FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

# A Geometric Analysis of Technological Heterogeneity in the Agricultural Sector: Evidence from Maize in Tanzania.

*Research Paper 12*
*January 2022*

**Authors: Karim Nchare, Marcel Vitouley, Heidi Kaila, Yanyan Liu.**

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY RESEARCH, CAPACITY, AND INFLUENCE (PRCI)
MICHIGAN STATE UNIVERSITY
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI
Cornell University
ReNAPRI
UNIVERSITY OF GHANA
ISSER Knowledge for Development
KU
RIS Research and Information System for Developing Countries
विकासशील देशों की अनुसंधान एवं सूचना प्रणाली

# FOOD SECURITY POLICY RESEARCH, CAPACITY, AND INFLUENCE (PRCI) RESEARCH PAPERS

This Research Paper series is designed to disseminate timely research and policy analytical outputs generated by the USAID-funded Feed the Future Innovation Lab for Food Security Policy Research, Capacity, and Influence (PRCI) and its Associate Awards and Buy-ins. The PRCI project is managed by the Food Security Group (FSG) of the Department of Agricultural, Food, and Resource Economics (AFRE) at Michigan State University (MSU) and implemented by a consortium of three major partners: the International Food Policy Research Institute (IFPRI), Cornell University, the Regional Network of African Policy Research Institutes (ReNAPRI), and the Institute for Statistical, Social, and Economic Research (ISSER) at the University of Ghana. The MSU consortium works with governments, researchers, and private sector stakeholders in Feed the Future focus countries in Africa and Asia to co-create a global program of research and institutional capacity development that will enhance the ability of local policy research organizations to conduct high-quality food security policy research and to influence food security policy more effectively while becoming increasingly self-reliant.

The papers are aimed at researchers, policy makers, donor agencies, educators, and international development practitioners. Selected papers will be translated into other languages.

Copies of all PRCI Research Papers and Policy Briefs are freely downloadable in pdf format from this link. Copies of all PRCI papers and briefs are also submitted to the USAID Development Experience Clearing House (DEC) at this link and to AgEcon Search.

# STATEMENT OF SUPPORT

This research is made possible by the generous support of the American people through the United States Agency for International Development (USAID) through funding to the Feed the Future Innovation Lab for Food Security Policy Research, Capacity, and Influence (PRCI) under grant 7200AA19LE000001. The contents are the responsibility of the study authors and do not necessarily reflect the views of USAID or the United States Government.  Published by the Department of Agricultural, Food, and Resource Economics, Michigan State University, Justin S. Morrill Hall of Agriculture, 446 West Circle Dr., Room 202, East Lansing, Michigan 48824, USA.

## Authors

Karim Nchare: African School of Economics - kfogam@africanschoolofeconomics.com

Marcel Vitouley: African School of Economics - mvetouley@africanschoolofeconomics.com

Heidi Kaila: Cornell University - hkk35@cornell.edu

Yanyan Liu: International Food Policy Research Institute - Y.Liu@cgiar.org

## Authors' Acknowledgments

We acknowledge the technical support provided by PRCI, and are grateful for comments and suggestions from Christopher Barrett, the STAARS and STAAARS+ fellows and mentors, as well as all attendees of feedback workshops. We are solely responsible for the content of this paper.

# ABSTRACT

This paper presents a new framework to measure farm-level heterogeneity, and productivity change, and to study the rate and direction of technical change within an agricultural sector. Building on the seminal works of Hildenbrand (1981) and Dosi et al. (2016), we show how, while relaxing most of the standard assumptions from production theory, discrete geometry is an effective tool for productivity analysis and technical change in agricultural economics. We apply the framework to a rich panel data from maize farmers in Tanzania to investigate the dynamics of technical heterogeneity and agricultural productivity growth.

**Keywords:** Discrete Geometry, Heterogeneity, Agricultural Productivity, Technical Change, Tanzania
**JEL** Codes: D22, O12, O55, Q12

# TABLE OF CONTENTS

# List of Tables



# List of Figures

# Acronyms

| | |
|---|---|
| AFRE | Agricultural, Food, and Resource Economics |
| DEA | Data envelopment analysis |
| FSG | Food Security Group |
| GPS | Global Positioning System |
| IFPRI | International Food Policy Research Institute |
| LSMS | Living Standards Measurement Study |
| LSMS-ARENA | Advancing Research on Nutrition and Agriculture project Phase II |
| LSMS-ISA | Living Standards Measurement Study–Integrated Surveys on Agriculture |
| MSU | Michigan State University |
| PRCI | Food Security Policy Research, Capacity, and Influence |
| R4D | Research for Development |
| ReNAPRI | Regional Network of African Policy Research Institutes |
| SSA | sub-Saharan Africa |
| STAAARS+ | Structural Transformation of African and Asian Agriculture and Rural Spaces |
| STAARS | Structural Transformation of African Agriculture and Rural Spaces |
| ths | Thousands |

# Introduction

Improving agricultural productivity is fundamental to achieving sustainable development, reducing poverty, and enhancing the living standards of most people in sub-Saharan Africa (SSA). According to a report from the African Development Bank (AfDB, 2013), agriculture accounts for at least 40 percent of exports, 30 percent of gross domestic product (GDP), up to 30 percent of foreign exchange earnings, and ensures employment for 70 to 90 percent of the labor force in sub-Saharan Africa (SSA). Despite moderate increases, SSA's agricultural productivity is still growing at approximately half the average rate of developing countries (Pratt, 2015). In recent years, many studies have demonstrated the heterogeneity of the smallholder production environment and technology (see Vanlauwe et al. 2019 anc citations therein). Using rich panel data from farms in Tanzania and Uganda, Gollin and Udry (2021) find that measurement error and unobserved heterogeneity in the characteristics of farm plots account together for a large proportion of the dispersion in measured productivity. This finding is vindicated by Maue et al. (2020), who establish that about half of this dispersion can be accounted for by measurement error in the output. After correcting for measurement error, Maue et al. (2020) find that dispersion in productivity among farms is significant and persistent over time. The results in both papers question the common implications of observed dispersion, such as the importance of misallocation of factors of production. For example, Gollin and Udry (2021) suggest that the potential for efficiency gains through the reallocation of land among farms and farmers may be relatively modest.

Yet agricultural research for development (R4D) that aims to identify and assess alternatives for increasing productivity has not consistently tailored its approaches to such heterogeneous conditions. As agricultural policies have focused on very general recommendations, there is increasing recognition that the heterogeneity of agroecologies and farms and farmers needs locally adapted solutions and tailored approaches (Giller et al., 2011). In fact, Smallholder farming environments in sub-Saharan Africa are characterized by (i) variable soil fertility conditions within short distances, (ii) variable access to resources for farming families within the same communities, (iii) variable enabling conditions for an increase in agricultural production per unit of inputs, (iv) including access to agro-inputs, markets and extension services. Moreover, farmers have (v) varying access to production resources including land, labor, and cash, (vi) different production objectives including food for subsistence and products for the market, (vii) varying capacities to absorb risk inherent to alternative management practices, with poorer households being more risk-averse, and (viii) diverse attitudes to farming and the role farming plays within their overall livelihood (Vanlauwe et al., 2019). Therefore, we need an analytical framework that allows each farm to be different in terms of its production technology. In order to formulate more tailored agricultural policies, this paper aims to answer the following questions: (i) how to measure an agricultural sector heterogeneity when farms are different over several dimensions? (ii) how to measure productivity and technical changes over time while accounting

for farms' heterogeneity? (iii) What factors are associated with the measured agricultural productivity?

To answer these questions we rely on a geometric approach to production analysis based on the seminal works of Hildenbrand (1981) and Dosi et al. (2016). Hildenbrand (1981) proposes an agnostic and data-oriented approach in which one can represent a production unit (within an industry) in the input-output space. The production possibility set of the industry is then represented geometrically by the space formed by the finite sum of all the line segments, linking the origin and the points representing each production unit withinthe industry (a Zonotope). Exploiting the properties of Zonotopes, Dosi et al. (2016) show how to obtain rigorous measures of intra-industry heterogeneity and productivity without imposing functional form or input substitution assumptions on the data like in standard production function analysis.

Our proposed approach is implemented using detailed farm and plot survey data from maize production in Tanzania. Agriculture contributes almost 30% to the country's GDP and is the predominant source of income for approximately 75% of the population (Van Dijk et al., 2017). Maize is the main staple food crop with 5.9 million tons produced in 2018 (FAO), and it is consumed and cultivated all over the country under varying agro-climatic and socio-economic conditions. Therefore, analyzing technological heterogeneity and productivity of households' maize production in Tanzania is policy-relevant in different areas including poverty reduction, food security, and the spatial distribution of agricultural crops. It is also methodologically interesting for us as it is typically expected that farmers growing maize are homogeneous in terms of production techniques.

We make two main contributions in this paper. The first contribution is methodological as we introduce in the agricultural economics literature, a new approach to investigate agricultural production when micro-data is available. Assessing intra-industry heterogeneity and within-industry productivity, as differentiated from measuring stated inefficiencies vis-a-vis some frontier distinguishes the geometric approach from the efficiency frontier approach (Farrell, 1957; Simar and Zelenyuk, 2011; Battese and Coelli, 1995). Although both approaches are non-parametric in nature, the emphasis of the latter is on measuring production units' inefficiency in terms of the distance from the efficient frontier, recovered by enveloping the data: the more distant a production unit from the frontier, the less efficient it is (Dosi et al., 2016). Furthermore, the geometric approach captures the variation of production techniques adopted by production units in any economic sector and allows one to determine the rate and direction of productivity change. Recent developments in robust frontiers analysis (Simar and Zelenyuk, 2011) have resolved many shortcomings of the traditional deterministic approach,such as the sensitivity to measurement errors and outliers: given that all farms are compared to the frontier, misspecification of it would heavily bias the entire analysis. Note, however, that our computed values for heterogeneity (volume of the zonotope) and productivity (angle from projections in the zonotope) are measures, not estimates. However, we will check in the empirical analysis how robust this productivity index is to the presence of measurement errors in one of the inputs (in our case, land size). Finally, the Zonotope approach allows for multi-output analysis and overcomes the shortcomings of existing

multi-output frontier models such as the Data envelopment analysis (DEA), the Stochastic distance function frontier, and the Stochastic ray frontier which generalized the multi-output ray production function using a polar-coordinate angle output vector (Löthgren, 1997).

The second contribution is empirical as we provide rigorous measures of farms' agricultural productivity and agricultural sector heterogeneity while accounting for farms' technological diversity. These measures are then used to identify the drivers of agricultural productivity inthe maize sector in Tanzania. The remainder of the paper is organized as follows. Section 2 presents in detail the discrete geometric approach to production analysis used in the paper. Section 3 describesthe dataset and provides some descriptive statistics. Section 4 provides empirical results associated with the discrete geometric approach, and Section 5 concludes.

