Maynooth
University
National University
of Ireland Maynooth

Department of Economics

Working Paper N320-24

Testing the GAEZ agronomic model in the fields: Evidence from Uganda

Bruno Morando[*]

February 2024

Abstract

The Global Agroecological Zones (GAEZ) model developed by FAO and IIASA provides granular crop-specific expected yields worldwide. An increasing number of papers in economics are using this dataset for a number of different purposes, such as deriving an exogenous source of variations in institutions or to parametrize computable general equilibrium models of trade in agricultural goods. However, as pointed out by the GAEZ creators, its results ”should be treated in a conservative manner and at appropriate aggregation levels, which are commensurate with the resolution of the basic data”. In this paper, I ”test” the GAEZ predicted yields by comparing them with the ones measured through agricultural surveys ”in the fields”. In particular, I use data from two different agricultural surveys in Uganda, where farmers typically grow a multitude of crops. In both instances, I find that GAEZ predictions often correlate negatively with yields, and that farmers' crop choices are more responsive to survey-based yields than to GAEZ's predictions. These findings suggest that, at least in the context of Uganda, predicted GAEZ yields fail to make reliable predictions at the granular level.

Key Words: GAEZ, Agriculture, Crop Choice, Uganda
JEL Codes: O55, Q10, R14

[*] Bruno Morando (bruno.morando@mu.ie) is at the Department of Economics in Maynooth University, Ireland.

1 Introduction

The use of spatial data in economics is becoming more and more common, and an increasing number of studies rely on them to develop innovative identification strategies and generate otherwise unobservable variables for their analysis (Donaldson and Storeygard 2016). The Global Agro-Ecological Zones (GAEZ) dataset, developed by FAO and IIASA, is perhaps the most prominent of such datasets in the agricultural, trade and development literature. Leveraging granular information on climate and soil characteristics, which are fed into a sophisticated agronomic model, the GAEZ provides expected crop-specific yields under different input intensities worldwide and at a very granular level.

A potential issue with spatial datasets is that while information is presented at a very disaggregated level and is typically easily accessible, the methodologies used to create the grids can lead to substantial measurement errors when the data are used for applications in economics. For example, Gibson et al. (2020) and Gibson et al. (2021) report some substantial issues in the of night-lights data typically used by economists to proxy for economic activities and urbanization. Similarly, Sun et al. (2018) find major inconsistencies when comparing different sources of rainfall data, which are increasingly used in the economic literature (Dell et al. 2014).

As pointed out in the GAEZ model documentation (Fischer et al. 2021), due to the low resolution/precision of some of the raw data used in the analysis as well as some generalizations required to harmonize the agronomic model for global application, some of these issues might occur when using the predicted yields (as well as other results) from the GAEZ model. Quoting the authors: (FAO & IIASA 2022) "global data sets used as inputs [. . .] are known to be of uneven quality and reliability. Hence, the results [. . .] should be treated in a conservative manner at appropriate aggregation levels." (FAQ 13).

In this paper, I "test" the predicted yields from the GAEZ grid by comparing them with yields derived from geolocated farm level data from two agricultural surveys in Uganda.[1] It is important to specify that since the analysis involves the comparison of yields across different cells in a relatively small country (Uganda is ≈ 240 km^2 and covers 2,442 GAEZ cells), the scope of this study is not to verify the "validity" of the GAEZ model, but rather evaluate to which extent variations in measured predicted yields within narrow geographical ares correlate with observed differences in crops yields.

There are two main challenges in doing so: 1) While the GAEZ model returns crop specific yields *for all the cultivated land* in each cell (with a dimension of ≈ 10 km^2 at the Equator), survey-based yields are only observed in plots *where each crop is grown.* This implies that

[1] As explained in the following, the country is a useful benchmark due to the fact that farm households typically grow a large number of different crops, providing me with an adequate sample size to perform the comparison for more crop varieties. Additionally, I have access to two different datasets (presenting different advantages and disadvantages) which can be used to cross check the results of the empirical analysis. Finally, the Ugandan Workd Bank's LSMS is considered one of the most reliable in terms of the agricultural survey as it records plot and crop specific input use (Gollin and Udry 2021).

where land is heterogeneous within cells and crops' allocation is non-random, actual and cell-level predicted yields are not directly comparable. 2) Measuring yields from agricultural surveys is notoriously problematic (Abay et al. 2019; Abay et al. 2021; Ayalew et al. 2023) as survey-based measures of land inputs and crop outputs are subject to numerous sources of measurement errors.

To address the first issue, I develop a methodology to correct GAEZ yields to allow for farmers' crop choice to follow cell-specific comparative advantage i.e. allocate crops to relatively more suitable fields in the same area. Since information on within cell suitability distribution is not available, I parametrize the joint density function of crop suitability using cross-cell empirical moments.

As for the yields measurement error, I exploit the large number of observations to generate a more credible measure of cell-level yields: first, I drop the 5% more extreme plot- (or household-) level yields in the distribution. Second, I only include in the analysis cells where at least 10 separate observations were available in order to prevent outliers to have a big effect on the estimates. Finally, I employ two different surveys (the World Bank's Ugandan LSMS and the UBOS' Ugandan Census of Agriculture -UCA-) each presenting some advantages and disadvantages with respect to the other in order to cross-check the results.

The empirical analysis indicates that, among the nine major crops considered, only in the case of groundnuts (and only in one of the two samples) is there a positive and statistically significant relationship between the GAEZ yields and the ones derived from agricultural surveys "in the fields". Surprisingly, I find a negative and statistically significant relationship between the GAEZ and actual yields of 5 crops both in the LSMS and in the UCA, indicating that the cross-cell yields as predicted by the GAEZ model and the ones observed in the data is negative. This does not change when controlling for land characteristics and input intensity or when the GAEZ yields are corrected to account for non-random allocation of crops within cells.

In order to corroborate these findings, I study the relationship between crop choice and measured yields across cells. The main hypothesis is that the ratio of land devoted to different crops in different cells should (at least partly) depend on their relative suitability. When using the GAEZ predicted yields as a proxy for suitability, the results indicate a null or often negative relationship between relative crop specific productivity. On the other hand, when using survey yields (from either the LSMS or the UCA) the relationship is positive for the majority of the crop pairs considered. Though this evidence is not conclusive, it corroborates the credibility of survey-generated yields.

While the GAEZ remains an incredibly valuable asset to economists, these findings suggest some caution where considering granular variation across cells, at least in the context of Sub-Saharan Africa.

The paper develops as follows: Section 2 provides a description of the GAEZ dataset and its use in the economics literature. Section 3 illustrates the methodology used to "correct" the GAEZ predicted yields to account for non-random within-cell crop allocation where land is heterogeneous. Section 4 describes the two datasets used to obtain data on cell-specific yields from the fields. The results of the comparison between actual and GAEZ data and the empirical analysis of crop choice and crop-specific yields are shown in section 5 and 6 respectively. Section 7 concludes.

2 The GAEZ dataset and its use in the economic literature

This section provides a brief description of the GAEZ dataset and how it has been used in the existing economics literature. A more thorough explanation of the GAEZ model's technicalities can be found in the model documentation of the most recent (4.0) version (Fischer et al. 2021), which is the one I consider in this paper.

The main purpose of the GAEZ is to provide granular level information of the agricultural suitability for different crops across fields all over the world. This is achieved by combining information on climate and soil characteristics and feeding them into a sophisticated agronomic model returning the expected yields for 53 different crops. In short, given the typical weather and terrain characteristics observed in a given area, the agronomic model predicts the amount of achievable production for every given crops.

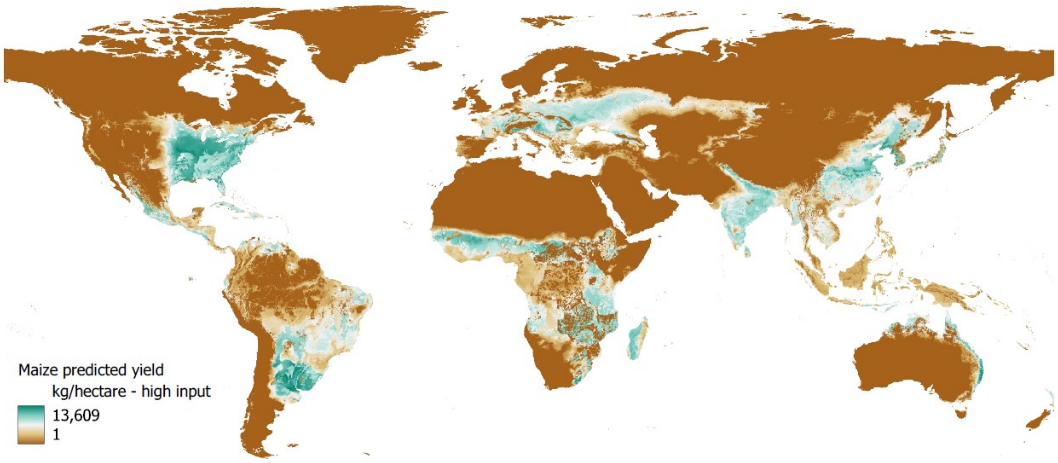
The actual process involve the combination of different raw datasets describing growing conditions which are progressively added up to estimate the severity of constraints to the production of different crops and eventually generate a grid displaying the distribution of crop-specific potential yields worldwide. For example, figure 1 shows the distribution of potential attainable maize yield where a high input intensity is applied.

Additionally, the model can generate the expected yields under different input intensities, irrigation facilities and climate projections (i.e. using projected temperature and precipitation variables to account for climate change rather than historical recordings). In the empirical analysis, I will focus on the specification with low input intensity, no irrigation and considering climate variables obtained from 1981-2010 historical data, as it is the one most closely aligned with the modal production techniques in the Ugandan agricultural sector. Finally, although most economic studies using the GAEZ are primarily interested in these predicted yields, it is worth mentioning that the dataset also provides estimates on actual cell-specific production and gaps between actual and potential yields. Figure 2, taken from the model documentation (Fischer et al. 2021) illustrates the structure of the GAEZ and the available outputs.

Technically, the outputs from the GAEZ model are presented in the form of fine grids (rasters), where each cell takes on a different value indicating the quantity of interest (e.g. maize suitability in the high input scenario). The size of the cells is of 5 arc-minutes, corresponding to roughly 10km^2 at the Equator. As shown in figure 3, this implies that the GAEZ can potentially identify variation in crop-specific yields at a very fine scale, as even a relatively small country like Uganda contains over 2,400 different cells.

