 2011–2013 are expected to be approximately 20 per cent higher (p < 0.01) than in 2002. Thus, the multivariate analysis highlights a trend in yields, which was not apparent in the bivariate analysis or analysis of the descriptive statistics.\footnote{\textsuperscript{11}}

Some of the above results tally well with those shown in several previous publications using the same Afrint data, mainly that socio-economic factors such as commercialization and pluriactivity have a strong impact on yields (Andersson, et al., 2011; Andersson et al., 2007; Djurfeldt et al., 2011, 2008; Jirström et al., 2011). In other words, if our sample is representative of wider trends, market integration drives the development of smallholder agriculture in SSA. Biophysical and farm management factors, as reflected in the NDVI small integral, apparently neither prevent nor encourage this development to a significant level.

4. Conclusion

The purpose of this paper was to measure the impact of different socio-economic and biophysical factors on maize yields using a combination of RS data and household survey data. We evaluated the relationship between the NDVI small integral during the growing season, commercialization, pluriactivity and distance to market and variation in maize yields during a 12 year period from 2002 to 2013. We find that, on average, the NDVI in areas managed by farmers (cropped pixel) is not significantly different from that in less managed areas (reference pixel). We also find that, while a significant positive relationship between the NDVI small integral and maize yields exists, it is weak, when examined individually. When examined along with socio-economic indicators, the weak significant relationship persists, but now we find that there is a stronger relationship between commercialization and pluriactivity and maize yields. Contrary to expectations, we find no significant relationship with distance to market.

We conclude that, at this scale of analysis, changes in biomass production, due to both biophysical conditions such as rainfall and temperature, and soil conditions driven by both these conditions and management factors, do not drive changes in yields to a significant level. Instead, trends in maize yields are more predominantly driven by socio-economic factors, like commercialization or market integration and rates of pluriactivity. Thus, yield estimation models focusing on biophysical drivers of yields while neglecting other drivers are likely to overestimate the effects of the former on yields. Models of yield estimation would thus benefit greatly from accounting for not only biophysical drivers of yield variation but also the socio-economic context within which maize is grown and harvested.

Notes

1. In the third round, data for four of the six countries was collected in 2013. The data for Mozambique and Tanzania was collected in 2015 instead of 2013. The third round of data collection will be referred to as 2013 in this study.

2. A second, more elaborate multivariate model containing fixed country effects for three of the independent variables (panel period, pluriactivity and commercialization) was also developed. Details have been excluded here; however, it has been included in Table A1. In instances where the findings from the first multivariate are modified by the second, this has been identified in a footnote.

3. As pointed out, in effect only year 2002.

4. The second multivariate model in Table A1 brings the variance between countries to light. Here the variance between several countries is statistically significant for a number of variables.

5. The second multivariate model in Table A1 shows a significant negative association between distance to market and yield. One way to interpret this is that differences between countries with small distances and good infrastructure and others are masked in the model above.

6. $\exp(0.058) - 1 = 5.9\%$. 
7. Here Mozambique and also Zambia pull down the positive value of the regression coefficient, perhaps because the former country is more subsistence-oriented. In the case of Zambia, the heavy dependence of maize production on government purchasing policy is possibly an explanation for the deviant pattern.

8. \[ \exp(0.068) - 1 = 6.6\% . \]

9. The generally negative effect of pluriactivity on yields \((\beta = -0.68)\) is masked somewhat by Malawi and to a smaller extent by Zambia. Referring to Djurfeldt, Djurfeldt, Hall, and Archila Bustos (2018) a possible explanation is that the former country is less affected by the agrarian change driven by the structural transformation in the other countries since 2002. As noted in footnote 7, Zambia may be exceptional due to its State-dominated purchasing policy for maize.

10. The second multivariate model in Table A1 does, however, as previously mentioned, show a significant negative association between travel time and yields. This is more in line with the expected patterns.

11. Malawi is a deviant country and pulls down the general effects and the regression coefficients for panel 2 and 3. In other words, stagnation in Malawi somewhat masks the overall progress in yields in the five other countries.