# A Geometric Approach to Production Analysis

We start this section by providing an empirical motivation for our methodological approach. To illustrate the phenomenon of persistent technological heterogeneity among households growing maize (single crop plot), we provide some empirical evidence focusing on the labor productivity (log) distribution by agroecological zones (AEZs). Labor productivity is defined as the ratio of total maize production over total days of labor. Details and summary statistics on these two variables are presented in Section 3. We consider five agro-ecological zones: Central Zone (Dodoma, Singida, Tabora), Coastal Zone (Dar es Salaam, Lindi, Morogoro, Mtwara, Pwani), Northern Zone (Arusha, Kilimanjaro, Manyara, Tanga), Lake Zone (Geita, Kagera, Mara, Mwanza, Shinyanga, Simiyu), Southern Highlands Zone (Iringa, Katavi, Kigoma, Mbeya, Njombe, Rukwa, Ruvuma)[1]. Figure A.1 represents the empirical distributions[2] (pooled and by rounds) of labor productivity in maize production by AEZs in Tanzania. It shows the coexistence of households with persistently different levels of productivity across and within agroecological zones. The observed heterogeneity in labor productivity in the maize sector is striking, exhibiting a ratio top to bottom greater than 5 to 1 (in logs). This is evidence that household-level maize growing techniques are not derived from the same production function. The empirical densities are also consistent with the spatial distribution of agricultural crops in Tanzania (See Figure A.2) as Zones where maize farming is dominant (Northern Zone, Southern Highlands Zone) display higher labor productivity on average.

In the remaining of this section, we provide a brief outline of the geometric approach to production analysis that will be used in the research project. Similar to Koopmans (1977), Hildenbrand (1981), and Dosi et al. (2016), the production activity, as describing the actual technique of firm/farm *i*, is represented by a vector

$$a_i = \left(\alpha_1^i, \dots, \alpha_l^i, \alpha_{l+1}^i, \dots, \alpha_{l+m}^i\right) \in \mathbb{R}_+^{l+m} \qquad \textbf{(1)}$$

The production unit *i* produces $(\alpha_{l+1}^i, \dots, \alpha_{l+m}^i)$ units of output during the current period by means of $(\alpha_1^i, \dots, \alpha_l^i)$ units of input. Then, one can characterize the short-run production possibilities of an industry with N units during the current period by a finite family of production activity vectors $\{a_i\}_{1 \le i \le N}$. Any production activity vector $\{a_i\}$ is associated with a line segment.

$$[0, a_i] = \{x_i a_i \mid x_i \in \mathbb{R}, 0 \le x_i \le 1\} \qquad \textbf{(2)}$$

[1] We do not consider the Western Zone since in our sample, it is only represented by a small number of households from the Kigoma region. Therefore, we include the Kigoma region in the Southern Highlands Zone

[2] estimated using Epanenchnikov Kernel

Hildenbrand (1981) defines the short-run total production set associated with the family $\{a_i\}_{1\leq i\leq N}$ as the Minkowsky sum of line segments generated by production activities $\{a_i\}_{1\leq i\leq N}$

$$Y = \sum_{i=1}^{N}[0, a_i] = \{y \in \mathbb{R}_+^{l+m} \mid y = \sum_{i=1}^{N} \emptyset_i a_i, 0 \leq \emptyset_i \leq 1\} \quad \textbf{(3)}$$

$Y$ is also called the Zonotope generated by the vectors $\{a_i\}_{1\leq i\leq N}$. Let $D$ denote the projection of $Y$ on the input space $\mathbb{R}_+^l$.

$$D = \{V \in \mathbb{R}_+^l \mid \exists X \in \mathbb{R}_+^m \, s.t. \, (V, X) \in Y\} \quad \textbf{(4)}$$

Hildenbrand (1981) then defines the short-run efficient production function $F : D \rightarrow \mathbb{R}_+^m$ as

$$F(V) = \max\{X \in \mathbb{R}_+^m \mid (V, X) \in Y\} \quad \textbf{(5)}$$

This definition implies that the maximum total output in an industry is achieved by allocating, without any restrictions, the level $(V_1, \ldots, V_l)$ of inputs in the most efficient way over the individual production units within the industry. However, the frontier associated with this production function is uninformative on the actual technological set-up of the whole industry and therefore, could not be the focal reference both for a normative or positive analysis (Hildenbrand, 1981).

From the Zonotope Framework, Dosi et al. (2016). define the main diagonal of a Zonotope Y as the line linking the origin $0 = (0, \ldots, 0)$ with its opposite vertex in Y. Because this diagonal expresses both the amount of inputs employed and outputs produced by the industry, it is called the production activity of the industry. In terms of vector, it is given by:

$$d_y = \sum_{i=1}^{N} \alpha_1^i, \ldots, \sum_{i=1}^{N} \alpha_l^i, \sum_{i=1}^{N} \alpha_{l+1}^i, \ldots, \sum_{i=1}^{N} \alpha_{l+m}^i \quad \textbf{(6)}$$

If all the firms in one industry were to use the same technology, their input-output ratio would be proportional implying that all the vector firms would lie on the same line. In this case (minimum heterogeneity case), the associated Zonotope would be of zero volume. Conversely (maximum heterogeneity case), the industry contains some firms with almost zero inputs but sufficient outputs and others with a large quantity of inputs but little outputs. In such a case, the generated Zonotope is close to a parallelotope. Building on these two extreme cases, Dosi et al. (2016) derive rigorous measures of heterogeneity and productivity.

Let $A_{i_1,\dots,i_{l+m}}$ be the matrix whose rows are vectors $\{a_{i_1},\dots,a_{i_{l+m}}\}$ and $\Delta_{i_1,\dots,}\Delta_{i_{l+m}}$, its determinant. The volume of the Zonotope $Y$ is given by

$$Vol(Y) = \sum_{1\le i_1\le\cdots,\le i_{l+m}} |\Delta_{i_1,\dots,}\Delta_{i_{l+m}}| \qquad (7)$$

It is a good candidate to assess heterogeneity within an industry as small volume corresponds to minimum heterogeneity and large volume corresponds to maximum heterogeneity. However, $Vol(Y)$ grows as the number of firms grows and also depends on the units in which inputs and outputs are measured. To overcome these shortcomings, Dosi et al. (2016) define heterogeneity as the ratio of the volume $Vol(Y)$ over the volume $Vol(P_y) = \prod_{i=1}^{N} |a_i|$ of an industry $P_y$ with production activity $d_y = \prod_{i}^{N} |a_i|$. This normalized ratio is called the Gini volume:

$$G(Y) = \frac{Vol(Y)}{Vol(P_y)} \tag{8}$$

We will now use the notation for the one-output case $(a_i \in \mathbb{R}_{+}^{l+1})$ but the concept presented here can easily be extended to the multiple-outcome case. The measure of the efficiency of the industry is the angle formed by the industrial production activity vector $d_y$ with the space generated by all inputs. This is because the higher the angle, the more the industry is able to produce more output with the same quantity of inputs. The measure of productivity fora given industry with $N$ production units is given by the tangent of that angle:

$$P = t_g\left(\Theta_{l+1}(d_y)\right) = \frac{\sum_{i=1}^{N} \alpha_{l+1}^{j}}{||pr_{(l+1)}(d_y)||}$$

where $||v||$ represent the norm of vector $v$ and $pr_{(l+1)}(a_i) = (a_1^i, \dots, a_l^i)$ for all $a_i \in \mathbb{R}_{+}^{l+1}$.

The framework allows us to compute the elasticity of substitution and to understand under what circumstances does the entry of a new production unit increase or decrease the heterogeneityof a given industry. To empirically applied this framework, we will use the Zonotope Stata command developed by Cococcioni et al. 2019.

**Farm-level measure of productivity and heterogeneity.** Similar to the aggregate level, the measure of the productivity of a farm/household *i*, with technology $a_i$, is given by the tangent of the angle formed by the vector $a_i$, with the space generated by all inputs. The measure of the heterogeneity of a farm/household *i*, with technology $a_i$, is given by the tangent of the angle formed by the projection of $a_i$ $pr_{(l+1)}(a_i)$ in the space generated by all inputs with the projection of $d_y(pr_{(l+1)}(d_y))$ on the same space. It measures to which degree the individual input combinations diverge from the industry average combination. The graphical intuition behind productivity and heterogeneity measurements at the farm level is illustrated in Figure 2.

# Empirical Strategy

We consider the following relationship between output and inputs within pure-stand maize plots:

$$Y_{iht} = (Ld_{iht}, La_{iht}) \tag{10}$$

where $Y_{iht}$ is the total quantity of maize harvested on plot *i* by household *h* in survey round *t* and measured in kilograms, $Ld_{iht}$ is the total planted area on plot *i*, and $La_{iht}$ is the total pre-harvest labor on plot *i*. It is measured as the total number of person-days spent on pre-harvest activities by either hired laborers or their own household members. Similar to Gollin and Udry (2021), the production function of Equation 10 is a simplification as it abstracts away from the multistage process associated with maize farming, which also includes the preparation of labor and land. Other potential inputs such as irrigation or machinery, commercial fertilizer, or other agrochemicals are excluded as only a small fraction of Tanzanian maize pure-stand plots use them.[33] Equation 10 with two inputs is also convenient for the application of the Geometric approach as it allows us to visualize graphically the shape of zonotopes in a three-dimensional space (one output and two inputs). Finally, we focus on maize as a single crop per plot to avoid the difficulty of measuring yield as physical quantities in presence of inter-cropping.

We apply the geometric approach to derive nonparametric measures of productivity at the plot level and technological heterogeneity measures at the zone level. Finally, we perform a comparative analysis of results derived from our approach with those obtained from the efficient frontier approach, the DEA analysis by comparing productivity changes obtained from the geometric approach with the Malmquist Index generated from the DEA analysis.

The next step in our empirical analysis is to use agricultural productivity obtained from the geometric analysis as the outcome variable in the following regression:

$$O_{ihzt} = \gamma X_{ihzt} + \delta_{zt} + \varepsilon_{ihzt} \tag{11}$$

where $O_{ihzt}$ is the agricultural productivity of plot *i* from household *h* in zone *z* onround *t*. $X_{ihzt}$ is a vector of plot, household, and community-level characteristics including access to markets, community infrastructure, household demographics, weather, soil, and land quality variables. $\delta_{zt}$ is zone-round fixed effects to control for time-variant zone-level unobservable including zonal policy shocks. Finally, $\varepsilon_{ihzt}$ is the error term. We cluster the standard errors

[3]Irrigation is used on less than 2% of plots. We do include inputs such as fertilizer (both organic and chemical used on 15% of plots, and pesticides on 11% of plots on the second step of the analysis, where we consider a wider range of factors related to agricultural productivity.

at the community level. The parameter of interest is $\gamma$, which captures the effects of time-variant factors, $X_{ihzt}$, on plot productivity.