A growing number of works in economics are using information from the GAEZ model. The

Figure 1: Maize suitability Worldwide (high input)



Source: GAEZ v4.0.

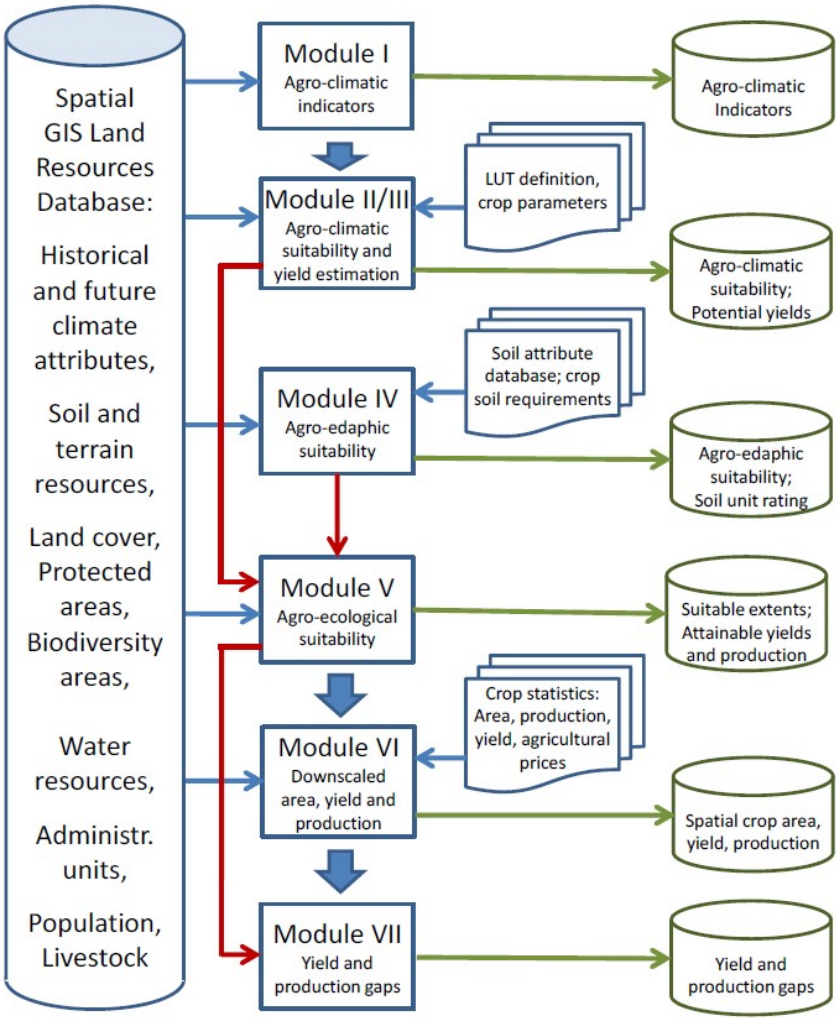
most desirable features of the dataset is that crop-specific information is available regardless on whether the crop is actually grown in the area or not and the fact that the estimates provided depend exclusively on exogenous soil and weather characteristics. Thus, it is relatively easy to leverage this information to generate credible counterfactual (in terms of alternative crop distributions and resulting yields/production) and/or as an exogenous source of variation for studies based on instrumental variables identification.

For example, Adamopoulos and Restuccia (2022) use GAEZ data on potential and actual yields to study the nature of the agricultural productivity gap between rich and poor countries, finding that it is not due to inherently different land quality but rather in inefficient production modes and geographical distribution of crops in the latter. Estimates of potential yields are used by Costinot and Donaldson (2016) and Sotelo (2020) to calibrate domestic trade models and in turn the gains from trade deriving from spatial integration across regional markets. Costinot et al. (2016) compares the existing system of agricultural comparative advantage with the one under expected climate change (i.e. derived from weather projections between 2030 and 2050) to study the production and welfare impact of global warming on different economies.

In other instances, estimates from the GAEZ have been used to ”test” theories on the long-run determinants of political institutions (Fernández-Villaverde et al. 2023 and preferences Galor and Özak 2016, sometime exploiting the so-called Columbian exchange (Mayshar et al. 2022; Nunn and Qian 2011). These studies are typically quite large in scale and do not heavily rely on granular differences in predicted yields.

Other studies instead base their identification strategies on more granular variation in the GAEZ model. Dias et al. (2023) use variation in the relative soybean yields between the low- and high-input scenarios in Brazil to derive exogenous variation in the profitability of glyphosate use across municipalities to study its impact on birth outcomes. In a similar way,

Figure 2: Overall structure of GAEZ data

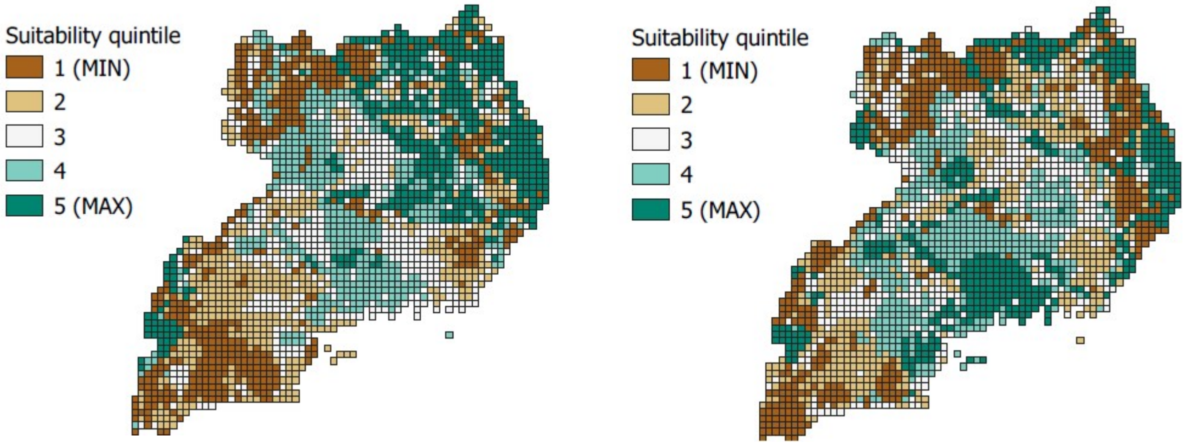

Spatial GIS Land Resources Database:
Historical and future climate attributes,
Soil and terrain resources,
Land cover, Protected areas, Biodiversity areas,
Water resources,
Administr. units,
Population, Livestock
Module I
Agro-climatic indicators
Module II/III
Agro-climatic suitability and yield estimation
LUT definition, crop parameters
Module IV
Agro-edaphic suitability
Soil attribute database; crop soil requirements
Module V
Agro-ecological suitability
Module VI
Downscaled area, yield and production
Crop statistics: Area, production, yield, agricultural prices
Module VII
Yield and production gaps
Agro-climatic Indicators
Agro-climatic suitability; Potential yields
Agro-edaphic suitability; Soil unit rating
Suitable extents; Attainable yields and production
Spatial crop area, yield, production
Yield and production gaps

Source: Fischer et al. (2021).

Lowes and Montero (2021) exploit variation in the relative suitability of cassava and millet across different sampling areas in former central African French colonies to predict exogenous variation in mandatory medical treatment aimed to tackle sleeping sickness (since cassava is less land intensive and requires to be soaked in water, tse-tse flies found a more suitable environment in zones were the crop was grown more widely) to examine the impact of these colonial practices on current trust in modern medicine. Berman et al. (2021) uses variation in the Harmonized World Soil Database (one of the main inputs of the GAEZ model) and main crops produced from the GAEZ model interacted with changes in fertilizer prices to estimate the quantitative impact of inequality on the offset of conflict in Africa.

In short, the GAEZ dataset enabled a wide range of research questions and its use produce a large number of interesting results, especially in the development economics and trade

Figure 3: Maize (left) and cassava (right) suitability in Uganda by quintiles



Source: Author's computation based on GAEZ v4.0.

literature. Due to the quality of the data as well as the state-of-the-art agronomic model that combines them, it represents a very valuable resource which can be easily accessed by scholars and used to answer more relevant questions. However, as explained in the model documentation (Fischer et al. 2021), there is some variation in the quality of the raw data used which can result in imprecise outputs, especially when using results at a very disaggregated level. This paper represents the first attempt to compare actual and predicted yields across different cells in a single country (Uganda) and study the empirical relationship between the production observed ”in the fields” and the one predicted by the model.

3 Within-cell heterogeneity and non-random crop distribution

When comparing the GAEZ predictions with actual yields, it is important to account for the fact that, while the GAEZ dataset provides expected yields for the whole cultivated area of a cell, data on actual yields are only available for the area where each crop is actually grown. This has two broad implications: first, it is only possible to test the GAEZ model in cells where a crop is actually grown, second, it is possible that the fields where a crop is grown in each cell differ systematically and in a non-random fashion from the average plot of land. For example, if farmers only grow maize in relatively more suitable fields in a given cell, this will result in a systematic difference between the GAEZ predictions and the actual yields which cannot be attributed to incorrect predictions from the GAEZ (which by definition only captures average yields in the cultivated area).

In order to account for this, I develop a methodology to correct the GAEZ predictions and account for non-random allocation of land to different crops *within cells* where there is some heterogeneity among fields in the same cell. The following section illustrates this methodology and the assumptions made in modelling the unobserved structure of crop-specific comparative advantage within cells and the subsequent farmers' choices.

3.1 GAEZ correction

Formally, the GAEZ dataset provides, for each crop $k \in K$ and cell $c \in C$, the expected yield in the cultivated area:

$$\text{GAEZ}_k^c = \int^{L^c} \hat{y}_k^c \; g(\hat{y}_k^c) \; d\hat{y}_k^c \tag{1}$$

where L^c denotes the total cultivated area in cell c, which is a continuum of fields or plots whose predicted yields for crop k are denoted by $\hat{y}_k^c$. $g(\hat{y}_k^c)$ represents crop k yield density function across the whole cultivated area in cell c and it is such that:

$$\int^{L^c} g(\hat{y}_k^c) \; d\hat{y}_k^c = 1 \tag{2}$$

On the other hand, the actual yields observed in each cell do not refer to the entire cultivated area, but only to the area in cell c devoted to the specific crop k (i.e. the aggregate yield is the average across the fields where the crop is cultivated, not of all cultivated fields). For example, the actual yield for crop k in cell c observed in the LSMS data can be described as:[2]

$$\text{LSMS}_k^c = \int^{L_k^c} y_k^c \; f(y_k^c) \; dy_k^c \tag{3}$$

where Y_k^c denotes the actual yields of the fields and L_k^c is the total land in cell c where crop k is grown, where by definition $L_k^c < L^c$. $f(y_k^c)$ is the density function of yields in the areas cultivated with the crop k, and is such that:

$$\int^{L_k^c} f(y_k^c) \; dy_k^c = 1 \tag{4}$$

It is clear that, even where the GAEZ estimated yields are correct ($\hat{y}_k^c = y_k^c \; \forall \; k, c$), the quantities expressed in equation 1 and 3 (i.e. the yields in all the cultivated area of the cell and the yields in the areas cultivated with the specific crops) differ as long as there is some heterogeneity in the yields within cell and the land allocation to crops is non-random.

