Contract Farming and Agricultural Dynamics in Western Kenya *

Lorenzo Casaburi† Michael Kremer‡ Sendhil Mullainathan§

Preliminary and Incomplete. Do not cite or circulate without permission.

Abstract

We use a newly collected plot level panel database compiled from the administrative records of a large Kenyan contract farming scheme to study productivity and land allocation patterns between 1987 and 2006. We find evidence that a negative rainfall shock reduces the amount of land allocated to cane in the subsequent replant cycle by 4 to 5%. However, smaller and more productive plots are less affected. Our estimates suggest that plots in the bottom 25th percentile of the plot size distribution and in the top 10th percentile of the productivity distribution expand following a shock. Combined with strong evidence of decreasing returns to land, these result suggests a potential "cleansing effect of recessions".

We then compare the observed land allocation across producers with the one that maximizes expected total output. We find that the optimization implies both a redistribution of land toward small producers and an increase in variance of plot sizes. The misallocation implies an average loss of about 5%-15% of total yearly production. Finally, we use a simulation to assess the dynamic impact of rainfall shocks. Shocks have a static negative effect on yields but a potentially positive dynamic effect via reallocation toward more productive and smaller producers. We show that a one-time aggregate rainfall shock that hits all the plots reduces contemporary output by 15% but raises output in the subsequent period by 2%.

*We thank Andrew Foster and seminar audiences at Harvard University for helpful comments. Helene Ba, Sayon Deb, Owen Ozier, Ravindra Ramrattan, and Ian Tomb provided excellent research assistance. We wish to thank the National Bureau of Economic Research Africa Project and the Sustainability Science Program at the Harvard University for funding this research and Innovations for Poverty Action Kenya for administrative support. The usual disclaimer applies.

†Harvard University, email: casaburi@fas.harvard.edu

‡Harvard University, email: mkremer@fas.harvard.edu

§Harvard University, email: mullain@fas.harvard.edu

1 Introduction

Increasing smallholder agricultural productivity is a key prerequisite for promoting growth in developing countries. This is even more relevant for Sub-Saharan Africa which, with some noticeable exceptions, has experienced an increase in the gaps in agricultural output per capita and per hectare relative to other regions (World Bank, 2008). Recent contributions have provided insights on the evolution of both input productivity and TFP at aggregate level (Block, 2011). At the micro level, the determinants of productivity of individual farms are at the core of a large body of literature. In this study, we focus instead on the role of land allocation as a source of aggregate productivity. We look at the evolution of production shares of high versus low productive farmers, and at the role of exogenous shocks, such as negative rainfall events, in shaping these patterns.

Understanding whether reallocation toward high productive users occurs in the agricultural sector is an important question for research and for policy. On the one hand, the literature has emphasized several factors that might break the relation between expected productivity and total production shares, such as liquidity constraints (Conning and Udry, 2007), "power" factors (Goldstein and Udry, 2008) and behavioral biases (Duflo, Kremer and Robinson, 2011). On the other hand, a wide range of studies points at the inefficiency deriving from the coexistence of high and low productive producers and at the potential gains from reallocation (Foster and Rosenzweig 2011 for India; Collier and Dercon 2009 for Sub-Saharan Africa). Finally, improving efficiency land of allocation is one of the goals of a wide range of policies, including land reforms and financial sector interventions.

A formal analysis of the efficiency and the dynamics of the allocation patterns requires detailed panel micro data for large samples of producers in an industry, and for a time series that goes beyond the low-T panels collected in standard household and agricultural survey projects. While the increased availability of long-term plant level panels has fostered research on these topics in the analysis of the manufacturing industries (Syverson, 2010; Hsieh and Klenow, 2009), data limitations are likely to be a major obstacle to progress in research in this direction when looking at the agricultural sector in developing countries.

Our study provides a contribution toward filling this gap, using a new plot-level database collected from the administrative records of a large sugar cane contract farming scheme, involving tens of thousands of farmers in Western Kenya. Sugar cane outgrowers schemes in Western Kenya are generally considered to be one of the most successful examples of contract farming in East Africa and a well known case study in the analysis of the shift toward commercial crops in areas previously focused primarily on food crops (Grosh, 1994). Our database covers sugar cane harvests between 1987 and 2006 for more

than 15,000 "accounts". This is a number close to the total population of sugar cane contracted farmers ("outgrowers") active at any time during the observation period in the target areas. While the compiled panel represents an important advancement for the study of long term dynamics, it has some important limitations that we discuss in the paper, such as the lack of labor data.

In the first part of the paper, we study how land allocation to sugarcane depends on transitory income shock and on differences in expected land productivity, either via heterogenous marginal returns to land or via heterogenous residual land productivity. We also test whether different producers react differently to an exogenous negative income shock. As a starting point, it is helpful to think of a basic production function $Y_t = T_t A_t^{\theta} X_t^{\gamma}$, where Y is cane production, A is the amount of land under cane cultivation, X is a vector of exogenous covariates, such as rainfall shocks, and T is the residual land productivity. Our preliminary step consists in the estimation of θ and γ, as well as of the vector T. We run several alternative specifications in order to take into account the potential endogeneity of A. In all the specifications, we find strong evidence of decreasing returns to land, with our preferred estimates of θ being between 0.572 and 0.614 . The presence of decreasing returns is confirmed when we use profits per hectare (excluding labor costs) as dependent variable rather than output. The result contributes to the the broad and controversial literature on the inverse relationship between plot size and productivity/profitability (Binswanger, Deninger, and Feder, 1995; Eastwood, Lipton, and Newell, 2010).