We conclude this section by discussing the implications of land size measurement errors for our regression analysis. Indeed, there is important empirical literature documenting the presence of land size measurement errors (See Dillon et al. (2019)) in the Living Standards Measurement Study–Integrated Surveys on Agriculture (LSMS-ISA) at the World Bank which is the data set used in this paper. As an input variable in the geometric approach, the fact that land size suffers from measurement errors implies that the obtained productivity index might be mismeasured. As this productivity measure is the dependent variable inour regression analysis, we must access how it could introduce biases in our estimates. If we assume that the error introduced by land size on the productivity index is classical, we will still obtain unbiased and consistent estimates but larger standard errors. However, as shown by Abay, Abate, Barrett, and Bernard (2019) and Abay, Bevis, and Barrett (2021), productivity measures are likely to suffer from nonclassical measurement errors associated with plot size which could result in severe bias in our estimates. Abay et al. (2019) argue that using an improved measure of the dependent variable to partially correct for potential bias is preferable to no correction. Building on that suggestion, we present and compare our empirical analysis estimates obtained with productivity measures generated using the self-reported and the GPS-measured plot sizes respectively.

# Data

## Data source

For the analysis, we use a nationally representative panel data set from Tanzania collected in four rounds[4]. These data were collected by the National Bureau of Statistics of Tanzania with the support of the World Bank's program on Living Standards Measurements Surveys LSMS-ISA. The survey includes data on all plots cultivated by the household including which household member(s) manage the plot, and which supply labor on the plot, as well as detailed information on inputs used and output harvested during the long rainy season.[5]

The survey also includes detailed characteristics of the households and their farming activities. We have information on the household size, each household member's age, education level, and relationship to the household head, as well as information on their supply of labor for household farming activities. Additionally for each household, the data includes plot-level information on crops cultivated, inputs used on plot and soil characteristics such as soil type and quality, and information on erosion. The data is merged with climate variables such as measures of rainfall. The decision-making unit in our analysis is the farming household.

We supplement our main dataset with LSMS-ARENA, an LSMS-ISA data compilation provided by the Advancing Research on Nutrition and Agriculture project Phase II (ARENA-II) at IFPRI. ARENA Phase II is an ongoing project from 2018 to 2020. One of ARENA's innovations is to merge a wide range of GIS indicators on agriculture, climate, demography, and infrastructure with LSMS surveys based on the latitude and longitude coordinates at the survey clusters.[6]

## Variable construction

**Variables in the first step analysis.** The first step of analysis is conducted at the plot level. Our dataset comprises all pure-stand maize plots, a total of 3,188 plots among 1,870 households across the four rounds. We remove some implausible observations from the sample. First, we drop households where the yield is larger than 8000 kg/ha.[7] We also remove implausibly large plots, which we determine using the following criteria: We drop plots where the land size is higher than the 99 percentile when the land has implausibly low yield (below 1 percentile). Such observations also need to rely on self-reported land size such that there is no GPS information

[4] The sample design allows analysis at four primary domains of inference, Dar es Salaam, other urban areas on mainland Tanzania, rural mainland Tanzania, and Zanzibar. The representativeness is similar
across all the four rounds of data (2008, 2010, 2012 and 2014).

[5] The data and documentation are available at http://surveys.worldbank.org/lsms/programs/integrated-surveys-agriculture-ISA/

[6] More information on ARENA II is available at https://www.ifpri.org/project/advancing-research-nutrition-and-agriculture-arena.

available. After these adjustments, our sample size is 2674 pure-stand maize plots among 1691 households.

We use the crop-plot level dataset from ARENA to create variables for the first step that are at the level of a maize purestand plot. This dataset includes all four rounds and both seasons (long rainy, and short rainy) in each round. First, out of a total of 62,868 observations, we delete all the crop-plot level observations where the crop is a tree crop. This reduces the sample size to 32,539. Next, removing the observations recorded in the short rainy season, we are left with 25,211 observations. Next, dropping all other observations than maize purestand leaves us with 3,188 such plots. These plots are distributed across 1,870 households in an unbalanced 4-wave panel, that has between 617 and 1083 observations per round. Next, there are a total of 130 observations where the quantity harvested is either missing (129 observations) or zero (1 observation). We delete these observations as well. We are left with 3,058 purestand plots. Finally, we restrict the dataset so that we drop observations where the self-reported land size is above the 99th percentile (all 27 plots where land sizeis over 20 acres, the maximum being implausibly high, 600), are deleted. These 27 plots are distributed across 25 households. Next, we also drop the observations with implausibly low yield. We delete the observations that are in the 1st percentile of the yield distribution. This removes all plots where yield (kg/acres) is below 9 and as low as 0.5. This removes 27 observations distributed across 26 households. We are left with a sample of 2,674 purestand maize plots distributed across a total of 1,691 households in 4 rounds. Now in each round, we have between 536 and 878 plots. Inside this dataset, we have created variables for yield, land size (both self-reported and a variable where we use the GPS and self-reported when GPS is unavailable), the amount of maize produced measured in kg, and the value of maize (using ARENA variable "estimated value of harvest in ths").

After creating these variables, we use the ARENA dataset which has input data at the plot level to create our input variables. In this step, we create variables for fertilizer use on the plot. We find that 83 percent of plots do not use any inorganic fertilizer, and 87 percent of plots do not use any organic fertilizer.[8] We create a variable for the quantity of total (inorganic and organic) fertilizer used on the plot in kg. We also create variables for family labor, and hired labor (a total in land preparation, weeding, and harvesting measured as days for both variables).

For several other plot-specific variables, we use the raw LSMS data, which we then merge with ARENA data for final analysis. These variables are used in the second step of the analysis. We have dummy variables for good and average soil quality, the omitted category being poor soil quality. We have variables denoting the distance to the road, distance to home, and distance to the market (in km). We also use the LSMS data at the plot level to construct variables for the plot manager's age and age squared, and whether the plot manager is female. Furthermore, we also use the LSMS data to construct variables at the household level (household size, female household head, years of schooling of household head, asset index, access to electricity, whether household experienced any shock and any asset losses). In the first step, we consider land size and labor as inputs in maize farming. For labor,

we consider both family labor and hired labor on plots, measured as days spent inland preparation, weeding, and harvesting. Due to some high- value outliers in the family labor variable, we censor the upper tail of the family labor distribution, such that we censor implausibly high values to an upper limit defined the following way: We consider household members eligible to work on the farm to be all household members who are 6 years or older to derive the maximum number of people who can supply family labor on the farm. We consider one person's maximum labor supply to be 120 days, corresponding to working each day for four months, the length of the agricultural season of the long rainy season. These limits together yield a household-specific maximum labor supply. There are 4 plots whose labor supply extends this limit, we replace those outlier values with this upper limit created. The final labor variable included in the analysis includes both family and hired labor with non-zero labor days. We do not censor outlier values of hired labor, as hired labor does not have a theoretical upper bound. We also dropped islands and the final sample size for our first step analysis comprises 1677 households growing maize on 2542 pure-stand plots.

**Variables in the second step analysis.** For the second step, we consider the plot, household, community-level characteristics, and geographical variables as factors influencing agricultural productivity at the plot level. The summary statistics are reported in Table 1. We can see that the average household size is approximately five members, with household head completing about 4.8 years of education.

The asset index is constructed using factor analysis of dummies denoting ownership of durable assets. These include household assets not related to farming activities, such as television or car, and the index is thus an indicator of household durable wealth. We also include total acres of land owned by households in the second step, this variable includes all land and is not restricted to maize production, nor farming. We can see that as many as 82.9 percent of households reported having been affected by a shock on assets or income. It is also worth noticing that a household has on average two to three plots.

Next we turn to plot-specific variables starting with soil quality. Most maize purestand plots are perceived to have either good (52 percent) or average (44 percent) soil quality, .14 percent of plots are declared to have erosion problems while only 6 percent of plots on average are steep. About 80 percent of plots are solely owned by the farmer while just 8 percent of them have a title. Regarding the usage of agrochemicals, pesticides are applied on about 8.5 percent of the plots on average, while 17.7 percent of plots have received inorganic fertilizers. Organic fertilizer use is similarly frequent at 13.6 percent. Almost 34 percent of plots use improved or purchased seeds. While 53.5 percent of plots have more managers than one, just 24.6 percent of plots have a female manager as the main manager. The average age of the main plot manager is 46 years.

At the community level, we include the length of the growing period (days), the average rainfall and temperature, as well as the elevation (above/below sea level) and the slope (gradient of steepness). The variables slope and elevation are the means for plots in that household. We also

include temperature and rainfall in the analysis. We take the monthly averages for the months from February to July that correspond to the agricultural season in question. The variables used are the demeaned version of this variable, where we subtract the mean (the country-level mean pooled across rounds) from the community average for that plot-round observation.

# Results

## Measures of agricultural productivity and technical heterogeneity

We start by providing a graphical illustration of Zonotopes. Figure 3 illustrates Zonotopes for maize growers in the Pwani (left) and Kigoma (right) regions using pooled data from 2008 to 2014. Based on the shape of their respective Zonotopes, it is clear that Pwani is more heterogeneous than Kigoma, but Kigoma is more productive. This graphical guess is vindicated by productivity coefficients which are respectively 4.52 and 5.09.

Next, in Table 2 we present the zonal level technological heterogeneity, agricultural productivity, and labor intensity measures. The first observation is that the chosen normalization approach for the volume of the zonotope seems to be effective, as there is apparently no relationship between the number of plots-generators and the heterogeneity coefficients. Our findings seem to be in accordance overall with the map of the spatial distribution of crops in Tanzania described in Figure 4. On average, the Northern and the Southern Highlands zones are consistently the most productive maize growing zones. On average, the Coastal zone exhibits the highest level of technological heterogeneity among maize growers while the Lake zone on average is the most labor-intensive.

In Table 3 we investigate the growth of productivity, heterogeneity, and labor intensity, thus the dynamics of the maize farming sector over time. Overall, productivity growth seems to be accompanied by an eventual decrease in labor intensity and a volatile heterogeneity. The first column corresponds to the growth rate between the years 2008 and 2010, the second between the years 2010 and 2012, and similarly for the third column. For example, between 2008 and 2010 Productivity grew by 6.67 percent across all zones. Across the sample, productivity growth is positive and increases at an accelerating rate. Heterogeneity displays a more volatile pattern: it first decreases, then increases, and then decreases again overall zones. The labor intensity growth rate shows a declining pattern going from a 22 percent increase between 2008-2010 down to a 7 percent decline between the last two waves.