Ideally, the right quantity to compare to actual yields observed in survey data is not the GAEZ output in equation 1, but the average predicted yields in the fields actually devoted to the crop:

[2] While I use LSMS in the notation, the same holds true for the other dataset considered: the UCA, and the methodology could be generalized to any survey-based measure of yields.

$$\int^{L_k^c} \hat{y_k^c} \; f(\hat{y_k^c}) \; d\hat{y_k^c} \tag{5}$$

Although this quantity is not observed, it is possible to generate an approximation by imposing some structure on the joint distribution of crop specific yields within a cell and a rule determining crop choices. By doing so, the raw output of the GAEZ can be corrected to account for non-random within-cell crop distribution where crop suitability in the same cells are heterogeneous.

In terms of the crop choices within cells, it is reasonable to assume that farmers aim to maximize the production and as such their decisions depend on the structure of comparative advantage across plots. Formally, the problem can be presented as:

$$\max_{o_k^c} \sum^{K} \frac{1}{GAEZ_k^c} \int^{L^c} o_k^c \; \hat{y}_k^c \; g(\hat{y}_k^c) \; d\hat{y}_k^c \tag{6}$$

where o_k^c is a selection variable taking value 1 where a plot is devoted to crop k and 0 otherwise. In short, the objective of the farmers is to allocate crops in a way that maximizes their total production, where different crops are aggregated by considering the deviation of the actual yields from the average achievable in the cell $GAEZ_k^c$.[3]

Two constraints are imposed on the farmers' maximization problem. First, that the share of land devoted to each crop in the cell is the one actually observed in the data, second, that only one crop is grown in each field. Formally:

$$\begin{aligned} &\int^{L^c} o_k^c \; g(\hat{y_k^c}) \; dy_k^c = l_k^c \;\; \forall k \\ &\sum^{K} o_k^c = 1 \end{aligned} \tag{7}$$

The solution to this maximization problem allows me to compute, for each crop k and cell c, the expected yield *in the area where the crop is grown* as opposed to the one for the whole cell, which can be compared to the actual yields obtained in the data. These "corrected" yields are denoted as $\widetilde{GAEZ}_k^c$, and formally are presented as:

$$\widetilde{GAEZ}_k^c = \int^{o_k^c=1} \hat{y}_k^c \; \phi(\hat{y}_k^c) \; d\hat{y}_k^c \tag{8}$$

[3] Note that other aggregation methodologies could be considered, such as the prices of different crops. However, in the context under analysis most crops are grown for self-consumption and as such market values might not mirror the shadow price of decision makers. Another issue with using market prices is that due to scarce regional market integration, they might differ across different cells and reliable data on crop-specific sub-national prices is typically hard to obtain.

where ϕ is the cell- and crop-specific yields density function *in the plots where the crop is grown* and is such that:

$$\int^{o_k^c=1} \phi(\hat{y}_k^c)\; d\hat{y}_k^c = 1$$

In order to generate the corrected GAEZ yields described in equation 8, it is not sufficient to state the farmers' maximization problem, but it is also necessary to impose a structure on the (unobserved) joint density function g^c of crop-specific yields within cells, since by definition the GAEZ dataset only provides the average. I assume that the joint density function takes the form of a multivaiate normal distribution $g^c \sim \mathcal{N}(\mu^c, \Sigma^c)$ where the vector μ^c is by definition the cell- and crop-specific yields as provided by the GAEZ dataset; i.e. $\mu = (GAEZ_1^c \ldots, GAEZ_K^c)$.

As for the variance-covariance matrix Σ^c:

$$\Sigma^c = \begin{bmatrix} \sigma_1^{c2} & \cdots & \rho_{1,k}\sigma_1^c\sigma_k^c & \cdots & \rho_{1,K}\ \sigma_1^c\sigma_K^c \\ \vdots & \ddots & \sigma_k^{c2} & \ddots & \vdots \\ \rho_{1,K}\ \sigma_1^c\sigma_K^c & \cdots & \rho_{k,K}\ \sigma_k^c\sigma_K^c & \cdots & {\sigma_K^c}^2 \end{bmatrix} \tag{9}$$

it is necessary to parametrize both cross-crops correlation factors ρ and the crop specific standard deviations within cells σ. The former are obtained as the cross-cell correlation in average yields of each crop-pair and are maintained fixed for every cell. As for the within-cell yields' standard deviations σ, they are computed as the standard deviations of the predicted crop-specific yields of cell c and the 8 bordering cells. The underlying assumption is that when there is higher dispersion in a crop's expected yield in the area (i.e. across neighbouring cells), the dispersion is also higher within cell. As a result, unlike the correlation factors ρ, these are crop and cell specific.[4]

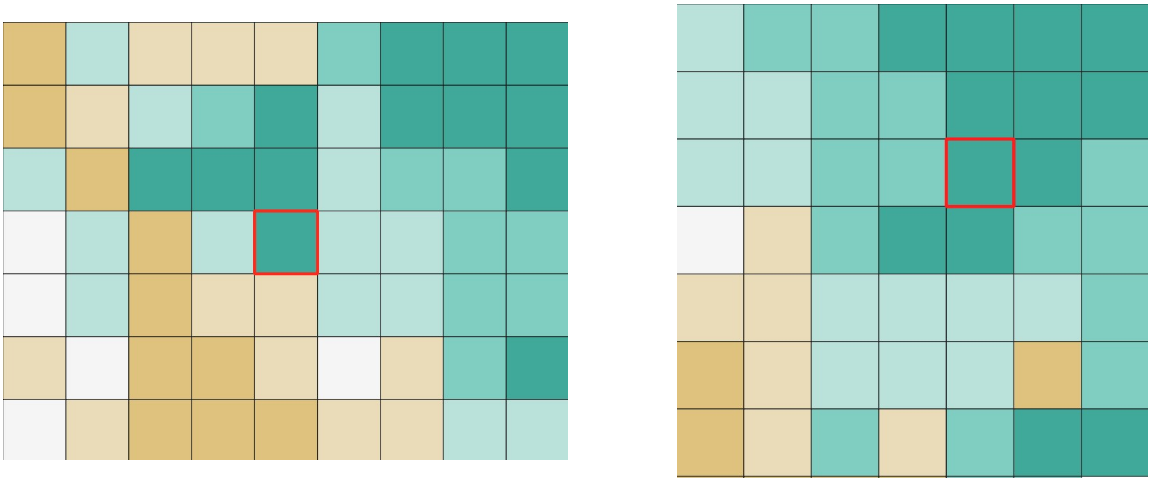
Figure 4 shows two cells with very similar maize yields (3078 and 3014 tonnes per hectare for the left and right panel respectively) but where the standard deviation differs significantly (798 and 110 in the left and right panel respectively) as a result of the larger dispersion of average yields when considering the neighbouring cells.

To sum up, this methodology allows me to correct the expected yields estimates made available by the GAEZ (which are based on the characteristics of the total farmed area in each cell) to account for heterogeneity of within-cell land quality and non-random within-cell crop choices.[5] As a result, the resulting crop and cell specific yields $\widetilde{GAEZ}_k^c$ are directly comparable to the actual yields which are by definition only available for the plots where crops

[4] It is worth pointing out that it might be the case that this might either under or overestimate the actual within cell variation in crops yields. However, in the absence of information on the distribution of within-cell yields, this allows me to fully characterize the joint density function using only available data from the GAEZ.

[5] The maximization problem is solved separately for each cell using a recursive algorithm implemented through MATLAB. For each cell, the joint yields distribution is simulated through 100,000 draws from the multivariate normal.

Figure 4: Example of high and low standard deviation cells



Source: Author's computation based on GAEZ v4.0.

are actually grown. As this methodology is based on some unverifiable assumptions on the farmers' decision making process as well as the within-cell joint yields distributions, in the empirical analysis I will consider both the original GAEZ estimates and my correction.

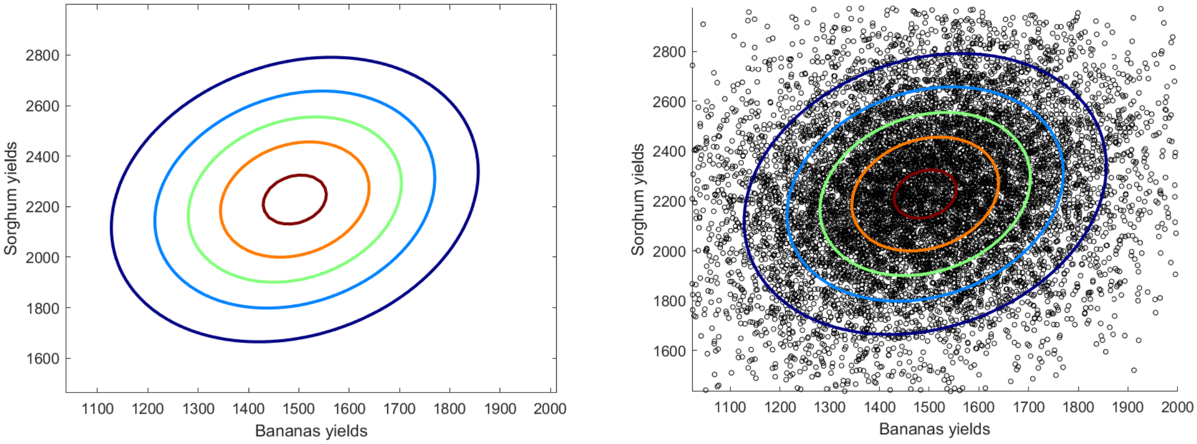
In the following subsection, the intuition of this methodology and the implications for the resulting correction is explained in the simplified case of two different crops.