Acknowledgments

The authors acknowledge the support and funding from The Swedish Research Council. We are grateful for comments from Agnes Andersson Djurfeldt and participants in the Development Research Day, Lund University 2017 and Rural Development Seminar, Department of Economic History Lund University. The assistance of Afrint team members is much appreciated: Daniel Bruce Sarpong, Fred Dzanku, Willis Oluoch-Kosura, Stephen K. Wambugu, Joseph Karugia, Wapulumuka O. Mulwafu, Peter Mvula, Peter Coughlin, Aida Isinika, Elbariki Msuya, Mukata Wamulume and Charlotte Wonani.

Declaration of interests

We hereby certify that all authors have seen and approved the final version of the manuscript being submitted. We warrant that the article is the authors’ original work, hasn’t received prior publication and isn’t under consideration for publication elsewhere. There is no financial/personal interest or belief that could affect our objectivity and no potential such conflicts exist.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Vetenskapsrådet.

ORCID

Göran Djurfeldt http://orcid.org/0000-0002-1139-2404
Ola Hall http://orcid.org/0000-0002-9231-4028
Magnus Jirström http://orcid.org/0000-0002-7053-4492
Björn Holmquist http://orcid.org/0000-0003-4316-2483
Sultana Nasrin http://orcid.org/0000-0001-7390-0542

References

Andersson, A., Djurfeldt, G., Holmquist, B., & Jirström, M. (2007). Green revolution stalled? The drivers of farm change in sub-Saharan Africa. The case of maize. Lund: Department of Sociology.

Andersson, A., Djurfeldt, G., Holmquist, B., Jirström, M., & Nasrin, S. (2011). A new era for sub-Saharan African agriculture? Changing drivers of maize production. In G. Djurfeldt, E. Aryeetey, & A. Isinika (Eds.), African smallholders - food crops, markets and technology (pp. 107–137). Wallingford, Oxon.: CAB International.

Andersson Djurfeldt, A., Dzanku, F.M., & Cuthbert Isinika, A. (Eds.). (2018). Agriculture, diversification, and gender in rural Africa: Longitudinal perspectives from six countries. Oxford: Oxford University Press.

Beza, E., Silva, J.V., Kooistra, L., & Reidsma, P. (2017). Review of yield gap explaining factors and opportunities for alternative data collection approaches. European Journal of Agronomy, 82, 206–222.
## Appendix

Table A1. Linear regression of village mean yields, logged, truncated sample, on the NDVI small integral, logged, panel period and socio-economic indicators, mixed model with country effects.

|                      | $\beta$ | SE  | Sig. |
|----------------------|---------|-----|-----|
| Intercept            | 5.08    | 0.80| *** |
| NDVI small integral, logged | 0.16    | 0.07| *   |
| Panel 2 for:         |         |     |     |
| Ghana                | 0.15    | 0.19|     |
| Kenya                | 0.30    | 0.12| *   |
| Malawi               | −0.99   | 0.17| *** |
| Mozambique           | 0.21    | 0.14|     |
| Tanzania             | 0.25    | 0.15|     |
| Zambia               | 0.59    | 0.17| *** |
| Panel 3 for:         |         |     |     |
| Ghana                | 0.33    | 0.19|     |
| Kenya                | 0.23    | 0.12|     |
| Malawi               | −0.99   | 0.17| *** |
| Mozambique           | 0.36    | 0.14| *   |
| Tanzania             | 0.39    | 0.15| *   |
| Zambia               | 0.69    | 0.17| *** |
| Pluriactivity for:   |         |     |     |
| Ghana                | −2.08   | 0.32| *** |
| Kenya                | −0.82   | 0.27| **  |
| Malawi               | 0.92    | 0.28| **  |
| Mozambique           | −0.67   | 0.20| *** |
| Tanzania             | −0.80   | 0.30| **  |
| Zambia               | 0.17    | 0.32|     |
| Commercialization for: |       |     |     |
| Ghana                | 1.05    | 0.20| *** |
| Kenya                | 1.29    | 0.35| *** |
| Malawi               | 4.19    | 1.30| **  |
| Mozambique           | −0.90   | 0.35| *   |
| Tanzania             | 1.21    | 0.54| *   |
| Zambia               | −0.44   | 0.41|     |
| Travel time to market, hrs. | −0.05  | 0.01| *** |

Note: All regression coefficients have been tested to establish if they are significantly different from zero. Significance levels are denoted as *, **, *** for 5, 1 and 0.1 per cent significance respectively.

$-2^{\log}\text{likelihood} = 461.094$