However, when looking at regions separately, no clear zonal patterns emerge over time. For example in terms of heterogeneity we see first a decrease in heterogeneity followed by an increase, or an increasing and then decreasing growth rate in all zones except in the central zone where the growth rates are positive throughout. In terms of productivity, we also see both negative and positive growth rates for all zones except Coastal, where productivity growth is always positive at an accelerating rate. In terms of labor intensity, similar up and down movements can be

observed, there is no single zone where the growth rate is not both negative and positive during the time period.

In Table 7, we compare our computation of productivity change as presented in Table 3 to those obtained with an efficiency measure, the Malmquist Index. In order to run the analysis on the same set of data, we balance the panel of plots in any of the three couples of years over which we investigate productivity change (2008-2010, 2010-2012, and 2012-2014). This ensures that in the estimates of Table 7 we consider the same number of observations both in the computation of our proposed measure of agricultural productivity changes in rates of growth for the Malmquist index. The two measures are often in agreement in suggesting a productivity increase (decrease) when productivity change is positive (negative) and Malmquist index is smaller (bigger) than one.

Finally, we compare the productivity measures obtained using self-reported plot size with the one obtained using GPS-measure plot size. Figure 5 shows the distribution of self-reported and GPS plot sizes for pure-stand maize plots in Tanzania. We can observe significant differences across these distributions. However, quite remarkably, the productivityindices computed using these different plot size measurements have almost identical empirical distributions (Figure 6). This finding suggests the potential robustness of the geometric approach to measurement errors in the input variables.

## Factors associated with higher agricultural productivity

In the first-step, we derive agricultural productivity measures at the plot level. To identify the factors associated with higher agricultural productivity, we regress productivity measures on plot, household, community-level characteristics, and weather conditions. Table 4 reports the estimated coefficients of equation (11), using the logarithm of plot-level productivity as the dependent variable. From columns 1 to 4, we sequentially include different sets of explanatory variables. Column 1 only controls for soil quality and other land characteristics. Column 2 adds the land manager's gender and age and other manager charachteristics. Column 3 further includes household demographics and welfare indicators. Column 4 is our full model which further includes terrain characteristics (elevation, slope) and weather controls. In all columns, we control for zone-specific year fixed effects and cluster the standard errors at the village level.

Our interpretation is based on the full model in Column 4 while noticing that the coefficient estimates are largely robust across all model specifications. The results suggest that the good quality of the soil has a significant and positive effect on productivity. As matter of fact, the soil being of good and average quality increases productivity by respectively 53.97 and 26.19 percent, respectively, compared to a plot with the soil of bad quality. A 1km increase in the distance from plot to road is associated with 14.7 percent increase in productivity. Futhuremore, an increase of 1km in the distance from the plot in the distance from the plot to the market is associated with a

4.3 percent increase in agricultural productivity. This result is consistent with the findings from Gollin and Udry (2021). We could explain this by the fact that when a plot is further from the market the manager is more inclined to invest more efforts and resources to compensate for the time and the transportation cost associated with the plot. Using inorganic fertilizer is associated with increase in agricultural productivity by 33.33 percent. While using improved seeds is associated with a 12.4 percent decrease in productivity.

We spot a U-shape relationship between the age of the manager and the dependent variable. Everything else being constant, an increase of 1 household member is correlated with 2.3 percent decrease in agricultural productivity. Asset index and farm equipment index are also positively associated with maize productivity. We also detect an inverted U-shape relation between productivity and the number of plots held by the household. The length of the growing period and the (demeand) temperature are negatively associated with productivity (resp. 0.4 and 6 percent). This suggests that the more days a given crop takes to grow the less productivity it yields. All the variables cited above are significantly associated with productivity.

Although the remaining set of variables is not significant in our regression, it is worth noticing that most of them correspond to findings in the existing literature. An average plot tends to have higher productivity if it has access to electricity, land with erosion problems, and the managers use pesticides in the main crop season. Alternatively, it tends to have lower productivity if it has land in steep locations but also if the household encountered shocks.

These results are mostly consistent with prior reasoning and the previous literature. Households are more efficient when they are equipped with higher productive assets. Households with the higher assets are also less likely to face budget constraints and can apply inputs at the right times to improve efficiency. Defect soil and land conditions are important factors associated with lower efficiency (Abay et al., 2019). Higher rainfall is associated with higher productivity in drought-prone locations. The inverse relationship between landholding and farm productivity has long been documented in the literature

To understand which groups of variables are more correlated with the distribution of the agricultural productivity index, we decompose the variance explained by the regression (measured by R-squared) into contributions over particular groups of regressors using Shapley values. They represent the mean marginal contribution of each group of variables to the overall model R-squared. The decomposition is presented in Table 5 and we observe that plot-level variables have the highest mean marginal contribution, explaining 29.16% of the observed variation, highlighting the importance of environmental conditions at the plot level. Managerial, household, and community variables have almost similar contributions explaining respectively 15%, 26%, and 13% of the variation.

Finally, as a robustness check exercise, we compare estimates obtained using productivity index generated by self-reported land size with those obtained using the productivity index generated by GPS-measured land size. Overall the results are similar and the estimates are of similar

magnitudes. The same patterns are observed when comparing the decomposition of variance explained in both regressions (See Table 5 and Table 9)

# Conclusion

This study introduces a new approach to measuring agricultural productivity while accounting for technological heterogeneity among farmers. The Zonotope approach allows the quantification of agricultural productivity and technological heterogeneity without relying on typical a priori functional form assumptions associated with methods such as the Stochastic Frontier Analysis widely use in the empirical literature. Empirical analysis hints at the robustness of the Zonotope approach to the presence of measurement errors in inputs (land size). Our findings suggest that even a homogeneous sector like maize production can conceal a variety of agricultural practices persistent over time which must be accounted for when designing agricultural policies to improve yields and households' welfare. We compare our measures of productivity change to those obtained with efficiency measures from Data Envelopment Analysis (Malmquist Index) and observe that the two measures are often in accordance. Using a regression analysis with the agricultural productivity index as the dependent variable, we find that the socio-economic, environmental, and geographical factors associated with higher agricultural productivity are consistent with findings in the existing literature. Finally, there are two potential extensions for this study. First, the Geometric approach can accommodate multiple-output production technology which could allow us to investigateinter-cropping and inputs allocation efficiency. Second, the Geometric approach can integrate the entry and exit of production units thus allowing us to investigate the diffusion of new agricultural technology.

# Tables

**Table 1: Summary statistics of maize pure stand plot characteristics**

| | 2008 | 2010 | 2012 | 2014 | Total |
|---|---|---|---|---|---|
| Maize harvested quantity(kg) | 554.8 | 476.4 | 579.1 | 727.6 | 572.1 |
| | (779.5) | (633.3) | (971.3) | (1114.1) | (884.5) |
| Land size | 2.256 | 2.208 | 2.530 | 2.443 | 2.369 |
| | (2.386) | (2.535) | (2.898) | (2.808) | (2.685) |
| Total days of labor (days) | 84.08 | 67.13 | 74.88 | 76.22 | 74.89 |
| | (82.48) | (63.67) | (74.77) | (76.67) | (74.14) |
| Soil quality is good | 0.524 | 0.433 | 0.471 | 0.411 | 0.461 |
| | (0.500) | (0.496) | (0.499) | (0.492) | (0.499) |
| Soil quality is average | 0.440 | 0.482 | 0.457 | 0.494 | 0.467 |
| | (0.497) | (0.500) | (0.498) | (0.501) | (0.499) |
| Having erosion problem | 0.140 | 0.165 | 0.119 | 0.143 | 0.140 |
| | (0.347) | (0.371) | (0.324) | (0.351) | (0.348) |
| Being steep | 0.0548 | 0.0511 | 0.0350 | 0.0243 | 0.0417 |
| | (0.228) | (0.220) | (0.184) | (0.154) | (0.200) |
| Distance from plot to home (Km) | 3.175 | 3.987 | 6.067 | 6.098 | 4.895 |
| | (5.280) | (20.49) | (31.88) | (23.42) | (23.73) |
| Distance from plot to the road (Km) | 1.900 | 2.027 | 2.452 | 2.683 | 2.261 |
| | (2.546) | (3.533) | (4.669) | (6.956) | (4.564) |
| Distance from plot to the market (Km) | 8.195 | 14.24 | 13.61 | 10.52 | 12.11 |
| | (8.990) | (18.57) | (16.24) | (14.78) | (15.67) |
| Solely owned by the household | 0.858 | 0.847 | 0.848 | 0.565 | 0.799 |
| | (0.349) | (0.361) | (0.359) | (0.496) | (0.401) |
| Plot has a title | 0.0416 | 0.0838 | 0.126 | 0.0375 | 0.0810 |
| | (0.200) | (0.277) | (0.332) | (0.190) | (0.273) |
| Use of Organic Fertilizer on the plot | 0.147 | 0.107 | 0.153 | 0.137 | 0.136 |

| | 2008 | 2010 | 2012 | 2014 | Total |
|---|---|---|---|---|---|
| | (0.355) | (0.309) | (0.360) | (0.344) | (0.343) |
| Use of Inorganic Fertilizer on the plot | 0.144 | 0.190 | 0.180 | 0.190 | 0.177 |
| | (0.351) | (0.393) | (0.384) | (0.393) | (0.382) |
| Use of Pesticides on the plot | 0.110 | 0.0795 | 0.0654 | 0.0993 | 0.0846 |
| | (0.313) | (0.271) | (0.247) | (0.299) | (0.278) |
| Improved seeds | 0.136 | 0.459 | 0.388 | 0.442 | 0.338 |
| | (0.343) | (0.500) | (0.488) | (0.497) | (0.473) |
| Multiple managers on the plot | 0.432 | 0.570 | 0.599 | 0.481 | 0.535 |
| | (0.496) | (0.495) | (0.490) | (0.500) | (0.499) |
| Main manager is a female | 0.265 | 0.250 | 0.209 | 0.289 | 0.246 |
| | (0.442) | (0.433) | (0.407) | (0.454) | (0.431) |
| Age of the main plot manager | 46.22 | 46.50 | 47.09 | 45.42 | 46.45 |
| | (15.47) | (15.38) | (15.99) | (15.00) | (15.55) |
| Age of the main manager (squared) | 2375.7 | 2398.5 | 2473.3 | 2287.2 | 2399.2 |
| | (1579.7) | (1554.3) | (1646.5) | (1499.2) | (1582.1) |
| Individual years of schooling | 4.768 | 4.794 | 4.891 | 4.991 | 4.857 |
| | (3.158) | (3.236) | (3.326) | (3.478) | (3.294) |
| Number of manager | 1.459 | 1.642 | 1.641 | 1.494 | 1.578 |
| | (0.553) | (0.619) | (0.561) | (0.531) | (0.577) |
| Household size | 5.514 | 5.665 | 6 | 5.294 | 5.680 |
| | (3.040) | (3.016) | (3.768) | (3.006) | (3.300) |
| Household head school years | 4.741 | 4.776 | 4.835 | 4.947 | 4.819 |
| | (3.121) | (3.241) | (3.332) | (3.476) | (3.290) |
| Having access to electricity | 0.0340 | 0.0469 | 0.0619 | 0.0883 | 0.0566 |
| | (0.181) | (0.212) | (0.241) | (0.284) | (0.231) |
| Asset index (asset dummies) | -0.363 | -0.399 | -0.358 | -0.376 | -0.374 |
| | (0.414) | (0.458) | (0.582) | (0.603) | (0.522) |