3.2 The case of two crops

In the case of two crops, the correction outlined in the previous section can be readily visualized in a two-dimensional graph. For example, one can consider the case of a cell where 50% of the land is used to grow bananas and the remainder to grow sorghum. While the GAEZ provides information on the expected yields of both crops, these measures might have to be corrected to account for the fact that a) there can be heterogeneity in the crop-specific yields within land and b) the choice of growing bananas versus sorghum within the cell is non-random and likely depends on their relative crop suitability (as per point a).

For each cell considered, the GAEZ provides only the means of the joint yields distribution, while the correlation factor $\rho_{B,S}$ as well as the standard deviations of bananas (σ_B) and sorghum (σ_S) need to be parametrized. As discussed above, the former is obtained as the cross-cell correlation factors between average crop yields and the latter as the standard deviation in the expected yields of the cell under analysis and the neighbouring ones. In this case, we consider a cell with median expected yields and standard deviations for bananas ($\mu = 1,492$; s.d $= 200$) and sorghum ($\mu = 2,228$; s.d $= 309$), whose yields have a cross-cell

Figure 5: Joint density of bananas and sorghum's yields



Source: authors' computation based on median value from GAEZ v4.0.

correlation of ≈ 0.2.[6]

These parameters can be used to generate the joint density function for the two crops' yields within the cell, which is displayed in figure 5 both with and without (in right and left panel respectively) the 10,000 draws used to perform the correction numerically. As expected, due to the positive relationship between the two crop yields, the distribution develops along the south-west to north-east axis, but due to the low magnitude of the correlation factor, there is some significant dispersion with some plots being relatively suitable for one crop but not for the other.

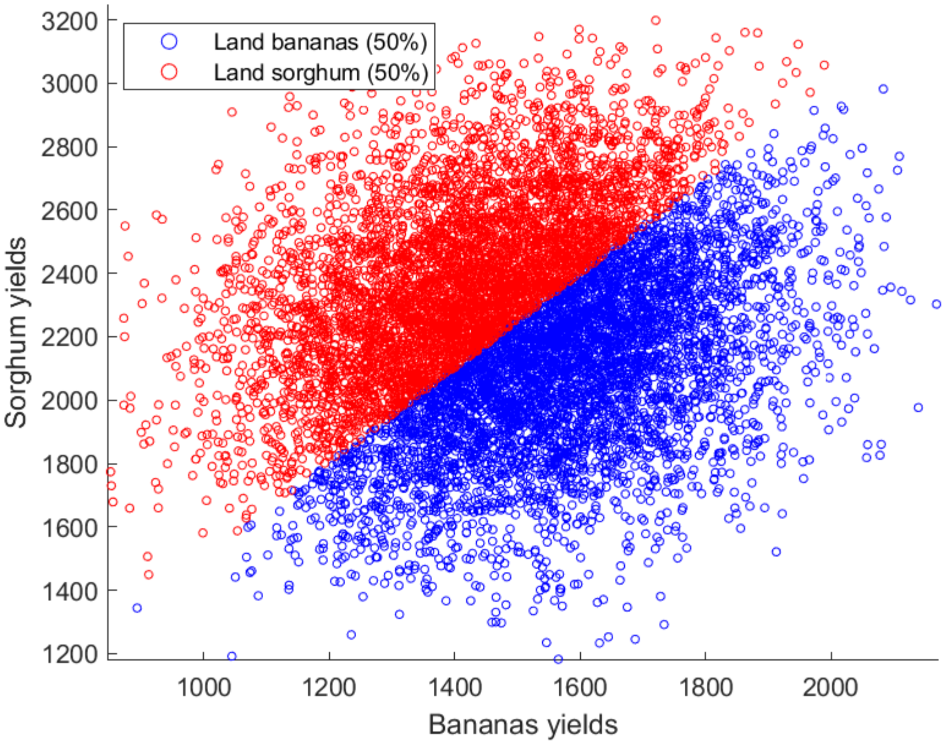
Figure 6 shows the efficient allocation of plots to the two different crops, where farmers maximize the aggregate normalized yields of the two crops (as per equation 6) and the land in the cell is equally split between them. As long as there is no perfect correlation between the two crop yields, this implies that the plots where each crop is grown are not only relatively more suitable, but also have higher expected yields than the average field in the cell in absolute value.

This is illustrated in figure 7, which shows the yields distribution of bananas and sorghum (in the left and right panel respectively) across the whole cell (the blue solid line) and only for the plots where the crop is grown (red dashed lines). The difference in the average represents the magnitude of the correction applied to the original GAEZ data to account for heterogeneity in plots within the cell and non-random allocation of land to crops. In this instance, both expected yields are revised upwards as maximizing farmers are able to obtain higher than "cell-average" yields by allocating relatively more suitable plots to each crop.

The magnitude of the correction depends on the dispersion in each crop's yield, their cor-

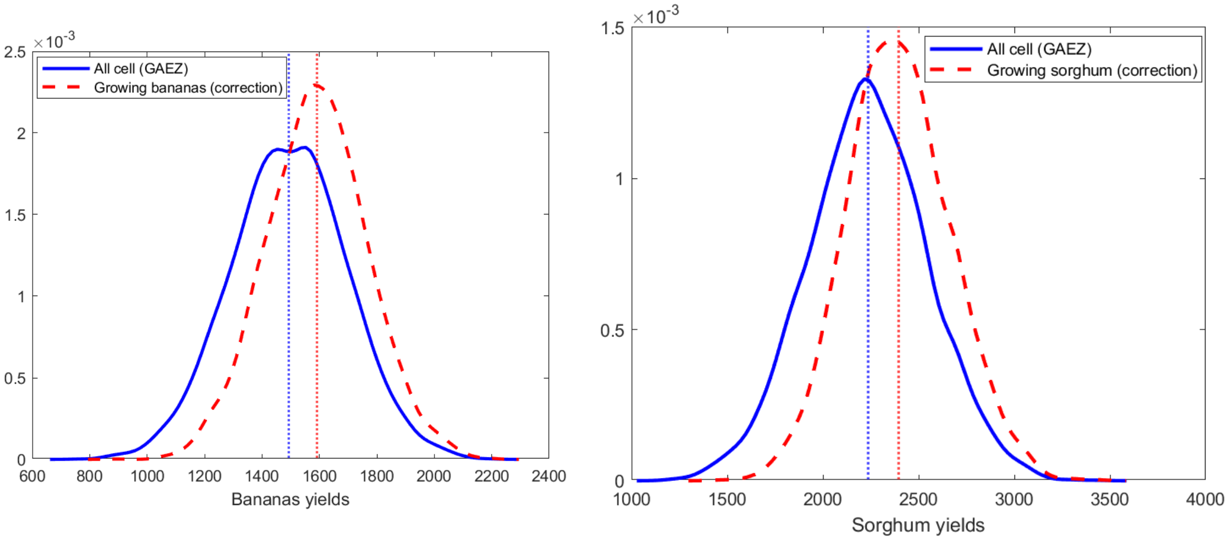
[6] Note that the two crops have among the lowest correlation among the ones considered in the analysis, indicating that good growing conditions for bananas are generally but non necessarily also good for sorghum.

Figure 6: Land allocation between bananas and sorghum (50%-50%)



Source: Author's own computation.

Figure 7: GAEZ Correction for bananas and sorghum expected yields



Source: Author's computation.

relation and the share of land devoted to each.[7] For example, where a crop is only grown in a very high share of the land in a given cell, the correction is less pronounced and vice-versa.[8] In a similar way, where the two crop yields have very high correlation, the plots with relatively higher suitability for a given crop will be evenly distributed across the crop yields distribution, and as such it is impossible for farmers to obtain significantly higher average yields by allocating crops in relatively more suitable cells.[9]

In the empirical analysis, the correction is estimated for each cell for up to 9 different crops. While the visualization and the numerical computation are not as straightforward, the underlying features of the methodology are the same. Similar to the example shown, on average the resulting expected yields *in the plots where each crop is grown* are higher than the cell-level averages, and this methodology effectively allows me to account for non-random crops allocation within cells and heterogeneity in crop suitability across plots.

4 Yields from survey data

In order to "test" the reliability of granular GAEZ predictions on crop-specific suitability, this paper compares the cell- and crop-specific yields from the agronomic model with the yields obtained in actual geolocated fields in Uganda as measured by micro-level surveys.[10] As argued before, one of the main issue with this consists in the difficulty in generating credible yields measures based on survey data (Abay et al. 2019; Abay et al. 2021; Ayalew et al. 2023). In order to partially address this, I compare the GAEZ dataset to actual yields obtained from two different surveys of Ugandan agricultural households, each presenting some advantages and disadvantages.

It is also worth pointing out that the literature on measurement errors is typically concerned with the computation of plot and/or holding specific yields in order to understand the actual nature of some empirical regularities in similar data such as the inverse size yields relationship. On the other hand, in this work the analysis is carried out mostly at the "cell-level" and as such the outputs and inputs are aggregated across households which operate in the same cells. This implies that the results are less sensitive to plot specific measurement errors. Additionally, in order to derive more credible measures of cell-specific yields, I drop the 5% most extreme plot or household specific observations from the analysis and only include in the final sample cell/crop pairs where at least 10 plot or household specific observations are available.[11]

[7] This implies that the correction is crop- and cell-specific. Note that if this were not the case, the uncorrected and the corrected GAEZ would be perfectly correlated and as such the correction would have no impact whatsoever in the regression analysis comparing actual yields with GAEZ predictions.

[8] In fact, in the limit case where the crop is grown in every plot, there is no need for any correction.

[9] In the appendix I provide two further examples illustrating these features of the correction.

[10] The country represents an useful case study as farmers typically grow a large number of crops, increasing the sample size for the analysis. Additionally, two separate surveys are available for roughly the same period, which allows me to perform a direct comparison using two different surveys.

[11] As explained in the following, in the LSMS data production and inputs are recorded at the plot and crop

4.1 Living Standard Measurement Survey

The first dataset considered is the World Banks' Ugandan Living Standard Measurement Survey (LSMS). I use three separate waves referring to 2010/11, 2011/12 and 2013/14 years.

The main advantage of this dataset is that information of agricultural output is available at the plot level. In particular, hours worked, seeds and fertilizer and other chemicals used can be attributed to a specific plot, allowing me to control for differences in input intensities. Where more than one crop is grown in the same plots, the inputs are distributed proportionally to the area of land farmed. Similarly, the output is recorded for each plot/crop pair, which results in a larger number of observations and a more precise estimation of the yields actually obtained in the fields. Other inputs and features such as self-reported land quality (on a scale from 1 to 3) and availability of the irrigation facilities are instead only available at the parcel level, where parcels typically include more than one plot/crop pair and are defined as uninterrupted land which is farmed by the household. For each wave, information on input and output for both agricultural seasons are included and are considered separately. Thus, it is possible that the same plot is used to grow different crops with different input intensity in the two seasons. In the empirical analysis, I will include yields from different seasons as separate observations.