We then model the allocation of land to cane in the subsequent replant cycle as a linear function of current rainfall shock, current plot size, and of our measure of expected land productivity estimated above, as well as time and geographical dummies. Several robust patterns arise from all the alternative specifications. The following figures refer to our preferred specification. First, accounts experiencing a negative rainfall shock in the current cycle shrink by 4 to 5 additional log points in the subsequent replant. The result points to a strong persistence between current shocks and investment decisions, a result consistent with the presence of imperfect capital markets and liquidity constraints. Assuming that the only channel through which rainfall shocks operate is their impact on income - an assumption that we qualify later in the paper - this implies approximately that a 10% negative income shock in the current cycle has a 2.5-3% negative effect on subsequent cycle plot size. We thus highlight a strong relation between weather variability and investment decisions, in the line of work pionereed by Rosenzweig and Binswanger (1993). Second, plots with larger sizes shrink relatively more: an increase in 10 log points in current plot size increases log plot size in the subsequent replant cycle by 7.5 log points. In the presence of decreasing returns to land, this contributes to a relative

increase in aggregate yields. Third, an increase in one standard deviation in our measure of the systematic component of unobserved land productivity increases plot size in the subsequent replant cycle by 5.6 log points, controlling for actual land size. Fourth, higher productive and smaller accounts shrink relatively less following a negative rainfall shocks. In addition, plots in the top 10th percentile of the residual land productivity and in the bottom 25th percentile of plot size distribution expand *more* following a negative rainfall shock than in normal circumstances. These results provide suggestive evidence of a "cleansing effect of recessions" (Caballero, 1994) in the sector. This result also shows that rainfall shocks have a negative static and possibly a positive dynamic effect, an issue we analyze in the second part of the paper.

In the second part of the paper, we use estimates from the above regressions in order to provide further insights on the efficiency of land allocation and the dynamics of aggregate productivity. First, we compare the properties of the actual land distribution with those of the "optimal" distribution. The latter is computed by solving the problem of a social planner who allocates land across the different accounts in each location so to maximize total production, given the total amount of land available. The algorithm uses the above estimates of the returns to land and of residual land productivity. Throughout the sample period, we find that the optimal land allocation implies, on average, a redistribution of land from large producers to some more productive small producers. However, the optimal distribution also presents a higher degree of inequality (Gini coefficient) as it exhibits high variance in land allocation among accounts starting with the same size. We quantify the average loss from this misallocation to be between 5% and 15%.

Finally, we delve into the role of rainfall shocks. As discussed above, our measure for negative rainfall shocks has a significant negative impact on production levels. On the other hand, it drives relative growth in the size of accounts with smaller plot size and higher residual land productivity. We attempt to quantify the compounded long run effect using a simulation. We perform a simple thought experiment and introduce an aggregate shock where every plot is hit by a poor rain regime in a given period. After drawing plot sizes and land productivity from their distribution in initial years of our sample and calibrating the model using estimates from the first part of the paper, we simulate the economy for three periods before the aggregate shock and 6 periods after, for a total of 10 periods. We find that, relative to the counterfactual with no aggregate shock, the aggregate shock scenario exhibits a fall of 15% in production levels in the period of the shock. Yet, production levels in the periods that follow the aggregate shock are above the counterfactual by about 2%. This preliminary result suggests that exogenous income shocks can be an important determinant of reallocation.

The remainder of the paper proceeds as it follows. In Section 2, we provide a back-

ground of the contract farming scheme we focus on. Section 3 provides an overview of the database and presents summary statistics. In Section 4 we estimate land productivity and we study the determinants of land usage decisions, highlighting the role of productivity, plot size, rainfall shocks and their interaction. Section 5 compares the properties of the actual land distribution with the optimal one and uses a simulation to quantify the impact of rainfall shocks on production and yields. Section 6 discusses potential policy implications of our preliminary findings and describes steps we plan to undertake in future work.

2 Background: Sugar Cane Contract Farming in Western Kenya

Contract farming is widespread in Africa and concerns primarily rural smallholders (Grosh, 1994). The outcomes of such schemes constitute an important and relatively understudied research question for economists. In Western Kenya, following the establishment of several sugar factories and outgrowers schemes in the Sixties and the Seventies, sugar cane has become the most common cash crop. In spite of important caveats, the establishment and expansion of these cane contract farming schemes is generally considered a major success story in the transition toward commercial agriculture in East Africa. The shift from food to market oriented agriculture is generally considered an important driver of structural transformation and growth, and the establishment of contract farming schemes in the area represented a major turning point in the regional economy, leading to increased incomes, services, input usage and non-agricultural employment (in the factory). Milling capacity expanded rapidly post-independence in response to targeted government investments with the construction of a total of six factories between 1965 and 1985. The sector undertook substantial reforms in the Nineties, with privatization of the government owned mills (with the government often retaining majority shares). Smallholder outgrowers contribute to about 92% of total sugar cane production, with the rest grown in the nucleus estates of the companies. In the Kenyan context, this has led to an expansion total sugar production from 369,000t in 1984 to 520,000t in 2008 as smallholder farmers have diversified away from food crops (Kenya Sugar Board, 2011).

In 1996, the company outgrowing scheme spanned across sixty-six sublocations, the Kenyan smallest administrative unit. The scheme included tens of thousand of smallholder farmers. The administrative unit used by the company, and thus the unit used for our analysis, is the "account". At any point of time an account is held by one or more contracting farmers. The name of the farmer(s) contracted on each account does vary over

time. However, our database is at account level, not farmer level. Another way to look at this is that we are in fact assuming that each account is contracted by a different household. Based on our fieldwork, we believe this to be a sensible approximation, although we certainly observe the presence of several large farmers who contract on multiple accounts. In addition, the contracted farmer is not necessarily the owner of the land. Land rental markets are quite developed in the area. Following a formal rental agreement, the tenant replaces the landlord on the contract. Even in the absence of this type of transaction, we observe frequent changes in the names of the contracted farmers. These patterns are likely to reflect transmission of plot management across members of the same households or inheritance episodes.