| | 2008 | 2010 | 2012 | 2014 | Total |
|---|---|---|---|---|---|
| HH, Any shock (0/1) | 1 | 0.813 | 0.754 | 0.812 | 0.829 |
| | (0) | (0.391) | (0.431) | (0.391) | (0.377) |
| HH, asset losses (0/1) | 0.567 | 0.385 | 0.336 | 0.322 | 0.393 |
| | (0.496) | (0.487) | (0.473) | (0.468) | (0.488) |
| Number of plots | 2.739 | 3.021 | 2.904 | 2.402 | 2.813 |
| | (1.383) | (1.670) | (1.669) | (1.229) | (1.556) |
| Number of plot (squared) | 9.412 | 11.91 | 11.22 | 7.276 | 10.33 |
| | (9.837) | (14.77) | (15.07) | (7.994) | (13.08) |
| Agricultural extension index (Source based) | 0.253 | 0.161 | 0.130 | 0.159 | 0.169 |
| | (0.496) | (0.394) | (0.376) | (0.401) | (0.415) |
| Farm implements index (diversity) | 1.401 | 1.482 | 1.827 | 2.073 | 1.686 |
| | (0.920) | (1.063) | (1.292) | (1.223) | (1.174) |
| Slope (gradient of steepness that measured in degree) | 2.044 | 1.835 | 1.802 | 1.762 | 1.853 |
| | (1.686) | (1.548) | (1.502) | (1.759) | (1.603) |
| Length of growing period in days | 199.7 | 201.4 | 201.2 | 202.7 | 201.2 |
| | (24.22) | (23.51) | (24.48) | (21.85) | (23.71) |
| Demeaned average temperature | -0.734 | 0.309 | -0.137 | 0.607 | -3.64e-08 |
| | (2.112) | (2.070) | (2.023) | (2.002) | (2.100) |
| Demeaned average rainfall | 11.64 | 6.596 | -11.94 | -0.831 | -9.35e-08 |
| | (22.09) | (31.56) | (15.57) | (29.20) | (26.45) |
| Observations | 2542 | | | | |

Note: Mean coefficients. Standard deviations are in parentheses.

**Table 2: Measures of Productivity, Heterogeneity, and labor intensity among Regions and Years in Tanzania**

| | Year | | | |
|---|---|---|---|---|
| **Zones** | **2008** | **2010** | **2012** | **2014** |
| **Agricultural Productivity** | | | | |
| Central | 5.247 | 5.537 | 7.828 | 6.098 |
| Coastal | 4.165 | 4.303 | 5.869 | 10.674 |
| Lake | 5.557 | 6.047 | 5.818 | 5.424 |
| Northern | 6.319 | 8.428 | 8.246 | 13.961 |
| S.Highlands | 8.486 | 9.301 | 8.873 | 11.081 |
| All zones | 6.590 | 7.030 | 7.634 | 9.028 |
| **Technological Heterogeneity** | | | | |
| Central | 0.161 | 0.164 | 0.189 | 0.196 |
| Coastal | 0.206 | 0.152 | 0.213 | 0.259 |
| Lake | 0.204 | 0.178 | 0.187 | 0.144 |
| Northern | 0.184 | 0.175 | 0.226 | 0.128 |
| S.Highlands | 0.193 | 0.175 | 0.198 | 0.198 |
| All zones | 0.204 | 0.186 | 0.220 | 0.199 |
| **Labor intensity** | | | | |
| Central | 0.026 | 0.036 | 0.042 | 0.029 |
| Coastal | 0.026 | 0.026 | 0.027 | 0.049 |
| Lake | 0.037 | 0.041 | 0.041 | 0.031 |
| Northern | 0.031 | 0.041 | 0.039 | 0.035 |
| S.Highlands | 0.024 | 0.029 | 0.026 | 0.025 |
| All zones | 0.027 | 0.033 | 0.034 | 0.031 |

**Table 3: Rates of Growth of Productivity and Heterogeneity among Regions and Years in Tanzania**

| Zones | Year 2010 | 2012 | 2014 |
|---|---|---|---|
| **Productivity** | | | |
| Central | 5.522 | 41.382 | -22.108 |
| Coastal | 3.310 | 36.406 | 81.857 |
| Lake | 8.825 | -3.788 | -6.771 |
| Northern | 33.376 | -2.155 | 69.305 |
| S.Highlands | 9.607 | -4.608 | 24.884 |
| All zones | 6.677 | 8.590 | 18.257 |
| **Heterogeneity** | | | |
| Central | 1.991 | 15.043 | 3.489 |
| Coastal | -26.049 | 39.870 | 21.379 |
| Lake | -12.838 | 5.365 | -22.919 |
| Northern | -4.668 | 29.179 | -43.543 |
| S.Highlands | -9.326 | 13.361 | 0.039 |
| All zones | -8.562 | 18.048 | -9.465 |
| **Labor intensity** | | | |
| Central | 39.788 | 15.433 | -29.474 |
| Coastal | -1.673 | 3.807 | 83.139 |
| Lake | 11.056 | -1.788 | -24.042 |
| Northern | 30.218 | -3.204 | -12.370 |
| S.Highlands | 23.509 | -11.042 | -3.664 |
| All zones | 22.144 | 2.276 | -7.157 |

**Table 4: Fixed-effects regressions with Productivity**

| | Log of productivity | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Soil quality is good | .5121*** | .5195*** | .5142*** | .5397*** |
| | (.0938) | (.0953) | (.0929) | (.0889) |
| Soil quality is average | .2585*** | .2651*** | .2439*** | .2619*** |
| | (.0777) | (.0810) | (.0824) | (.0859) |
| Having erosion problem | .0737 | .0780 | .0472 | .0504 |
| | (.0562) | (.0535) | (.0557) | (.0522) |
| Being steep | -.2021* | -.2382** | -.2170** | -.1273 |
| | (.1174) | (.1055) | (.0940) | (.1094) |
| Distance from plot to home (Km) | -.0007 | -.0009 | -.0012 | -.0015 |
| | (.0017) | (.0016) | (.0017) | (.0018) |
| Distance from plot to the road (Km) | .0127* | .0118* | .0126** | .0147*** |
| | (.0067) | (.0067) | (.0063) | (.0056) |
| Distance from plot to the market (Km) | .0068*** | .0057** | .0049** | .0043* |
| | (.0022) | (.0023) | (.0024) | (.0025) |
| Solely owned by the household | -.1100* | -.0447 | -.0667 | -.0535 |
| | (.0589) | (.0687) | (.0713) | (.0682) |
| Plot has a title | -.1085 | -.1071 | -.1675* | -.1595 |
| | (.0955) | (.0941) | (.0982) | (.0983) |
| Use of Organic Fertilizer on the plot | .0964 | .0960 | .0268 | .0244 |
| | (.0840) | (.0831) | (.0779) | (.0782) |
| Use of Inorganic Fertilizer on the plot | .4341*** | .4002*** | .3532*** | .3333*** |
| | (.0746) | (.0735) | (.0731) | (.0713) |
| Use of Pesticides on the plot | .0854 | .0767 | .0307 | .0381 |
| | (.1083) | (.0950) | (.0898) | (.0849) |
| Improved seeds | .2276*** | .1946*** | .1317** | .1240** |
| | (.0522) | (.0576) | (.0619) | (.0600) |
| Multiple managers on the plot | | -.1243 | -.2076 | -.1715 |
| | | (.1594) | (.1571) | (.1591) |
| Main manager is a female | | .0111 | .0329 | .0409 |
| | | (.0689) | (.0696) | (.0692) |
| Age of the main plot manager | | -.0252** | -.0299** | -.0300** |
| | | (.0123) | (.0130) | (.0124) |
| Age of the main manager (squared) | | .0002 | .0002 | .0002 |
| | | (.0001) | (.0001) | (.0001) |
| Individual years of schooling | | .0282*** | -.0308 | -.0357 |

| | Log of productivity | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | | (.0096) | (.0452) | (.0454) |
| Number of manager | | .1669 | .2477* | .1979 |
| | | (.1312) | (.1293) | (.1291) |
| Household size | | | -.0245*** | -.0232** |
| | | | (.0086) | (.0100) |
| Household head school years | | | .0504 | .0571 |
| | | | (.0424) | (.0428) |
| Having access to electricity | | | .2228 | .2013 |
| | | | (.1908) | (.1920) |
| Asset index (asset dummies) | | | .1814** | .1770** |
| | | | (.0800) | (.0825) |
| HH, Any shock (0/1) | | | -.0500 | -.0319 |
| | | | (.0822) | (.0844) |
| HH, asset losses (0/1) | | | .0021 | .0188 |
| | | | (.0679) | (.0702) |
| Number of plots | | | .0830 | .0920 |
| | | | (.0628) | (.0622) |
| Number of plot (squared) | | | -.0123* | -.0127* |
| | | | (.0071) | (.0071) |
| Agricultural extension index (Source based) | | | .0523 | .0410 |
| | | | (.0502) | (.0497) |
| Farm implements index (diversity) | | | .1744*** | .1616*** |
| | | | (.0281) | (.0280) |
| Slope (gradient of steepness that measured in degree) | | | | -.0176 |
| | | | | (.0228) |
| Length of growing period in days | | | | -.0044*** |
| | | | | (.0017) |
| Demeaned average temperature | | | | -.0607*** |
| | | | | (.0191) |
| Demeaned average rainfall | | | | -.0034 |
| | | | | (.0023) |
| Observations | 1986 | 1974 | 1932 | 1913 |
| $R^2$ | .1196 | .1485 | .1906 | .2054 |
| $R^2$adj | .1051 | .1318 | .1700 | .1832 |

Standard errors in parentheses $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

**Table 5: Decomposition of R-squared for productivity regression analysis**

| Input | Shapley |
|---|---|
| Plot Characteristics | 29.156 |
| Manager Characteristics | 14.994 |
| Household Characteristics | 25.966 |
| Community Characteristics | 12.486 |
| Zones-Year fix-effects | 17.397 |