On the other hand, the dataset also presents some criticalities. First of all, not all the plots are measured through GPS, which can result in some measurement errors where the farmers' self reported estimates are not precise. Similarly, the output is often indicated in non-standard units for which conversion measures are not always available. In the benchmark analysis, I include both parcels estimated through GPS technology and through farmers' estimation and consider median crop and production unit specific conversion factors where not available in the data.

Another relevant concern relates to the geolocation of the production units (which is crucial to match actual yields to the prediction of the GAEZ). In particular, in order not to disclose the identity of the survey participants, the available coordinates of the sample's enumeration areas were modified, so that the actual location of the households is within a 5 or 10 km radius (depending on whether they are urban or rural) from the one indicated in the data. In the benchmark analysis, I assume that the households actually reside in the cell indicated by the available coordinates, though there is a non-trivial probability that they are actually located in a neighbouring cell. Since the correction applied to the GAEZ estimates allows for some within cell heterogeneity in crop yields based on the expected yields in the neighbouring cell, potential errors in the exact location of the farms are less concerning.

In the analysis, I only consider the nine most important crops by acreage in Uganda, which accounts for nearly 95% of the total use of agricultural land once the areas left fallow are

level, while in the case of the UCA the production needs to be aggregated at the household/holding level.

dropped from the sample in the six agricultural seasons under analysis.[12] As stated above, the 5 percent most extreme yield values are dropped from the analysis in order to account for potential measurement errors from misreporting of production and/or area farmed.

The final sample consists of over 37,000 crop/plot specific observations from the six agricultural seasons under analysis, across 278 different enumeration areas and 241 10 × 10 square kilometre cells. The geographical coverage of the LSMS sample is shown in figure 8, as Uganda consists of 2,442 separate cells, the LSMS data covers about one tenth of the area. [13] Each observation indicates the crop, the production quantity, the area farmed as well as the quantity of inputs employed in the production, namely the hours of labour, organic fertilizer, pesticide, and seeds applied in the plot. Information on the self-reported land quality (on a scale from 1 to 3) and availability of irrigation facilities is also available.

Table 1 reports the average yields from the GAEZ model (uncorrected and corrected) as well as the ones obtained from the LSMS. While there are observations from over 200 cells, when restricting the sample to areas with at least 10 separate observations for each crop, the sample is reduced. While for widely grown crops like maize and cassava this does not represent a serious issue, the number of available cells remaining for relatively less widespread crops like sorghum and millet is significantly reduced. This represents a potential drawback of the analysis when considering these crops. Also, it is clear that the correction illustrated in Section 3 revises the average GAEZ yields upwards (indicating that on average farmers who allocate land non-randomly can achieve higher yields for all crops) but the correlation between uncorrected and corrected estimates is rather large. Finally, In some instances the LSMS yields appear to be significantly higher than the ones in the GAEZ. As explained above, this is realistically a consequence of the fact that while the GAEZ reports units of final dry grain/produce, LSMS observations indicate the simple weight of the produce at

[12] The nine crops considered are maize, cassava, bananas (which in turn include three different types of bananas: plantain, brewing bananas and "sweet" bananas), beans, sweet potatoes, sorghum, coffee, groundnuts and millet).

[13] The number of enumeration areas is larger than the number of cells as some enumeration areas fall within the same cell.

harvest.

Table 1: Average yields by crop - LSMS

	GAEZ yield	Corrected yield	Correlation	LSMS yield	N cells	Area share
Maize	1338	1401	0.94	1664	196	17.8%
Cassava	2794	2838	0.96	2544	173	17.3%
Bananas	1626	1641	0.95	4092	148	20.4%
Beans	457	474	0.95	818	180	14.5%
Sweet Potatoes	2068	2094	0.94	2960	176	10.0%
Sorghum	929	931	0.87	599	54	4.4%
Coffee	473	489	0.89	1265	77	6.2%
Groundnuts	366	385	0.92	979	93	5.6%
Millet	216	221	0.91	883	54	3.8%

Source: Author's calculation based on GAEZ and LSMS data. The yields represent the cross-cell average from the GAEZ data without (column 1) and with correction (Column 2) as well as from the LSMS survey. They are expressed in kilos per hectare. GAEZ and LSMS data cannot be compared directly as the output in the former is expressed in processed product while the LSMS reports the weight of the harvested crop. Data are only presented for cells where there were at least 10 plot/crop specific observations. Correlation refers to the correlation coefficient between the uncorrected and corrected GAEZ yields. Area share refers to the fraction of the agricultural area taken up by each crop in the LSMS survey.

4.2 Ugandan Census of Agriculture

The second data source used is the Ugandan Census of Agriculture (UCA) collected by the Ugandan Bureau of Statistics (UBOS) and referring to the two agricultural seasons between 2008 and 2009. The main advantage of this dataset compared to the LSMS is the much larger coverage and sample size (see figure 8). In this case, nearly 14,000 household/crop pair specific observations are included from nearly 30,000 holdings over 1308 cells.

However, the amount of single observations included is relatively lower (around 140,000) due to the fact that 1) only two agricultural seasons are available and 2) output are aggregated at the holding/season level rather than at the plot level, as information on the production is not collected for each plot but only for each separate crop grown. This implies that if an households grows the same crop on more than one plot in a given agricultural season, only one observation is generated. In a similar way, the information on the inputs used is less precise than in the LSMS. In particular, most observations only include information on the number, age and gender of workers and whether they were employed full or part-time in the farm, and information on the use of fertilizer, chemicals and irrigation availability is only included as a binary variable at the holding level.[14]

[14] Units of labour are therefore converted in male full-time adult equivalent, where part-time workers are assumed to work half the hours of a full time worker, and women and children are assigned a weight of 0.8 and 0.55 respectively based on their relative median wage as observed in the LSMS survey for 2010.

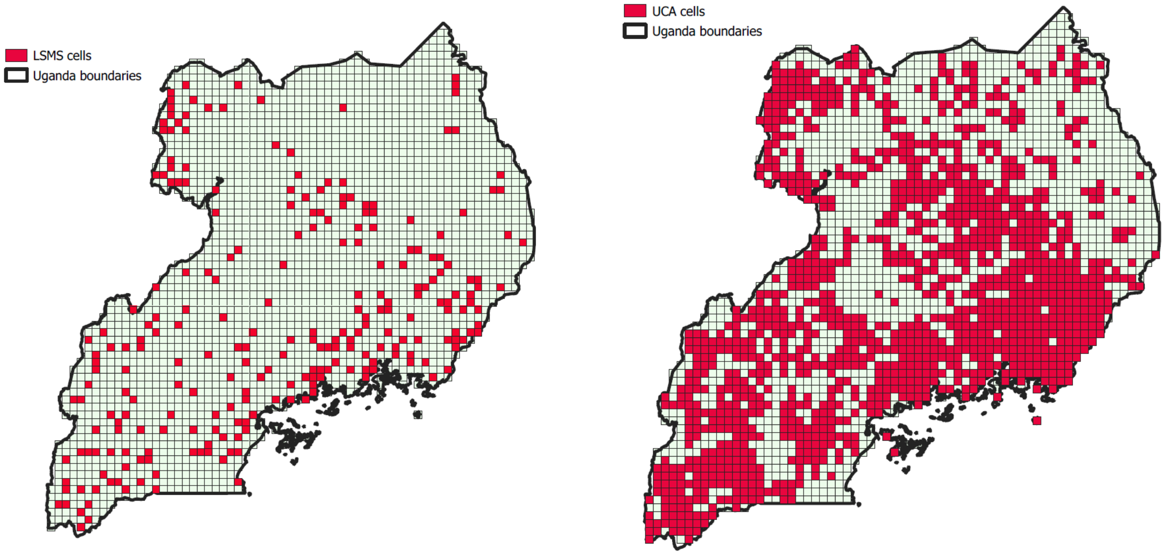
However, compared to the LSMS, there are three relevant advantages. First, the location of the enumeration areas is not modified, which implies that the correct coordinates are indicated. Second, the size of the plots is estimated using GPS techniques in the wide majority of the cases and the crop specific share devoted to each crop is measured by surveyors rather than estimated by the farmer. Finally, the output was converted in dry grain/final produce rather than reported in harvested units. The resulting average yields for the UCA are shown in table 2. It is clear that in this instance the correspondence between actual and GAEZ yields is closer. Similar to the case of the LSMS, the correction revises the GAEZ yields upward for the majority of the cell, while the correlation between the original GAEZ yields and the corrected ones is still rather high. Due to the relatively lower number of observations per cell, the number of cells for which at least 10 crop specific observations for a given crop falls more significantly than in the LSMS case. However, due to the higher number of households surveyed, the resulting final sample is larger for each of the nine crops considered.

Table 2: Average yields by crop - UCA

	GAEZ yield	Corrected yield	Correlation	UCA yield	N cells	Area share
Maize	1318	1367	0.94	1353	743	17.9%
Cassava	2855	2882	0.90	2296	314	10.5%
Bananas	1723	1739	0.90	2304	289	6.9%
Beans	459	478	0.94	739	663	11.9%
Sweet Potatoes	2245	2273	0.90	2265	306	8.0%
Sorghum	922	924	0.89	803	298	12.0%
Coffee	454	464	0.87	826	308	6.7%
Groundnuts	393	400	0.90	562	362	13.6%
Millet	241	248	0.79	881	207	12.4%

Source: Author's calculation based on GAEZ and UCA data. The yields represent the cross-cell average from the GAEZ data without (column 1) and with correction (Column 2) as well as from the LSMS survey. They are expressed in kilos per hectare. Data are only presented for cells where there were at least 10 household/crop specific observations. Correlation refers to the correlation coefficient between the uncorrected and corrected GAEZ yields. Area share refers to the fraction of the agricultural area taken up by each crop in the UCA survey.

Figure 8: Cells covered by LSMS and UCA sample



Source: Author's computation based on LSMS and UCA data.