Each account is typically matched to one (sub)parcel as defined by the Kenyan land registry. Different accounts can share the same parcel in cases where a parcel gets split into two parts, for instance between two brothers or between a landlord and a tenant. Potential changes in managers and land sizes of the account make it a unit of observation we believe to be roughly comparable to the plant in standard manufacturing data. Accounts are aggregated into *fields*, sets of plots that are usually treated homogeneously for input provision, in order to exploit economies of scale. The average number of plots per field in our data is 13.36 (s.d. 6.98).

The company and the contracting farmer write a contract that typically spans for one replant cycle. Each of these include the initial planting harvest and several *ratoon* harvests, usually two or three. Ratooning leaves the lower parts of the plant along with the root uncut at the time of harvesting. The main benefit of ratooning is that the crop matures earlier. The yield of the ratoon crop decreases after each cycle. Formally, the company takes the choice to enter another ratoon cycle vs. replanting at field level, based on yields from the last harvest agronomic analysis. However, farmers' opinion, which can certainly differ from the company's best interests, can have an important role in shaping the final decision. The duration of each harvest cycle spans between 18 and 27 months, though early or late harvesting can occur following specific raw material demand from the mill. Planting and harvesting occur in staggered fashion throughout most of the year, in order to provide a constant supply of cane to the processing mill. The length of the harvest cycle is a major difference from the other major crop in the area, maize, which is harvested twice a year. The difference in harvest durations is an important factor in shaping the farmer's decision to allocate land to one of the two crops.

Farmers are paid based on the tonnage of cane provided at harvest time. Cane prices are homogeneous for all the farmers that harvest at the same time. Price changes are typically announced a few weeks before their implementation. Timing of the changes are plausibly orthogonal to the characteristics of the farmers who are approaching harvest in

that specific period. The relevant price for a given farmer is the one set at the harvest time, not the one holding at the beginning of the cycle. The company provides inputs on credit. These include land preparation (ploughing, harrowing) in the replant cycles, fertilizer (DAP and UREA), harvesting and transport to the mill. The unit cost of transport per ton of cane varies according to discrete transport zones. The farmer's main duties include weeding (several times along the harvest cycle) and fertilizer application, both of which are important determinants of the final yield level.

3 Database Description

3.1 Sample

The target population for the database included of all the accounts that had processed at least one payment between 1997 and 2006 in sixteen target sublocations. Administrative paper records had to be located in the company register, scanned, and entered. The register is divided in a main section that includes folders for active accounts and a storage space for accounts that stopped contracting with the company. Among the target accounts, 93.7% were located. About two-thirds of the attrition comes from two sublocations. In future work, we will assess robustness of our results after excluding these two sublocations. In addition, while searching into the deactivated section of the register, additional accounts operating in the target sublocations between 1987 and 1996, but no longer active by 1997, and thus excluded by the initial target population, were located.

The final sample is comprised of 14,516 accounts, close to a full census of the population of accounts in the target sublocations. The database is based on the forms the company record at each harvest. It includes information spanning between 1987 and 2006. The staggered fashion in which harvesting occurs implies that we have a continuous flow of observations across months and years. In this regard, the structure of our database is somewhat unconventional. We do not observe all of our sample within the same time "period" (e.g. month or year). For the purpose of our analysis, we thus define lags and forwards for each account as referring to the previous or subsequent harvest of the account, rather than to the previous year. The data include information on plot sizes at each harvest. These can change from harvest to harvest, due to the outcome of the maize vs cane allocation choice or the fragmentation of the plot across different household members. A larger fraction of plot changes occurs at replanting time, though we do observe changes in sizes during ratooning, too.

While the census nature and long time series component make the database unique for the goal of studying agricultural dynamics in developing countries, it is important to stress

several limitations we face, compared with standard plot survey data. First, we have little information about household characteristics of the contracting farmer. Second, we lack labor data. Third, we do not have soil quality measures. Fourth, by looking at harvest and input provision dates, we estimate that in about 5% of the cases a certain harvest is missing from our database. This can occur if the form is missing from the account folder in the registry or if its quality makes it unfit for data entry (image deteriorated, blurred printing, waning ink). The likely random nature of these attrition patterns limits the concerns arising from it. Fifth, our field work suggests that in a non trivial share of cases the farmer(s) appearing on the contract develop informal leasing arrangements with third parties who manage the plot. This type of arrangement, unlike the formal leasing cases described above, is not recorded in the forms.

3.2 Summary Statistics

Our initial sample size includes 14,516 accounts and 114,072 account-harvest observations. We implement a two-step trimming procedure. First, we drop observations that are not matched to rainfall stations. Second, we replace extreme outliers in plot size, production, inputs, cycle and yields as missing values. Details are provided in the Appendix **?? **APPENDIX TO BE WRITTEN**. As a result of this procedure, we have a sample of 14,460 accounts and 112,324 harvests which have non-missing values for the variables used in our analysis. Figure 1 shows the number of harvest observations per harvest year.

Each account has an average of 7.87 (s.d 3.29) harvests and 2.76 (s.d. 1.22) replants cycle observations. Figure 2 graphs the patterns of entry and exit. Entry of new accounts occurs throughout the sample period, while the last observed harvest per account is concentrated toward in the last couple of years of our sample, suggesting that the accounts are not leaving the scheme. The average duration of a contract is 2.56 cycles (s.d 1.13).