**Table 6: Estimated Productivity and Standard Error using bootstrap with 100 replications**

| Region | Productivity | Std. Error |
|---|---|---|
| dodoma | 6.716 | 0.701 |
| arusha | 11.024 | 1.511 |
| kilimanjaro | 13.894 | 11.423 |
| tanga | 7.142 | 0.725 |
| morogoro | 7.373 | 1.090 |
| pwani | 4.520 | 1.401 |
| DAR ES SALAAM | 8.780 | 11.284 |
| lindi | 2.973 | 0.205 |
| mtwara | 3.655 | 0.280 |
| ruvuma | 8.175 | 0.653 |
| iringa | 10.948 | 1.062 |
| mbeya | 7.644 | 0.506 |
| singida | 4.246 | 0.218 |
| tabora | 6.832 | 0.298 |
| rukwa | 12.240 | 0.923 |
| kigoma | 5.092 | 1.378 |
| shinyanga | 6.244 | 0.266 |
| kagera | 2.138 | 0.132 |
| mwanza | 4.583 | 0.512 |
| mara | 6.713 | 0.904 |
| manyara | 14.489 | 5.608 |
| njombe | 6.156 | 2.448 |
| katavi | 8.762 | 5.614 |
| simiyu | 7.227 | 0.983 |
| geita | 2.882 | 0.301 |

**Table 7: Rates of Growth of Productivity and Malmquist index of productivity among Regions and Zones in Tanzania**

| | Productivity growth | | | Malmquist Index | | |
|---|---|---|---|---|---|---|
| Unit | 2008-2010 | 2010-2012 | 2012-2014 | 2008-2010 | 2010-2012 | 2012-2014 |
| Regions | | | | | | |
| dodoma | 0.534 | 0.264 | -0.406 | 1.055 | 1.179 | 0.707 |
| arusha | -0.272 | 0.119 | 0.153 | 0.800 | 0.902 | 1.533 |
| kilimanjaro | -0.200 | 1.714 | -0.321 | 0.742 | 1.797 | 1.055 |
| lindi | 0.077 | -0.120 | -0.318 | 0.975 | 1.162 | 0.571 |
| mtwara | -0.469 | 0.254 | -0.382 | 0.574 | 1.284 | 0.439 |
| ruvuma | 0.241 | 0.098 | 0.961 | 1.557 | 1.050 | 1.898 |
| iringa | 0.013 | -0.120 | 0.250 | 0.795 | 0.876 | 1.235 |
| mbeya | 0.587 | -0.153 | 0.911 | 0.965 | 0.955 | 1.831 |
| tabora | -0.211 | 0.570 | -0.313 | 0.715 | 1.488 | 0.897 |
| rukwa | -0.137 | 0.088 | -0.124 | 0.786 | 1.437 | 0.816 |
| kigoma | -0.024 | 0.054 | -0.569 | 1.302 | 0.801 | 0.786 |
| shinyanga | 0.087 | -0.132 | -0.227 | 1.062 | 0.873 | 0.967 |
| mwanza | -0.466 | 0.708 | -0.221 | 0.371 | 1.829 | 0.894 |
| manyara | 0.387 | 0.018 | 0.366 | 1.387 | 1.090 | 1.478 |
| Zones | | | | | | |
| Central Zone | 0.055 | 0.414 | -0.221 | 0.990 | 1.414 | 0.827 |
| Coastal Zone | 0.033 | 0.364 | 0.819 | 1.042 | 1.331 | 1.659 |
| Lake Zone | 0.088 | -0.038 | -0.068 | 1.088 | 0.962 | 0.973 |
| Northern Zone | 0.334 | -0.022 | 0.693 | 1.334 | 0.978 | 1.693 |
| S.Highlands Zone | 0.096 | -0.046 | 0.249 | 0.986 | 1.011 | 1.289 |
| all zones | 0.067 | 0.086 | 0.183 | 1.018 | 1.086 | 1.216 |

**Table 8: Robustness: Factors associated with Productivity.**

| | Log Productivity (Self-reported) | | | Log Productivity (GPS measures) | | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Soil quality is good | .4384*** | .4392*** | .4211*** | .4372*** | .4373*** | .4380*** | .4199*** | .4361*** |
| | (.1140) | (.1242) | (.1170) | (.1155) | (.1142) | (.1242) | (.1171) | (.1156) |
| Soil quality is average | .2053* | .1992* | .1907* | .1908* | .2039* | .1976* | .1891 | .1892* |
| | (.1049) | (.1156) | (.1141) | (.1125) | (.1054) | (.1161) | (.1146) | (.1130) |
| Having erosion problem | .0398 | .0522 | .0405 | .0646 | .0405 | .0529 | .0411 | .0650 |
| | (.0856) | (.0755) | (.0772) | (.0710) | (.0855) | (.0753) | (.0770) | (.0708) |
| Being steep | -.1342 | -.2228 | -.2664* | -.1281 | -.1335 | -.2219 | -.2658* | -.1277 |
| | (.1884) | (.1717) | (.1527) | (.1573) | (.1883) | (.1715) | (.1528) | (.1573) |
| Distance from plot to home (Km) | .0041 | .0023 | -.0005 | .0010 | .0042 | .0024 | -.0004 | .0011 |
| | (.0076) | (.0078) | (.0077) | (.0080) | (.0076) | (.0078) | (.0077) | (.0080) |
| Distance from plot to the road (Km) | -.0016 | .0015 | .0043 | .0078 | -.0019 | .0012 | .0040 | .0076 |
| | (.0089) | (.0098) | (.0096) | (.0091) | (.0089) | (.0098) | (.0095) | (.0091) |
| Distance from plot to the market (Km) | .0056** | .0041 | .0029 | .0019 | .0056** | .0041 | .0029 | .0019 |
| | (.0027) | (.0028) | (.0026) | (.0028) | (.0027) | (.0028) | (.0026) | (.0028) |
| Solely owned by the household | -.2542*** | -.1881** | -.1878** | -.1700** | -.2544*** | -.1884** | -.1880** | -.1702** |
| | (.0872) | (.0903) | (.0877) | (.0835) | (.0870) | (.0902) | (.0875) | (.0833) |
| Plot has a title | -.0106 | -.0206 | -.0780 | -.0624 | -.0141 | -.0244 | -.0818 | -.0661 |
| | (.1258) | (.1229) | (.1235) | (.1268) | (.1252) | (.1224) | (.1231) | (.1263) |
| Use of Organic Fertilizer on the plot | .0728 | .0776 | .0102 | .0316 | .0698 | .0744 | .0072 | .0284 |
| | (.0945) | (.0928) | (.0908) | (.0943) | (.0941) | (.0925) | (.0903) | (.0938) |
| Use of Inorganic Fertilizer on the plot | .4052*** | .3992*** | .3727*** | .3701*** | .4040*** | .3979*** | .3711*** | .3681*** |
| | (.0937) | (.1000) | (.1063) | (.0995) | (.0932) | (.0995) | (.1061) | (.0993) |
| Use of Pesticides on the plot | -.0523 | -.0368 | -.0718 | -.0876 | -.0575 | -.0421 | -.0768 | -.0930 |
| | (.1254) | (.1153) | (.1136) | (.1243) | (.1242) | (.1142) | (.1129) | (.1235) |

| | Log Productivity (Self-reported) | | | | Log Productivity (GPS measures) | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Improved seeds | .2088*** | .1820*** | .1288* | .1263* | .2100*** | .1832*** | .1301* | .1277* |
| | (.0617) | (.0623) | (.0701) | (.0688) | (.0617) | (.0623) | (.0701) | (.0688) |
| Multiple managers on the plot | | .0892 | .0042 | .0335 | | .0916 | .0068 | .0364 |
| | | (.1802) | (.1695) | (.1695) | | (.1802) | (.1696) | (.1694) |
| Main manager is a female | | .0428 | .0521 | .0555 | | .0434 | .0526 | .0559 |
| | | (.0876) | (.0802) | (.0811) | | (.0875) | (.0802) | (.0810) |
| Age of the main plot manager | | -.0210 | -.0267 | -.0266* | | -.0209 | -.0266 | -.0265* |
| | | (.0144) | (.0161) | (.0151) | | (.0144) | (.0161) | (.0151) |
| Age of the main manager (squared) | | .0001 | .0001 | .0001 | | .0001 | .0001 | .0001 |
| | | (.0001) | (.0002) | (.0001) | | (.0001) | (.0002) | (.0001) |
| Individual years of schooling | | .0052 | -.0685 | -.0718 | | .0053 | -.0686 | -.0719 |
| | | (.0126) | (.0504) | (.0513) | | (.0126) | (.0501) | (.0509) |
| Number of manager | | .0524 | .1129 | .0746 | | .0510 | .1115 | .0730 |
| | | (.1522) | (.1512) | (.1433) | | (.1522) | (.1511) | (.1432) |
| Household size | | | -.0175 | -.0163 | | | -.0175 | -.0162 |
| | | | (.0108) | (.0115) | | | (.0108) | (.0115) |
| Household head school years | | | .0675 | .0731 | | | .0677 | .0734 |
| | | | (.0508) | (.0518) | | | (.0504) | (.0515) |
| Having access to electricity | | | .4484** | .4232** | | | .4478** | .4223** |
| | | | (.2167) | (.2093) | | | (.2168) | (.2094) |
| Asset index (asset dummies) | | | .0655 | .0745 | | | .0659 | .0750 |
| | | | (.1034) | (.1030) | | | (.1036) | (.1031) |
| HH, Any shock (0/1) | | | -.0931 | -.0665 | | | -.0926 | -.0658 |
| | | | (.0983) | (.1007) | | | (.0984) | (.1007) |
| HH, asset losses (0/1) | | | .1199* | .1278* | | | .1176 | .1255* |
| | | | (.0708) | (.0720) | | | (.0708) | (.0721) |

| | Log Productivity (Self-reported) | | | | Log Productivity (GPS measures) | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Number of plots | | | .0978 | .0924 | | | .0984* | .0928 |
| | | | (.0590) | (.0586) | | | (.0588) | (.0584) |
| Number of plots (squared) | | | -.0122** | -.0105* | | | -.0122** | -.0105* |
| | | | (.0058) | (.0057) | | | (.0058) | (.0057) |
| Agricultural extension index(Source | | | .0335 | .0086 | | | .0340 | .0093 |
| based) | | | (.0807) | (.0840) | | | (.0810) | (.0843) |
| Farm implements index (diversity) | | | .1542*** | .1441*** | | | .1534*** | .1432*** |
| | | | (.0321) | (.0325) | | | (.0323) | (.0327) |
| Slope(gradient of steepness that mea- | | | | -.0187 | | | | -.0185 |
| sured in degree) | | | | (.0265) | | | | (.0265) |
| Length of the growing period in days | | | | -.0056*** | | | | -.0057*** |
| | | | | (.0016) | | | | (.0016) |
| Demeaned average temperature | | | | -.0432* | | | | -.0438* |
| | | | | (.0230) | | | | (.0230) |
| Demeaned average rainfall | | | | -.0028 | | | | -.0027 |
| | | | | (.0025) | | | | (.0025) |
| Observations | 1307 | 1304 | 1290 | 1287 | 1307 | 1304 | 1290 | 1287 |
| $R^2$ | .1112 | .1442 | .1869 | .2038 | .1112 | .1442 | .1867 | .2037 |
| $R^2$adj | .0889 | .1185 | .1555 | .1702 | .0889 | .1184 | .1553 | .1701 |

Standard errors in parentheses * p < 0.10, ** p < 0.05, *** p < 0.01

Table 9: Decomposition of R-squared for productivity regression models: Self-reported (sr) vs GPS

| Input | Shapley-sr | Shapley-gps |
|---|---|---|
| Plot Characteristics | 25.706 | 25.732 |
| Manager Characteristics | 19.167 | 19.174 |
| Household Characteristics | 22.280 | 22.183 |
| Community Characteristics | 12.759 | 12.814 |
| Zones-Year fix-effects | 20.088 | 20.096 |

# Figures

Figure 1: Example of a 3D Zonotope, source Hildenbrand (1981). This zonotope has been generated by 4 vectors $a_1$, $a_2$, $a_3$, and $a_4$ representing farms (or farms' characteristics). The combination (Mikowski sum of the segment line) of two vectors yields a parallelogram, and all the vectors combined, give the zonotope.