5 Predicted vs survey yields

The core objective of this paper is to check whether there is any correlation between the yields predicted by the GAEZ model and the ones actually observed in the geolocated fields. This is done simply by estimating equation 10 for each of the nine crops considered, where the dependent variable is the actual yield measured in cell c for crop k and the independent variable is the "corrected" GAEZ yields. Cell and crop-specific controls X are included in order to account for confounding factors such as input intensities and season-specific shock. More specifically, the vector of controls includes the amount of labour/seeds, organic fertilizer and pesticide used per unit of land, as well as the availability of irrigation and self-reported land quality.[15]

$$\text{Survey Yield}_k^c = \beta_0 + \beta_1 \widetilde{\text{GAEZ}}_k^c + X_k^c \gamma + \epsilon_k^c \tag{10}$$

Since the correction of the GAEZ yields is based on assumptions on the farmers' objective function and the unobservable joint distribution of crops' yields within cells, I also estimated an alternative, naïve specification where I simply include the original GAEZ expected yields as main independent variable. As shown in tables 1 and 2, the correlation between the two tends to be rather high for every crop and in both sample, so it is perhaps unsurprising that the results are comparable across the two specifications.

[15] The variables are aggregated across different observations based on the size of the plot. In the case of the UCA, the use of fertilizer and pesticide is only included as a dummy, so rather than quantities I consider the share of land where such inputs are utilized.

$$\text{Survey Yield}_k^c = \beta_0 + \beta_1 \text{GAEZ}_k^c + X_k^c \gamma + \epsilon_k^c \tag{11}$$

The estimates for β_1, expressed in terms of standard deviations, are shown in table 3 for the LSMS sample and in table 4 for the UCA. Surprisingly, the coefficients indicate a mostly negative relationship between survey yields and the ones predicted by the GAEZ model regardless of the specification chosen, whether controls are included and the sample chosen. With the exception of sorghum in the LSMS sample, the coefficients are either insignificant or negative.

Table 3: Yields comparison - LSMS

	Correction		Naïve			
	No Controls	Controls	No Controls	Controls	N	Average obs
Maize	-0.04	-0.06	-0.09	-0.09	196	32
Cassava	-0.04	-0.09	-0.06	-0.11	173	22
Bananas	-0.27***	-0.26***	-0.27***	-0.26***	148	49
Beans	-0.23***	-0.23***	-0.29***	-0.28***	180	36
Sweet Potatoes	-0.08	-0.08	-0.05	-0.04	176	23
Sorghum	-0.27*	0.02	-0.27**	0.03	54	18
Coffee	-0.50***	-0.45***	-0.51***	-0.47***	77	24
Groundnuts	0.16*	0.23**	0.19*	0.29***	93	17
Millet	-0.43***	-0.40**	-0.46***	-0.36*	54	16

Source: Author's calculation based on GAEZ and LSMS data. The figures represent the average change (expressed in standard deviations) in the LSMS cell-specific yield for a standard deviation change in the GAEZ predicted yields. The first two columns use corrected yields to account for within crop heterogeneity and non-random crops distribution. The third and fourth use the uncorrected GAEZ yields. Controls include labour, organic fertilizer, pesticides and seeds per unit of land and average quality and irrigation availability of the plots where each crop is grown. N and average obs refer to the number of cells included in the regression and the average number of plots used to obtain the LSMS yields. Cells/crop pairs with less than 10 plots observations were dropped. The stars refer to the significance of the coefficients, *** p <0.01, ** p < 0.05, * p <0.1.

It is important to point out that, while these results suggest that the GAEZ yields are a poor predictor of the ones observed in the data, there are a number of reasons why this can be the case. First, this can be the result of measurement error in the survey yields. In spite of the effort made to reduce these concerns, it could still be the case that the survey data are a poor proxy for the actual harvests obtained by the typical farmers in the fields. Another possibility is that the correction applied to account for the intrinsically different nature of the data compared (i.e. the fact that survey yields are only observed where the crops are grown and not throughout the cell) is not sufficient/able to address the issue making the empirical analysis less reliable. Third, it might be the case that the controls included do not fully account for differences in the technology used by different farmers and that as a result the impact of the inherent suitability of land/climatic conditions (captured by β_1) cannot be captured by the regressions. This can be particularly problematic in the case of Sub-Saharan

Africa where actual technical production has been shown to be very far from the maximum potential (Adamopoulos and Restuccia 2022).

Table 4: Yields comparison - UCA

	Correction		Naïve			
	No Controls	Controls	No Controls	Controls	N	Average obs
Maize	-0.02	-0.02	-0.02	-0.02	743	30
Cassava	0.01	-0.01	0.03	0.01	314	25
Bananas	-0.04	-0.01	-0.05	-0.02	289	31
Beans	-0.11***	-0.10**	-0.11***	-0.10**	663	29
Sweet Potatoes	-0.12**	-0.08	-0.09	-0.06	306	23
Sorghum	-0.23***	-0.18***	-0.23***	-0.16***	298	22
Coffee	-0.09**	-0.15**	-0.13**	-0.18***	308	30
Groundnuts	-0.08	-0.08	-0.05	-0.05	362	19
Millet	-0.17***	-0.16***	-0.21***	-0.21***	207	18

Source: Author's calculation based on GAEZ and UCA data. The figures represent the average change (expressed in standard deviations) in the UCA cell-specific yield for a standard deviation change in the GAEZ predicted yields. The first two columns use corrected yields to account for within crop heterogeneity and non-random crops distribution. The third and fourth use the uncorrected GAEZ yields. Controls include labour, organic fertilizer, pesticides and seeds per unit of land and average quality and irrigation availability of the plots where each crop is grown. N and average obs refer to the number of cells included in the regression and the average number of plots used to obtain the UCA yields. Cells/ crop pairs with less than 10 plots observations were dropped. The stars refer to the significance of the coefficients, *** p <0.01, ** p < 0.05, * p <0.1.

Nevertheless, these findings suggest caution when considering differences among cells which are close to each other. In order to further check these results, in the following section I estimate the relationship between cell-specific land use shares and relative crop yields as measured by the GAEZ and in the actual sample. The main hypothesis is that, where farmers follow comparative advantage in making land use decisions, it is possible to compare the validity of survey-derived versus GAEZ yields by considering the empirical relationship between crops' shares and these yields.

6 Land use and comparative advantage

The results in the previous section indicate some substantial differences between the expected yields indicated by the GAEZ and ones measured based on micro-level surveys from geolocated farms. In most instances, the relationship between the two is shown to be either insignificant or negative, suggesting lower yields in areas where according to the GAEZ the agronomic conditions are more viable. However, due to the well-known issues in measuring yields on the basis of survey data, the results are not fully convincing.

It is possible to double check those findings by looking at how observed cell-level crop choices are affected by yields as measured in the GAEZ and in the surveys. More precisely, where farmers respond to agronomic comparative advantage, the spatial distribution of crops is expected to depend on their relative suitability. I test this hypothesis for the eight most commonly observed crop pairs using a relative suitability index derived alternatively from the GAEZ (equation 12) and the yields from the survey data (equation 13).[16] I use relative rather than absolute crops yields to capture the structure of comparative advantage, and consider the deciles in the distributions in order to reduce the impact of outliers (typically cells where the share devoted to one crop is very small).

$$\text{Decile}^c\left(\frac{\text{Area}_i}{\text{Area}_j}\right) = \beta_0 + \beta_1\text{Decile}^c\left(\frac{\text{GAEZ}_i}{\text{GAEZ}_j}\right) + \epsilon^c \tag{12}$$

$$\text{Decile}^c\left(\frac{\text{Area}_i}{\text{Area}_j}\right) = \beta_0 + \beta_1\text{Decile}^c\left(\frac{\text{Survey}_i}{\text{Survey}_j}\right) + \epsilon^c \tag{13}$$

The estimates for β_1 and their significance are shown in table 5 and 6 for the LSMS and the UCA sample respectively, and can be interpreted as the average change in the decile distribution of relative land share for a crop due to a unit (decile) change in the distribution in relative suitability across the considered cells.

[16] In this case, in order to have sufficient cell-level observation, I include also cells with less than 10 crop specific observations. The eight crop pairs are chosen to maximize the sample size (i.e. number of cells where both crops are grown).

Table 5: Crop choice and estimated yields - LSMS

	GAEZ	LSMS
Maize/Beans	-0.19***	0.13**
	(0.07)	(0.06)
	[0.04]	[0.02]
Maize/Sweet Potatoes	-0.10	0.16**
	(0.07)	(0.07)
	[0.01]	[0.02]
Maize/Cassava	-0.13*	0.10*
	(0.07)	(0.06)
	[0.02]	[0.01]
Cassava/Sweet Potatoes	-0.34***	-0.23***
	(0.06)	(0.07)
	[0.11]	[0.05]
Beans/Sweet Potatoes	0.04	0.03
	(0.07)	(0.07)
	[0.00]	[0.00]
Cassava/Beans	-0.16**	0.30***
	(0.08)	(0.06)
	[0.03]	[0.09]
Maize/Groundnuts	-0.08	0.05
	(0.07)	(0.07)
	[0.01]	[0.00]
Maize/Bananas	0.38***	0.32***
	(0.06)	(0.07)
	[0.15]	[0.11]

Source: Author's calculation based on GAEZ and LSMS data. Standard errors in parentheses are robust, the numbers in square brackets are the R-squared. *** p <0.01, ** p < 0.05, * p <0.1.

In the case of table 5, where the GAEZ yields are used to capture comparative advantage, the coefficient is estimated to be negative and statistically significant for 4 out of 8 of the crop pairs considered, and only positive in the case of maize vs cassava. This indicates that farmers' crop choices are often against the underlying structure of comparative advantage as captured by the GAEZ expected yields. Interestingly, when using the yields based on LSMS, the relationship is reversed in 3 instances, where the relationship turns from negative to positive. This implies that farmers' choices seem to be more aligned with agronomic comparative advantage when considering yields estimated from the LSMS survey.

Table 6: Crop choice - UCA

	GAEZ	UCA
Maize/Beans	-0.23***	0.06*
	(0.03)	(0.03)
	[0.05]	[0.03]
Maize/Sweet Potatoes	-0.14***	0.15***
	(0.04)	(0.04)
	[0.02]	[0.02]
Maize/Cassava	-0.05	0.05
	(0.05)	(0.05)
	[0.02]	[0.02]
Cassava/Sweet Potatoes	-0.05	-0.06
	(0.05)	(0.05)
	[0.02]	[0.03]
Beans/Sweet Potatoes	-0.04	0.16***
	(0.04)	(0.04)
	[0.02]	[0.05]
Cassava/Beans	-0.19***	0.13***
	(0.05)	(0.05)
	[0.04]	[0.05]
Maize/Groundnuts	0.05*	0.04
	(0.03)	(0.03)
	[0.03]	[0.02]
Maize/Bananas	0.31***	0.17***
	(0.05)	(0.05)
	[0.10]	[0.07]

Source: Author's calculation based on GAEZ and UCA data. Standard errors in parentheses are robust, the numbers in square brackets are the R-squared. *** p <0.01, ** p < 0.05, * p <0.1.