Table 1 summarizes the variables used in the analysis. For each harvest we observe production level (in tons and revenues), plot size, value of fertilizer and other inputs, an indicator for plant vs. ratoon cycle.

We construct a measure of duration in months of the contract. For the plant cycles, the start month is defined as the month when the seedcane is provided. For the ratoon cycles, the start month is mostly defined as the month of the harvest of the previous cycle. However, in those cases where we identify a missing end-of-harvest statement or in which the first observation for an account is a ratoon (and thus we do not observe seedcane provision date), we impute duration using median duration by planting cycle, sublocation and harvest year (10,001 cases). The average month duration is 22 (s.d 5.17) for plant cycle and 19.75 (s.d. 4.76) for ratoon cycle.

Finally, we obtained monthly rainfall data from forty-four rainfall stations in the outgrowing scheme. As the locations are concentrated in a relatively narrowly defined geographical area, a large portion of the variation (69.53%) in monthly rainfall data in our sample comes from the time series component of the rainfall station level panel, rather than from the cross-sectional dimension. We obtained georeferenced data for the rainfall stations and for a subset of the target fields. We then match fields for which we do not have georeferenced data using handwritten maps used by the company field extension staff. For each field we thus define the two closest rainfall stations. Twenty of the forty-four rainfall station appear in our matched database. Finally we merge the fields with the rainfall data according to the cycle starting month we computed as described above. For each field we thus compute the average rain in the first ten months in the two closest rainfall stations. In addition, we compute binary dummy variables identifying shocks. In particular, throughout the paper, we define a "rainfall shock" in the first or second closest rainfall station if total rain in the first ten months is in the 10th percentile for that station, using the whole time series of data at the station for the 1987-2006 period. For our main regression we use a binary dummy indicating whether the shock occurred in both stations as a simple measure of rainfall shock. This joint event occurs in 5.3% of the harvests and in 6.9% of the cases in the plant cycle harvests. An important point for the remainder of our analysis is that rainfall measures, both levels and binary indicators, are not autocorrelated. A shock in the current harvest does not predict the occurrence of a shock in the subsequent harvest.

Figures 3 and 4 show basic patterns in yields (total production over total cultivated area) and plot sizes. First, we observe a gradual reduction in the average account plot size. Second, yields present a clear negative trend, with the exception of 1994-1997. The trend is consistent with data reported in sectoral publications (Kenya Sugar Board, 2005), although the levels in our database are 10-15% lower than the aggregate levels reported there. Declining soil fertility and continuous sugarcane monoculture are often reported as primary causes of these trends. We lack soil quality data and thus we cannot reach any conclusion of the role of land fertility in shaping these patterns. Changes in the values of observable variables do not capture these changes. Thus, in our model, these patterns are primarily captured by harvest year dummies, which exhibit a clear negative trend.

4 Production and Land Allocation Choices

Changes in aggregate productivity can arise from changes in individual farm productivity ("within" changes) or in the relative expansion of more productive firms ("between" changes). In this section, we focus on the second question, looking at whether (expected)

productivity is a driver of plot size expansion and exit. Differences in expected productivity arise both from different plots sizes, and thus different marginal returns to land, and by different residual land productivity. We also aim to compare the effect of expected productivity in shaping these "investment" decisions with the effect of exogenous factors, such as negative rainfall events, that are likely to affect decisions primarily through a one time income shock channel. Finally we are going to study whether there is an interaction between exogenous negative shocks and relative expansion of the best producers. We thus study whether accounts with different plot sizes or land productivity react differently to adverse rainfall events in terms of land allocation decisions. The analysis we propose to implement requires a clear definition of the measures of productivity and relevant exogenous shocks. We focus on this task in Section 4.1.

4.1 Measuring Land Productivity and Rainfall Shocks

Modelling production is the obvious first step toward the estimation of the degree of returns to land and the residual ("unobserved") land productivity. We propose the following specification:

$$Y_{ilrht} = T_{ilrht} A^{\theta}_{ilrht} g(R_{lrht}), \tag{1}$$

where Y is the volume of cane production, A is the size of the plot, R is a measure of rain quantity for the harvest cycle that enters production through the function $g(\cdot)$ and T is the "residual land productivity", which we describe in detail below. At this stage, it is useful to introduce the subscript notation used in the rest of the paper: i is the account identifier, l is the location, r is the replant number of the account, h is the harvest number within a certain replant cycle and t is the harvest year. The notation is somewhat non-standard and it is driven by the nature of the data, specifically by the fact that we observe accounts whenever they harvest and we define forward and lags based on these occurrences. We use an example to clarify the notation. If a certain account i in location l in replanting cycle r completes the plant harvest in 2000 and the first ratoon harvest in 2003, the two observations will be denoted, respectively, as $y_{ilr,1,2000}$ and $y_{ilr,2,2003}$, and the second will be considered the forward of the first in the analysis. The forward of (r, h) is $(r, h+1)$ in the case in which the subsequent harvest is in the same replant cycle (ratoon harvest) and $(r+1, 1)$ in the case in which it is the first harvest of the subsequent replant cycle. We use the notation $(r, h+1)$ to refer to the general forward. Similarly, $(r, 1)$ will define the first harvest (plant harvest) of the current replant cycle, while $(r+1, 1)$ will define the first harvest in the subsequent replant cycle. In the remainder of the paper, we will omit some of the subscripts if that does not preclude clarity of the presentation. We will

use "," to separate subscript terms when necessary.