Source Hildenbrand (1981).

**Figure 1: Example of a 3D Zonotope**

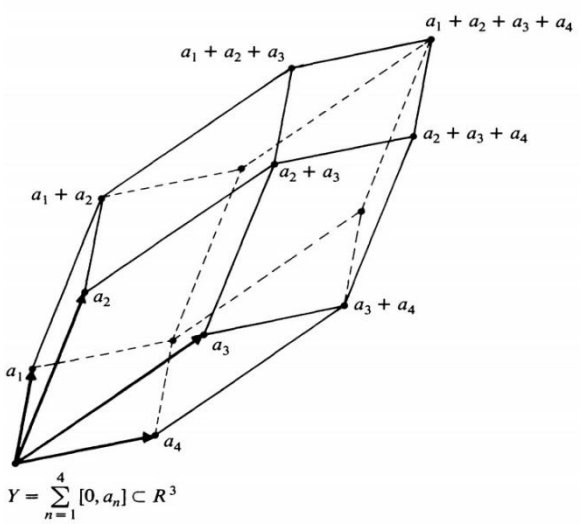


Figure 2: A graphical 3D illustration of the geometric approach. For simplicity, we only plot one vector of plot characteristics ($a_i$) and illustrates the measures obtained from the geometric analysis. $d_y$ is the diagonal of the zonotope from projecting all the vectors in the inputs-output space ($L_a, L_d, Y$). $pr_{-3}(d_y)$ and $pr_{-3}(a_i)$ represent respectively the projection of $d_y$ and $a_i$ in the inputs-space. The tangent of $\theta_3(d_y)$ is the productivity index of the whole sector made up of the farms $a_i$.

**Figure 2: A graphical 3D illustration of the geometric approach.**

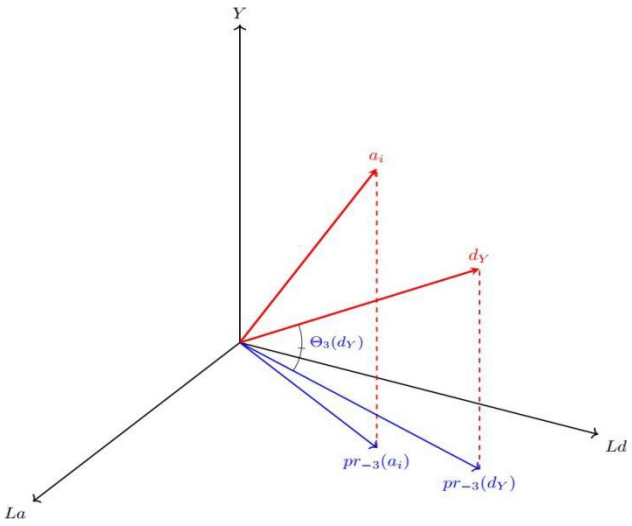
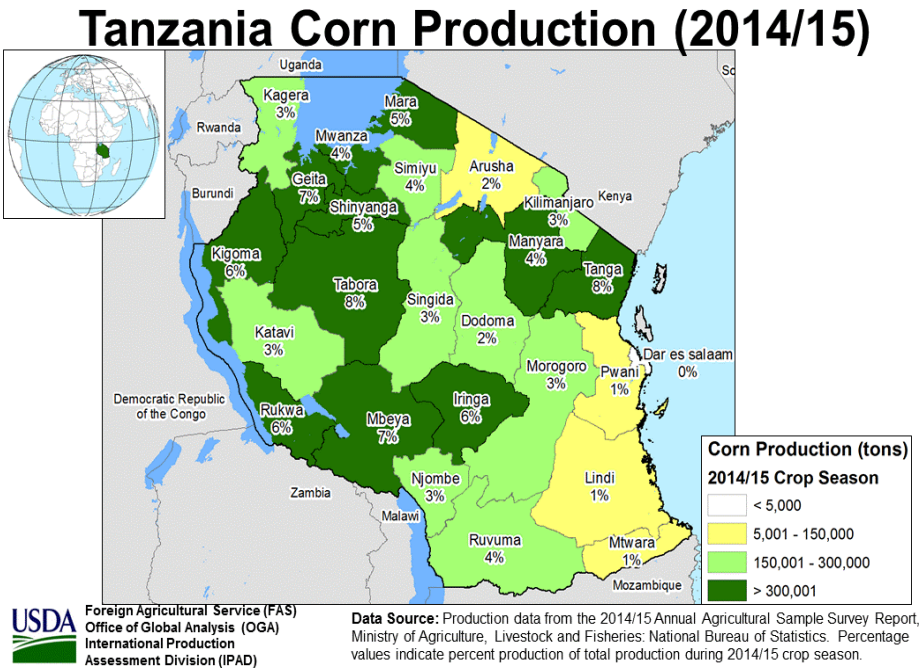



Figure 3: Zonotopes of Pwani (left) and Kigoma (right) - rendering at the same scale. The shape of the zonotopes provides graphically an idea of how heterogeneous is the maize sector in each of the regions.

**Figure 3: Zonotopes of Pwani (left) and Kigoma (right)**

**Figure 4: Map showing the spatial distribution of Maize crop in Tanzania in 2014/15.**




Source: https://ipad.fas.usda.gov/rssiws/al/eafrica_cropprod.aspx

**Figure 5. Distribution of self-reported and GPS plot sizes**

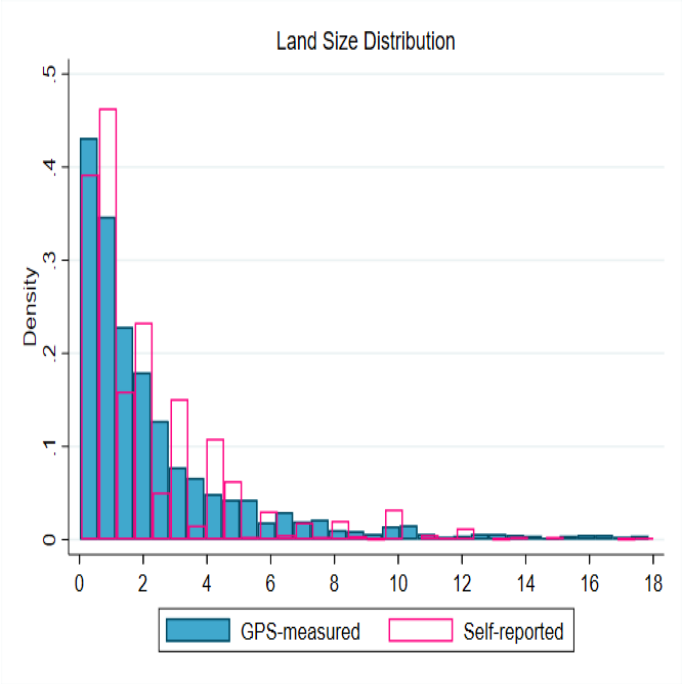



**Figure 6: Comparing empirical densities of log productivity index (GPS- Measured and self-reported)**

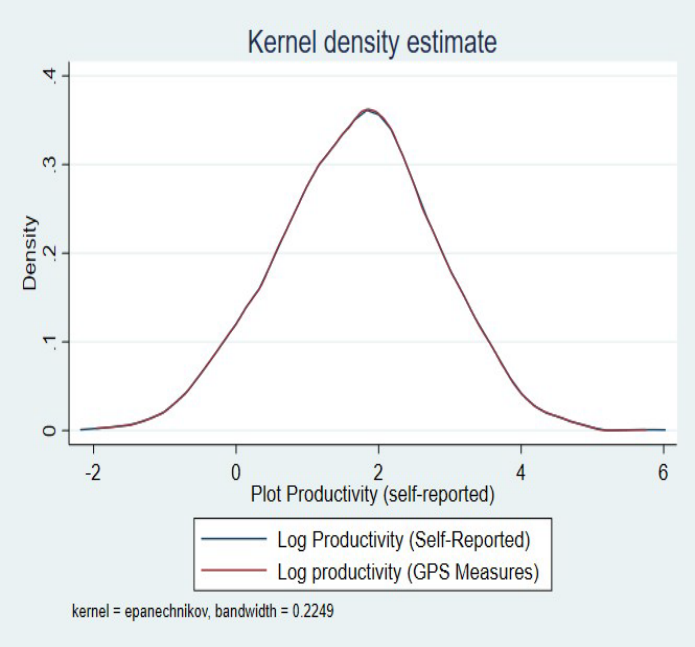


**Figure 7. Empirical Distribution of Labor Productivity in Maize Production by Agroecological in Tanzania**

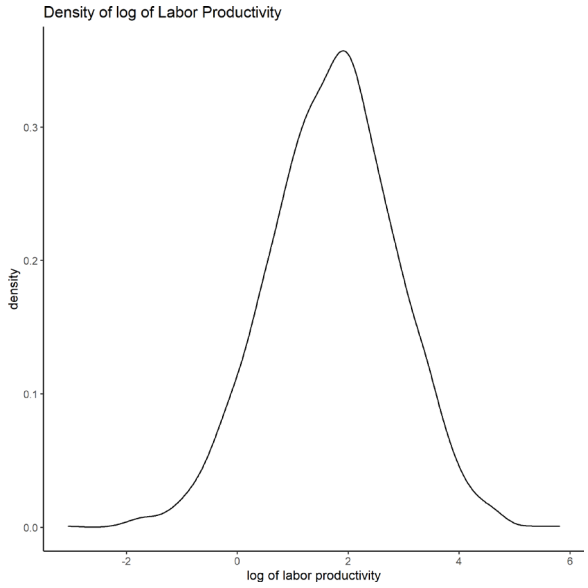



**Figure 8: Empirical Distribution of Labor Productivity in Maize Production by Agroecological per zone and round in Tanzania**