Table 6 shows a very similar relationship when considering the UCA sample. In this case, the relationship is negative in 3 out of 8 crop pairs considered and positive in only 2 in the case of the GAEZ yields, while it is positive in 5 out of 8 crop pairs (and never negative) in the case of the UCA.

These results potentially indicate that, rather than being a mere result of measurement error, the difference in actual versus GAEZ predicted yields could actually capture some measurement error in the latter. In fact, farmers' crop choices are shown to be more in line with comparative advantage where they are generated on the basis of the relative crops' yields from the survey data than from the GAEZ (where in most instances they instead seem to act against it).

It is however worth pointing out that this evidence is not conclusive. Especially in the case of smallholder farmers in developing countries, it is likely that crop choices are not fully determined by exogenous comparative advantages where lack of integration across regional markets and subsistence constraints can determine substantial deviations from the efficient crop choices (Fafchamps 1992; Omamo 1998; Morando 2022; Li 2023).

7 Concluding remarks

An increasing number of studies in economics are using newly available spatial data in order to address new research questions/exploit innovative identification strategies. The GAEZ dataset, providing granular data on expected and crop-specific yields worldwide, has witnessed a surge in use by development and trade scholars as it is able to generate counterfactual scenarios and is ”exogenous” (i.e. orthogonal to economic agents' decisions and institutions). While the underlying state-of-the-art agronomic model and the ease of access make the GAEZ an incredibly valuable resource, it is important to stress that, in the word of its own developers, its results ”should be treated in a conservative manner and at appropriate aggregation levels, which are commensurate with the resolution of the basic data”.

This study represents a first attempt to test the GAEZ yield predictions ”in the fields” by comparing its predicted cell-specific yields with the ones actually observed in geolocated farms from two different surveys in Uganda. In particular, the goal is to assess the accuracy of the GAEZ where used to compare growing conditions within areas in a relatively small country - an admittedly ambitious bar.

In order to allow for a direct comparison between GAEZ yields (based on achievable production in all the farmed area) and survey-generated yields (by construction, based only on fields where the crops are grown) I develop a theoretical model to account for non-random selection of fields within cells.

The empirical analysis shows very little or negative correlation between actual and predicted yields across cells in Uganda for the nine most widely grown crops. When studying the relationship between crop choices and the structure of comparative advantage implied by yields, I find no or negative response of relative land shares to relative crop yields when using GAEZ yields, and a mostly positive relationship when using survey-generated yields. This suggests that the lack of correspondence between GAEZ and actual yields could be due to issues in the granular-level predictions of the former rather than mere measurement errors and misreporting in the latter.

While the GAEZ remains a valuable resource for economists, this paper confirms that caution should be used where leveraging information within narrower areas as they might be subject to some issues, plausibly due to the quality of the raw data fed into the agronomic model. Further research should explore the generalizability of these findings to different contexts and attempt to identify the adequate level of aggregation of the output from the GAEZ.

References

Abay, K. A., Abate, G. T., Barrett, C. B., and Bernard, T. (2019). Correlated non-classical measurement errors,'Second best'policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics*, 139:171–184.

Abay, K. A., Bevis, L. E., and Barrett, C. B. (2021). Measurement error mechanisms matter: Agricultural intensification with farmer misperceptions and misreporting. *American Journal of Agricultural Economics*, 103(2):498–522.

Adamopoulos, T. and Restuccia, D. (2022). Geography and agricultural productivity: Cross-country evidence from micro plot-level data. *The Review of Economic Studies*, 89(4):1629–1653.

Ayalew, H., Chamberlin, J., Newman, C., Abay, K. A., Kosmowski, F., and Sida, T. (2023). Revisiting the size–productivity relationship with imperfect measures of production and plot size. *American Journal of Agricultural Economics*.

Berman, N., Couttenier, M., and Soubeyran, R. (2021). Fertile ground for conflict. *Journal of the European Economic Association*, 19(1):82–127.

Costinot, A. and Donaldson, D. (2016). How large are the gains from economic integration? Theory and evidence from US agriculture, 1880-1997. Technical report, National Bureau of Economic Research.

Costinot, A., Donaldson, D., and Smith, C. (2016). Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy*, 124(1):205–248.

Dell, M., Jones, B. F., and Olken, B. A. (2014). What do we learn from the weather? The new climate-economy literature. *Journal of Economic literature*, 52(3):740–798.

Dias, M., Rocha, R., and Soares, R. R. (2023). Down the river: Glyphosate use in agriculture and birth outcomes of surrounding populations. *Review of Economic Studies*, page rdad011.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198.

Fafchamps, M. (1992). Cash crop production, food price volatility, and rural market integration in the third world. *American journal of agricultural economics*, 74(1):90–99.

FAO & IIASA (2022). Global agro-ecological zones (GAEZ v4)-faq version 2.0.

Fernández-Villaverde, J., Koyama, M., Lin, Y., and Sng, T.-H. (2023). The fractured-land hypothesis. *The Quarterly Journal of Economics*, 138(2):1173–1231.

Fischer, G., Nachtergaele, F. O., van Velthuizen, H., Chiozza, F., Francheschini, G., Henry, M., Muchoney, D., and Tramberend, S. (2021). Global agro-ecological zones (GAEZ v4)-model documentation.

Galor, O. and Özak, Ö. (2016). The agricultural origins of time preference. *American Economic Review*, 106(10):3064–3103.

Gibson, J., Olivia, S., and Boe-Gibson, G. (2020). Night lights in economics: Sources and uses. *Journal of Economic Surveys*, 34(5):955–980.

Gibson, J., Olivia, S., Boe-Gibson, G., and Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, 149:102602.

Gollin, D. and Udry, C. (2021). Heterogeneity, measurement error, and misallocation: Evidence from African agriculture. *Journal of Political Economy*, 129(1):1–80.

Li, N. (2023). In-kind transfers, marketization costs and household specialization: Evidence from Indian farmers. *Journal of Development Economics*, page 103130.

Lowes, S. and Montero, E. (2021). The legacy of colonial medicine in Central Africa. *American Economic Review*, 111(4):1284–1314.

Mayshar, J., Moav, O., and Pascali, L. (2022). The origin of the state: Land productivity or appropriability? *Journal of Political Economy*, 130(4):1091–1144.

Morando, B. (2022). Aggregate productivity and inefficient cropping patterns in Uganda. *Journal of Productivity Analysis*, 58(2-3):221–237.

Nunn, N. and Qian, N. (2011). The potato's contribution to population and urbanization: Evidence from a historical experiment. *The quarterly journal of economics*, 126(2):593–650.

Omamo, S. W. (1998). Transport costs and smallholder cropping choices: An application to Siaya District, Kenya. *American Journal of Agricultural Economics*, 80(1):116–123.

Sotelo, S. (2020). Domestic trade frictions and agriculture. *Journal of Political Economy*, 128(7):2690–2738.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, 56(1):79–107.

Appendix

As explained in section 3.2, the magnitude of the correction applied to the original GAEZ yields in order to account for within-cell heterogeneity and non-random crop selection depends on the share of land devoted to each crop and on the correlation of the two crops' yields. In this section, this is illustrated by looking at the case where the same crops as the original example (bananas and sorghum) are considered, but 80% of the land is used to grow bananas, and by considering a different choice of crops (bananas and sweet potatoes) which present a larger correlation factor.

When 80% of the land is used to grow bananas, the crop allocation of the plots in the cell is described in figure A1. Similar to the case of figure 6, the plots are allocated depending on their relative suitability. However, since the vast majority of the fields are devoted to bananas, the "cut-off" to grow sorghum (in terms of relative suitability) is larger.[17]

The resulting correction is shown in figure A2, where the change in the average yield between all the plots in the cell (blue solid line) and the plots where the crops are grown (red dashed line) represents the magnitude of the correction. Clearly, similar to the case of figure 7, in both instances the expected in the plots where the crop is actually grown are higher than the average across all fields in the cell. However, in this instance the correction is considerably larger in the case of sorghum and barely noticeable in the case of bananas. This is a result of the latter crop being grown in the wide majority of the fields, which implies that the expected yields is converging to the one observed when considering the whole cultivated area. On the other hand, since sorghum is only grown in one fifth of the plots, only the ones with a sizeable comparative advantage (which implies also an absolute advantage where there is not perfect correlation between the two crops' yields) are used to grow sorghum, resulting in a larger divergence between expected yields across the cell and for the plots where the crop is actually grown.

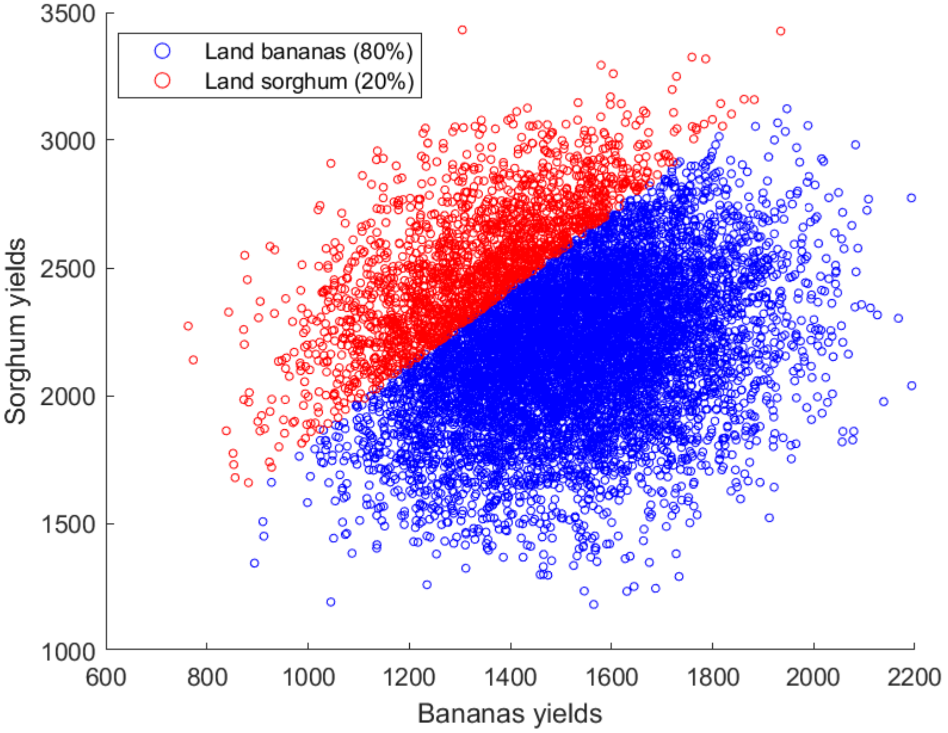
The correlation between two crops' yields plays a central role in determining the magnitude of the correlation. Intuitively, as the correlation factor approaches 1, the structure of comparative advantage between two crops converges to a perfect linear relationship where the trade-offs between the crops is the same in each field. In this extreme situation, farmers are indifferent in terms of the allocation of plots to different crops (once the share to be used for each crop is fixed) and as such the expected yields in fields where the crops are grown (which are selected randomly) are the same as the expected yields in the cell.