As we mentioned above, the lack of household input (labor) data, soil quality information and household characteristics implies we cannot work with a full-fledged multi-input production function. It is thus important to clearly state what the terms in Equation (1) capture. The key insight is that the effect of unobserved inputs on production will enter the parameter θ to the extent to which the inputs are correlated with plot size, and the residual T for the rest (we assume R to be uncorrelated with any observable). We cannot interpret the parameter in the same way we would do if we were estimating a "long" production function. Rather, θ captures the "gross' returns to land after taking into account the correlation with of all the other inputs. Similarly, T, the residual land productivity, is determined by multiple factors, arising from household or land characteristics. T thus measures per-hectare productivity after netting out the marginal returns to land of the account harvest, as determined by the plot size in the harvest, and the effect of other exogenous shocks.

Next, we seek to identify a measure of exogenous rainfall shocks that has a sizable and significant impact on production volumes. We took a first pass at the identification of such shocks by examining the relation between output and our rainfall databases. Using the fact that rainfall is uncorrelated with chosen inputs (as these are chosen before rain regimes are realized) we ignore plot size and heterogenous productivity and run a semiparametric partial linear regression of log output on average rain in the first ten months of the harvest cycle in the rainfall station closest to the field and harvest year fixed effect:

$$y_{ilrht} = g(R^1_{lrht}) + \zeta_t + \epsilon_{ilrht} \tag{2}$$

Figure 5 highlights the heterogeneity in slope in plant vs. ratoon harvest. To include this finding in a parametric setting, in the rest of the paper, we use a simple binary rainfall shock indicator which takes value one if mean rain in the first ten months of the cycle is in the 10th percentile in the 1987-2006 distribution in both the two closest rainfall stations and we allow for an interaction with the plant harvest indicator. As shown in Table 1 this event occurs in 5.3% of the cases for the whole sample and in 6.9% of the cases in the planting cycles. We assume the indicator to enter linearly in the logged production function estimation. The observable inputs that linearly affect log output are thus log plot size, a plant cycle indicator, the binary rainfall shock, the interaction between rainfall shock and planting cycle, geographical and harvest year fixed effects.

Secondly, we attempt to estimate the residual land productivity and, furthermore, to decompose it in a *systematic* (permanent or autocorrelated across harvests) and an *idiosyncratic* (transitory) component, where the first is observed by the producer, but not

to the econometrician, at the time of input decisions, while the second is not. To clarify, we rewrite Equation (1) in logged (lower case) terms, expanding the X term:

$$\begin{aligned} y_{ilrht} =& \gamma_1 \cdot rainshock_{lrht} + \gamma_2 \cdot plantcycle_{ilrht} + \gamma_3 \cdot rainshock_{lrht} * plantcycle_{ilrht} \\ &+ \theta \cdot a_{ilrht} + \underbrace{\zeta_t + \eta_{ilrht} + \epsilon_{ilrht}}_{=\ln T_{ilrht}}, \end{aligned} \tag{3}$$

where $\ln T_{ilrht}$ has been decomposed in a vector of harvest year dummies, ζ_t, permanent or autocorrelated residual land productivity component, η_{ilrht} and a noise term, ϵ_{ilrht}.

We are interested in the estimates of θ, γ and η. We follow alternative estimation strategies which pose different assumptions on the nature of η. The estimation of the parameters in a simple OLS procedure is biased by the standard simultaneity concerns due to the fact that input decisions depend on the expected productivity the producer has at the time of input demand (Mundlak, 2001; Griliches and Mairesse 1998; Ackerberg, Bernard, Berry and Pakes, 2007). These concerns are of particular importance for us as we aim to estimate a measure of unobserved productivity to use in other sections of the paper. We thus implement several alternative estimators. First, we propose a standard *account* fixed effect estimation, which assumes that the non-idyonsincratic component of productivity stays constant across specifications: $\eta_{ilrht} = \eta_i$. The long time dimension of our database and the frequent changes in the names of contracted farmers within the same account make this assumption less realistic than in the standard high-N, low-T panel databases. Thus, as a second attempt to mitigate the endogeneity concern, we estimate a model with *account*replant* fixed effect, where the fixed effect takes a different value every time the same account enters a new replant cycle: $\eta_{lirht} = \eta_{ih}$. Third, we follow Blundell and Bond (1998, 2000) by assuming that $\eta_{ilrh} = \eta_i + \mu_{ilrh}$ and $\mu_{ilrh} = \rho\mu_{ilr(h-1)} + \nu_{irh}$. In order to obtain consistent estimates in the presence of serial correlation, we first estimate an augmented model that includes first lags of inputs and of output by System GMM and then apply a minimum distance procedure in order to recover the coefficients on observable inputs and the correlation parameter ρ. The implementation of such procedure is computationally demanding. So far, we have run the estimation on a random sample of 20% of the data. The results related to this estimation procedure should thus be interpreted with this caveat.

In addition to the simultaneity problem discussed above, production function in an unbalanced panel might suffer from the selection bias firstly described in Olley and Pakes (1996). The fixed effect estimators and the Blundell and Bond procedure take this into account if we are willing to assume that exit is determined by the systematic component of the unobserved productivity, η, and not by the idiosyncratic shocks, ϵ. We plan to further

address both the endogeneity and the selection concerns by applying Olley and Pakes (1996) semiparametric estimation procedure in future work. In addition, as a robustness check we plan to provide some insight on the magnitude of the selection problem by running the estimation of the production function in Equation (3) regression on two alternative samples. The first one drops accounts that leave the scheme before 2004. The second also drops those accounts that enter the scheme after 1990.

Throughout the paper, unless indicated otherwise, we use two-way clustering of standard errors. First we allow correlation at field level. Second, we allow correlation of the errors within the quarter to capture the fact that rainfall shocks are highly spatially correlated across different rainfall stations at a given point in time.