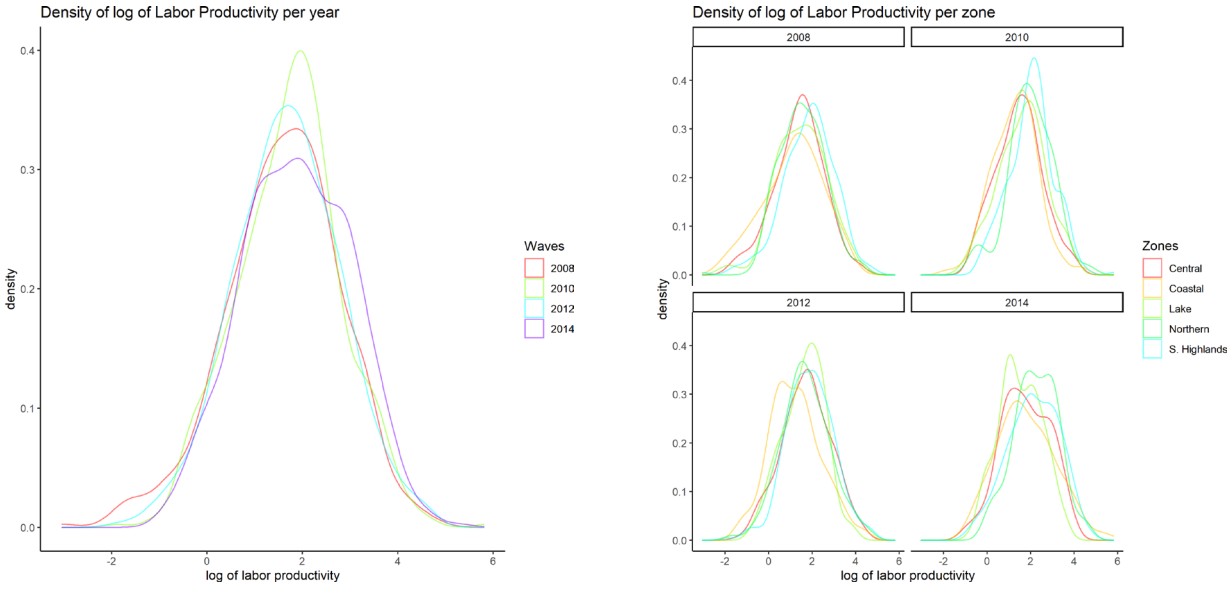


**Figure 9: Comparing the empirical densities of the Zonotope (log productivity) and the DEA (log efficiency) approaches per round**

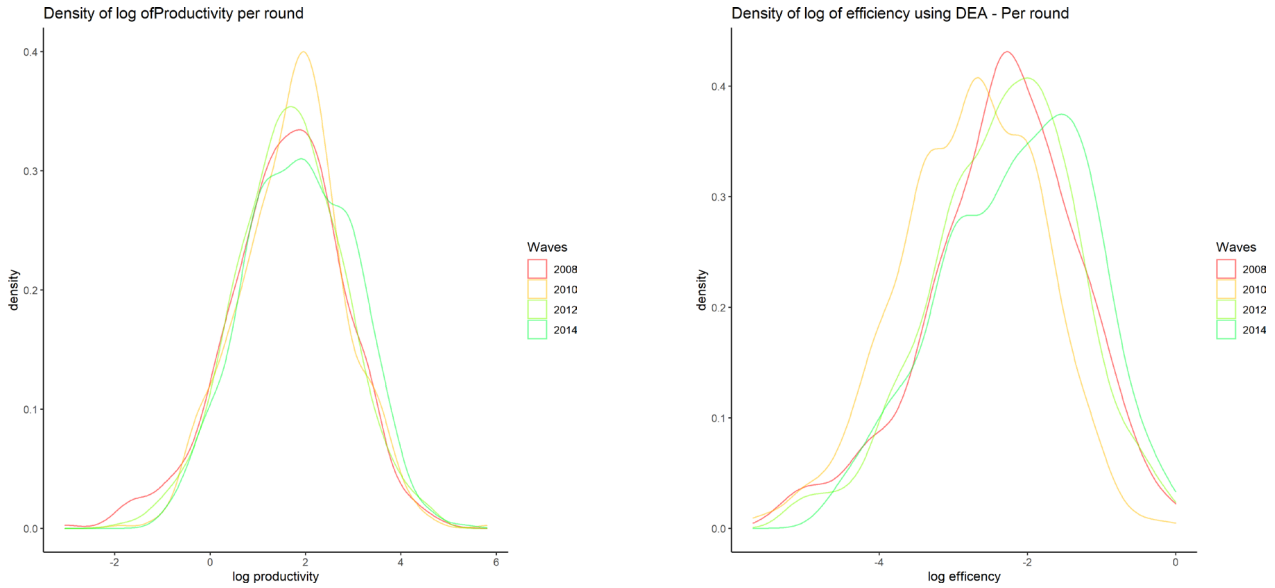
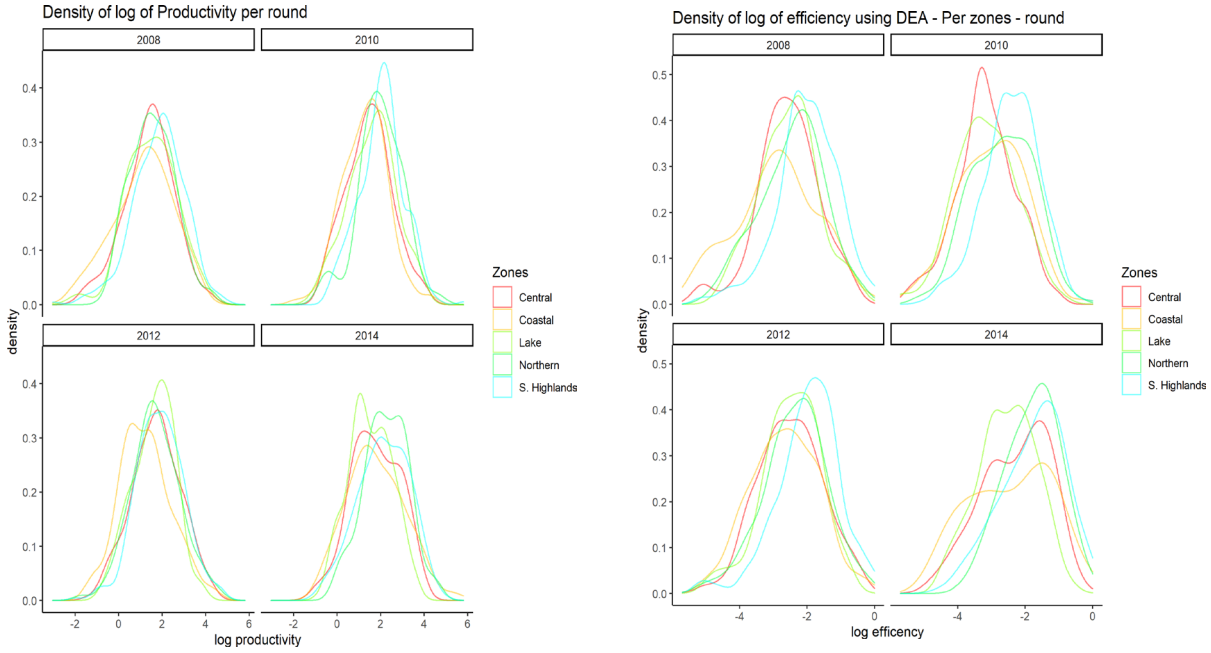



**Figure 10: Comparing the empirical densities of the Zonotope (log productivity) and the DEA (log efficiency) approaches per zone and round.**



# References


Abay, Kibrom A, Gashaw T Abate, Christopher B Barrett, and Tanguy Bernard (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, Kibrom A, Leah EM Bevis, and Christopher B Barrett (2021), "Measurement error mechanisms matter: Agricultural intensification with farmer misperceptions and misreporting." American Journal of Agricultural Economics, 103, 498–522.

AfDB (2013), "At the center of Africa's transformation, strategy for 2013–2022." African Development Bank.

Battese, George Edward and Tim J Coelli (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data." *Empirical economics*, 20, 325–332.

Dillon, Andrew, Sydney Gourlay, Kevin McGee, and Gbemisola Oseni (2019), "Land measurement bias and its empirical implications: evidence from a validation exercise." *Economic Development and Cultural Change*, 67, 595–624.

Dosi, Giovanni, Marco Grazzi, Luigi Marengo, and Simona Settepanella (2016), "Production theory: accounting for firm heterogeneity and technical change." *The Journal of Industrial Economics*, 64, 875–907.

Farrell, Michael James (1957), "The measurement of productive efficiency." *Journal of the Royal Statistical Society: Series A (General)*, 120, 253–281.

Giller, Ken E, Pablo Tittonell, Mariana C Rufino, Mark T Van Wijk, Shamie Zingore, Paul Mapfumo, Samuel Adjei-Nsiah, M Herrero, Régis Chikowo, Marc Corbeels, et al. (2011), "Communicating complexity: integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development." *Agricultural systems*, 104, 191–203.

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

Hildenbrand, Werner (1981), "Short-run production functions based on microdata." *Econometrica: Journal of the Econometric Society*, 1095–1125.

Koopmans, Tjalling C (1977), "Examples of production relations based on microdata." In *The Microeconomic foundations of macroeconomics*, 144–178, Springer.

Löthgren, Mickael (1997), "Generalized stochastic frontier production models." *Economics Letters*, 57, 255–259.Maue, Casey C, Marshall Burke, and Kyle J Emerick (2020), "Productivity dispersion and persistence among the world's most numerous firms."

National Bureau of Statistics, United Republic of Tanzania (2016), "Basic information documentnational panel survey (nps 2014-2015)."

Pratt, Alejandro Nin (2015), "The challenge of increasing agricultural productivity in Africa south of the Sahara." Accessed 22 January 2020,https://www.ifpri.org/blog/ challenge-increasing-agricultural-productivity-africa-south-sahara.

Simar, Léopold and Valentin Zelenyuk (2011), "Stochastic fdh/dea estimators for frontier analysis." Journal of Productivity Analysis, 36, 1–20.

Van Dijk, Michiel, Tom Morley, Roel Jongeneel, Martin van Ittersum, Pytrik Reidsma, and Ruerd Ruben (2017), "Disentangling agronomic and economic yield gaps: An integrated framework and application." Agricultural Systems, 154, 90–99.

Vanlauwe, Bernard, RIC Coe, and Ken E Giller (2019), "Beyond averages: new approaches to understand heterogeneity and risk of technology success or failure in smallholder farming." Exl Agriculture, 55, 84–106.