This can be shown by examining the case of two crops whose yields are highly correlated: bananas and sweet potatoes ($\rho \approx 0.8$). In this case, the joint distribution in an hypothetical cell is shown in figure A3:[18]

[17] Notice that when solving the problem numerically in the empirical analysis, the crops' shares are exogenously given as they are chosen to match the ones actually observed in the data.

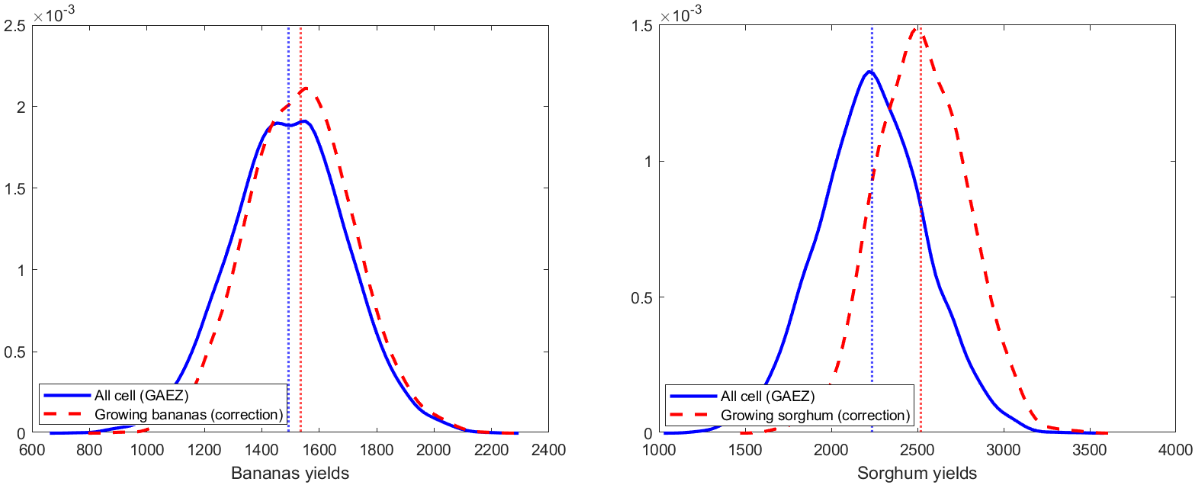
[18] This is generated by setting mean and standard deviation equal to the median across cells of both crops. Namely, the mean yield is 1142 for bananas and 777 for sweet potatoes. The standard deviation is 200 for bananas and 119 for sweet potatoes, and the correlation factor is 0.784.

Figure A1: Land allocation between bananas and sorghum (80%-20%)



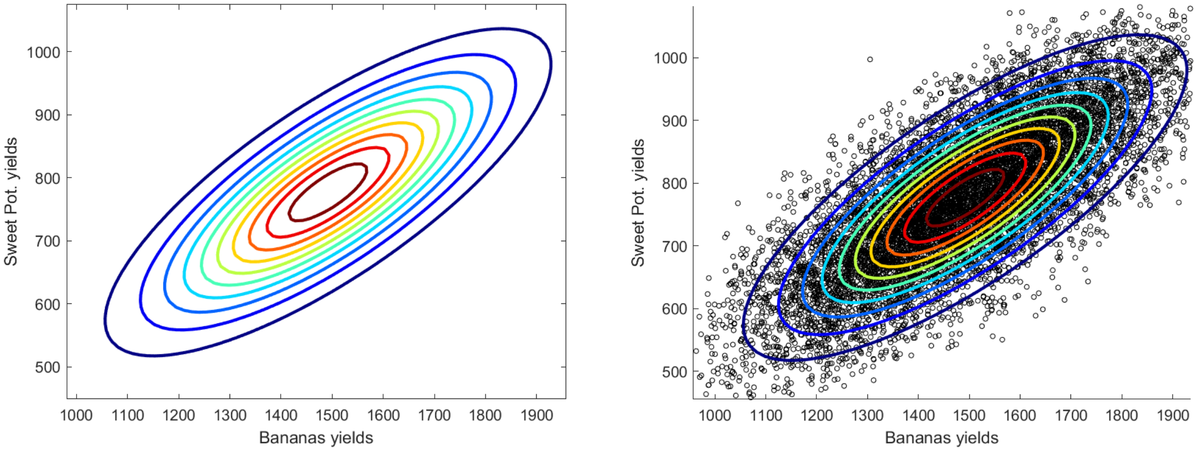
Source: Author's own computation.

Figure A2: GAEZ Correction for bananas and sorghum expected yields (80%-20%)



Source: Author's computation.

Figure A3: Joint density of bananas and sweet potatoes' yields

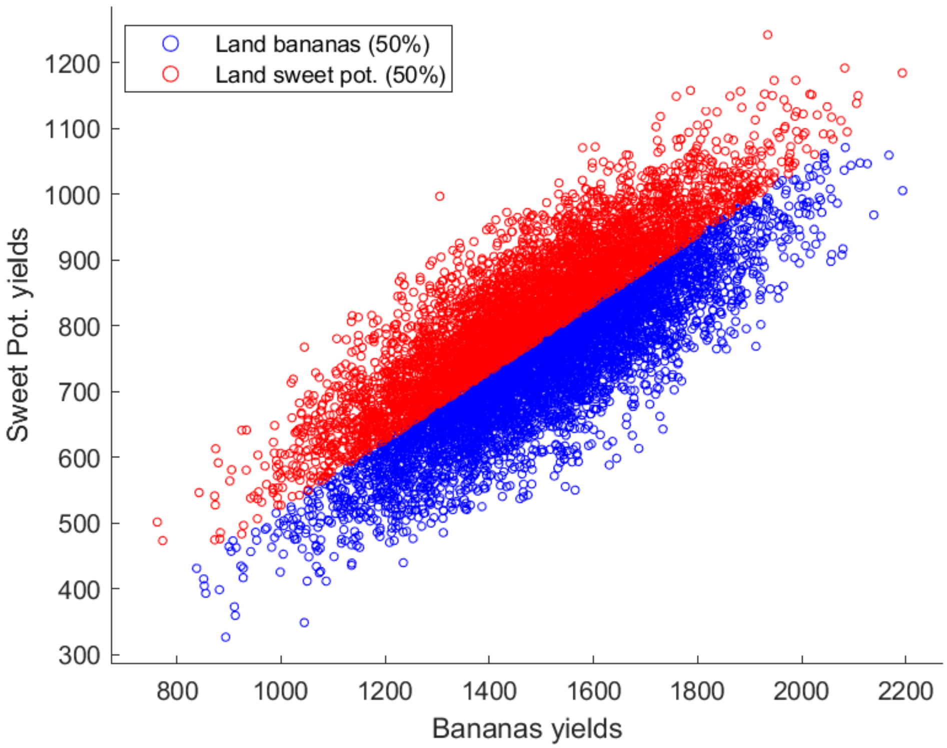


Source: authors' computation based on median value from GAEZ v4.0.

In this instance, the relationship between the two crops' yields is very tight indicating that the growing conditions that affect favourably bananas are similarly beneficial to sweet potatoes. Thus, when the farmers allocate the two crops across plots in the same cells (the optimal allocation is shown in figure A4), there is relatively little margin to improve the aggregate yields as there are few plots where comparative advantage for either crop is proounced.

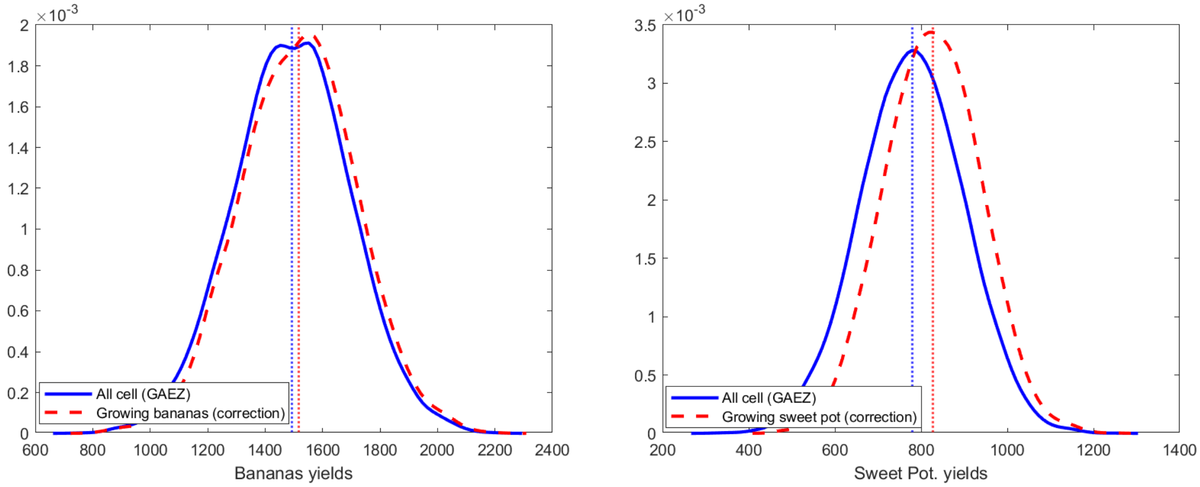
As a result, the correction to be implemented in this case is less pronounced. As shown in figure A5, the average yield obtained in the plots where each crop is grown is very similar to the average yield across the cell. This reflects the fact that relatively more suitable fields for either crop are found across the whole crop's yields distribution due to the high correlation between yields.

Figure A4: Land allocation between bananas and sweet potatoes



Source: Author's own computation.

Figure A5: GAEZ Correction for bananas and sweet potatoes expected yields



Source: Author's computation.