Table 2 reports the estimation outcomes. We highlight three main results. First, the coefficients on rainfall shocks confirm the patterns highlighted in Equation (2): rain has a stronger effect during planting cycle. Second, we get largely different coefficients on log plot size across specifications. Consistently with the findings in the firm level data literature (Ackerberg et al. 2007), the FE estimations produce possibly unrealistically low values of the coefficients on land. Given the sizable difference in coefficient, it will be important to assess whether subsequent stage regressions, which use estimates from the alternative production function estimations, deliver consistent results.

Third, the harvest year dummies (not reported) present a strong and systematic negative trend, consistently with Figure 4. When talking to the company staff, the primary source of this pattern is reported to be soil depletion. While this is a first order determinant of yields, we are not able to say much about the specific nature of these patterns. We attempted to explain these patterns by allowing for time varying coefficients in plot sizes (finding no clear evidence of such changes), and by controlling for the number of years a certain account has been in the sample. Yet, these steps did not reduce the magnitude and the strong trend of the harvest year dummies. In the rest of the paper, we often report predicted production values after purging the harvest year fixed effects.

We briefly discuss several robustness checks we implemented but do not report in the main body of the paper ***ROBUSTNESS CHECK APPENDIX TO BE WRITTEN***. First, we ran our regression on several alternative samples: i) dropping harvests for which we had to impute harvest lag , ii) dropping the two sublocations for which we had high attrition, iii) balanced panel which drops accounts whose first harvest is after 1990 and last harvest is before 2005. None of these alternative samples delivers significantly different estimates of the coefficients θ and γ. Second, we allowed the impact of rain to vary with plot size, without finding a significant coefficient in the interaction term. Third, we use an alternative dependent variable: profits per hectare. This is the difference between cane revenues and the costs of all the inputs charged by the company, including interest. Notice

again that this measure does not include labor costs. The measure is deflated by using cane prices charged in that harvest. We report coefficients comparing them to the ones observed when the dependent variable is yield (tons/ha). Using OLS, an increase in 10 log points in plot size reduces yields by 1% of its mean and profit per hectare by 0.4% of its mean. Using Account FE, the effects magnify: an increase in 10 log points in plot size reduces yields by 4% of its mean and profit per hectare by 3.7% of its mean. Thus, the evidence for decreasing returns to scale is found both when looking at land productivity and when focusing on profitability. Fifth, we include lagged values of the plant cycle, the rainfall shock and the interaction of the two as additional controls. We find that current rainfall shock indicators do not predict production level in the subsequent cycle, once the other regressors are controlled for. Combined with the finding that current rain does not predict future rain, this is an important point for the rest of our analysis. It suggests that experiencing a negative rainfall shock event in the current cycle does not affect the distribution of productivity in the subsequent cycle.

4.2 The Land Allocation Decision

After identifying an appropriate measure of rainfall shocks and developing alternative estimates of residual productivity in the production function regressions, we now focus on the determinants of land allocation to sugar cane.

Our baseline estimating model is:

$$a_{ir+1,1} = \delta_0 + \delta_s shock_{ir1} + \delta_e \eta_{ir+1,1} + \delta_{se} shock_{ir1} * \eta_{ir+11} + \delta_a a_{ir1} + \delta_{sa} shock_{ir1} * a_{ir1} + u_{ir1} \tag{4}$$

Estimation of Equation (4) uses only data from the first harvest (the plant harvest) of each replant cycle and thus obtains identification from accounts that have at least two replant cycles (81.5% of the accounts). We use several alternative measures of the systematic component of the residual land productivity, each of them being related to the four different production function regression specifications presented in the previous section. Given the sizable differences in the estimates of the coefficients on observable inputs (i.e. plot sizes), it is particularly important to address how the estimates of the coefficients in Equation (4) vary across alternative specifications. In the first specification, we just assume that the producers expect their residual land productivity to be equal to the residual productivity observed in the previous plant harvest. Yet, such an approach bears a major drawback. It ignores the fact that the unobserved productivity shapes the input usage decisions in the first stage, but then tests for the presence of this relation. The second measure of unobserved productivity derives from the Account FE regression. In this case

the first stage regression assumed that the component of unobserved productivity known to the producer was constant across periods (η_i). Similarly, the third measure assumes the unobserved productivity known to the producer to be constant within replant cycles (η_{ir}). Finally, we employ a fourth measure based on the Blundell and Bond estimation, which assumes $\eta_{irh} = \eta_i + \mu_{irh}$ with μ_{irh} AR(1) with autocorrelation ρ . We thus construct the unobserved productivity expected by the producer: $\eta_{ir+1,1}$. We proceed in the following way. We derive the total residual from the first stage estimate η_{irh}. Given that we ran the estimation only for 20% of the data, we computed predicted output for the observations out of the sample using the coefficients of the regression and then defined the residuals. We then estimated the time invariant component η_i using a within estimator with serial correlation, imposing the estimate of ρ from the first stage regression. We then compute $\eta_{ir+1,1} = \eta_i + \rho\mu_{ir1}$. Tables 3 and 4 present summary statistics and correlations across the different unobserved productivity measures.

Before presenting the results, we emphasize a few important caveats. First, our second stage regressions include estimates derived in the first stage production function regressions. Therefore, concerns related to the presence of generated regressors arise. As it is well known (Pagan 1984) parameters estimates in models that include generated regressors are consistent but the standard errors are invalid since they do not take into account sampling variation in the first stage estimates. We plan to tackle the problem in future work using proper adjustment procedures. Second, while exit is a rare event, as shown in Figure, 2 we still have some attrition. Due to the unbalanced nature of our panel data, the standard selection concerns arise as we only observe accounts in the harvests cycles where they allocate a positive amount of land to sugarcane. Our data lack a variable that shapes the presence in the scheme but not the land use intensive margin. As a consequence, we cannot implement standard unbalanced panel data correction procedures. In future work, we will partially explore the severity of the selection problem by running the regression on the subsamples that include only accounts that do not leave the scheme and that do not enter in late years of our sample. Third, in the below discussion on cleansing recessions, we will assume that the unobserved productivity is an account fixed characteristics, which does not change as a certain plot expand or contracts. For instance, this implies we are assuming homogenous land quality within an account. Given the small size of most plots in our sample, this looks like a reasonable assumption. Obviously, account expansion will still lead to a decline in expected yields, via decreasing returns to land.

Table 5 presents results for each of the unobserved productivity measures we described above. Several clear patterns emerge across all the specifications. First, a rainfall shock experienced in the plant harvest in a given cycle reduces plot size in the subsequent replanting by 4 to 5 log points. We interpret this as strong evidence of persistence of

exogenous income shocks. With imperfect capital markets, income shocks, and thus cash flow, might affect investment decisions, a result commonly found in the investment literature (Fazzari, Hubbard, and Petersen, 1988). An important point to stress is that experiencing a rainfall shock in the current period affects neither productivity in the subsequent harvest nor the distribution of rain. In other words, it does not change the objective productivity distribution in the next harvest, keeping other factors constant. Thus, the principal channel through which rainfall shocks operate is likely to be their effect on current farmer's profits. Using the estimates from the Blundell and Bond production function regression, we find that a 10% exogenous one time income shock in the current cycle reduces the amount allocated to cane by 2.5-3% in the subsequent cycle. Second, all of our measures of unobserved productivity are strongly correlated with the amount of land allocated to sugar cane. For instance, controlling for current plot size, an increase in one standard deviation in the productivity measure defined using the Blundell and Bond procedure raises plot size by 5.6%. Third, we find evidence that different types of producers react differently to exposure to rainfall shock. Accounts with smaller plots and higher unobserved productivity are less negatively affected by rainfall shocks. Notice that the coefficient on the interaction term between rainfall shock and log plot size is always strongly significant. On the contrary, the interaction term between unobserved productivity and rainfall shock is not significant at conventional confidence levels in most specifications. Using the estimated coefficients, we find that the sign of the impact of rainfall shocks on land allocation actually switches, becoming positive, for the top 10th percentile of the unobserved productivity distribution according to the account FE procedure and for the bottom 25th percentile of plot sizes. The two one-tailed tests give p-values of 13% and 2%, respectively.

We interpret these results as suggestive evidence that "better" accounts, with either higher residual productivity or higher marginal returns to land, expand relatively more in "recessions". The result does suggest that rainfall shocks, while definitely having a negative static effect on productivity, might have a positive dynamic effect. We will attempt to quantify the relative magnitude of these two forces in the simulation exercise we present in the next section.

5 Aggregate Productivity: Misallocation and Dynamics

In this section, we use the results from the regressions presented above to shed light on aggregate productivity patterns. First, we compute the distribution of land across producers

chosen by a production-maximizing social planner (the "optimal" distribution). We emphasize the role of heterogeneous unobserved productivity and decreasing returns to scale as forces pushing, respectively, toward concentration and fragmentation. We compare the properties of the optimal land distribution with the ones of the actual distribution and quantify the losses from misallocation. Second, we provide further insights on the role of bad rainfall shocks and on the impact of rain shocks via a negative static effect on production levels and a positive dynamic effect on reallocation of land shares toward better and smaller producers. It must be noted that the results in this section are extremely preliminary. The findings we present are highly dependent on the assumptions and the values of the parameters chosen for each exercise and as such require further robustness checks that we plan to implement in future work.

5.1 Optimal Land Allocation: Fragmentation or Consolidation?

We study the problem of a social planner who aims to maximize total production taking as given the amount of land under production in each sublocation in each harvest year. The social planner problem of optimal land allocation incorporates the trade-off between heterogenous productivities, which push toward land concentration in the hand of the more productive accounts, and decreasing returns to land, which, on the contrary, push toward land fragmentation. We are interested in studying where in this trade off the actual distribution stands relative to the optimal one. First, we frame the social planner problem, ignoring all the exogenous variables, such as rainfall, geographical and time dummies. In a given year t and sublocation l, the social planner can allocate land across all the producers who replant in that year, taking as given the total amount of land allocated to sugar cane:

$$\begin{aligned} \{A^*_{ilt}\} &= \arg\max_{\{A_{ilt}\}} \sum_{i \in L_t} T_i A^{\theta}_{ilt} \\ s.t. &\sum_{i \in L_t} A_{ilt} = A_{lt} \end{aligned} \tag{5}$$

where A_{ilt} is the plot size of account i in year t in sublocation l, A_{lt} is the total amount of land replanted allocated to cane plant in t, and T_i is the account productivity. We use our estimates from the Account FE production function regression to assign the values of θ (0.572) and T_i plugging our estimates . We then solve the allocation problem and compute the optimal land amount for each producer active in year t, which follows the

following first order condition:

$$\frac{A_j^*}{A_k^*} = \left(\frac{T_j}{T_k}\right)^{\frac{1}{1-\theta}} \tag{6}$$

Accounts with higher T are allocated larger plot sizes with an intensity that is increasing in the degree of the returns to land.

It is important to notice that, in performing this exercise, we are making a stronger assumption about the nature of T, relative to the one we made when looking at individual land allocation decisions. We assume that T represents a producer specific characteristics which stays constant as the producer takes over other land. In other words, we assume that heterogeneity in T comes purely from management and labor, not from land quality. We further discuss this assumption at the end of this section.

We compare the properties of the distributions of A_i and A_i^* across locations and years. First, Figure 6 shows a nonparametric regression of the optimal plot size against the actual plot size. We find that the predicted optimal plot size is increasing in the actual one. However, the graph shows that, on average, plots with size below median value have optimal size above their actual one, and viceversa for plots with size above median. This suggests that, for the given amount of land, total production could increase by shifting land from large accounts to smaller ones.

Second, for each harvest year, we compute the Gini Coefficient for the actual land distribution and for the optimal land distribution. The results are reported in Figure 7. We find that the standard deviation and the Gini Coefficient for the optimal land distribution are consistently higher than the ones for the actual land distribution, even using a relatively low value of θ such as 0.572. The redistribution toward small plots and the increase in the inequality in the optimal plot size distribution are only apparently in contradiction with each other. Figure 8 clarifies, by focusing on one location in a given year. The reallocation of land in the optimal allocation occurs between large producers and *some* small producers. This implies that accounts having the same actual size end up with substantial difference in optimal plot size, with some plots becoming smaller for each initial size level. This drives up variance, and thus inequality, in the optimal distribution. In the location "R" in Figure 8 the Gini Coefficient is 0.27 in the actual land size distribution and 0.48 in the optimal one. At the same time, the graph clearly displays the average reallocation of land toward small producers in the optimal distribution.

Finally, we quantify the losses from the reported misallocation. For each sublocation and year we compute expected productivity TL^θ both using the actual land sizes and the optimal ones. Total output losses are found to span between 5 and 10%.

As we discussed above, the exercise assumes that the unobserved productivity is a producer's characteristic rather than a land one (i.e. soil fertility). This is a problematic

assumption if we think that productivity primarily arises from soil characteristics. In order to address this concern, we slightly modify the social planner problem and allow for reallocation of land only among producers belonging to the same field. For this case, the assumption of similar land quality at a given point in time is more realistic. The main results of our exercise - that optimal distribution implies reallocation toward small producers and an increase in inequality - are robust to this modification.

5.2 Static and Dynamic Effects of Rainfall Shocks: Simulation

In this section, we attempt to shed further light on the role of rainfall shocks in shaping the aggregate levels of production and yields. The regressions presented in Section 4 highlighted that rainfall shocks have a negative static effect on yields. However, rainfall shocks also foster reallocation of production shares toward accounts with higher unobserved productivity and smaller plot sizes. We develop a simple simulation in order to quantify the net effect of these forces. For simplicity, we assume that all the accounts in the economy harvest a plant cycle in every period t and do not enter ratoon harvest cycles. In the simulated economy, production occurs according to the function:

$$\begin{aligned} y_{ift} =& \beta_0 + \beta_s RainShock_{ft} + \beta_a LogPlotSize_{ift} + \eta_i + \mu_{ift} \\ & \mu_{ift} = \rho \mu_{ift-1} + \nu_{ift}; \\ & \nu_{ift} \sim \mathcal{N}(0, \sigma_\nu^2), \end{aligned} \tag{7}$$

where i is the account, f is the field and t is the time period.

Plot sizes evolve according to the equation:

$$\begin{aligned} LogPlotSize_{ift+1} = & \delta_0 + \delta_s * RainShock_{ft} + \delta_e(\eta_i + \rho\mu_{ift}) + \delta_{se} RainShock_{ft} * (\eta_i + \rho\mu_{ift}) \\ & + \delta_a LogPlotSize_{ift} + \delta_{sa} RainShock_{ft} * LogPlotSize_{ift} + \xi_{ift}; \\ & \xi_{ift} \sim \mathcal{N}(0, \sigma_\xi^2) \end{aligned} \tag{8}$$

We simulate for 10 periods an economy that evolves according to Equations (7) and (8). Each simulated economy is populated by 10,000 accounts grouped in fields of 10 accounts each. We calibrate the parameters of Equations (7) and (8) using the estimates from the regressions in Section 4. For each simulation, we draw log plot sizes and unobserved productivity η_i from a bivariate normal distribution with mean and covariance matrix computed using the first four years of data of our sample. Similarly, we draw all the idiosyncratic terms using variance estimates from the regressions.

In order to disentangle the static and dynamic effects of rainfall shocks, we perform a simple thought experiment. We compare total production and yields under two scenarios, both spanning for 10 periods, $t = 1, 2...10$. In the first one, throughout the 10 periods, rainfall shocks occur with the frequency observed in the data, 7%. In the second one, we allow an increase of the frequency of shocks in one period, $t = 4$. Specifically, we assume shocks involve *all* of the plots, a five-fold increase relative to the observed frequency. We label this case as the "aggregate shock" scenario. For each of these two cases we run 50 simulations for each t. Importantly, we assume that all the model parameters, including those capturing responses to the shock events, are not affected by the one period change in the prevalence of the shocks.

Figures (9) and (10) present the results of the exercise. Period 4, in which the one-time increase in shock frequency occurs, presents a remarkable drop in both total production and yields, consistently with the strong negative effect of shocks on production levels highlighted by the production function estimation. The contemporary decrease in production levels following a shock is about 15%. However, in period 5, both production and yields bounce back *above* the counterfactual levels realized in the scenario with no aggregate shock. Production increases by about 2% while yields go up by 1.5%. In addition, production levels remain above the scenario with no aggregate shock during the subsequent periods, although the difference decreases over time. The increases occur as the one-period aggregate shock fosters reallocation of land toward producers with higher unobserved productivity and higher marginal returns to land.

The finding that the static shock dominates seems uncontroversial. On the contrary, more work is required to assess the quantitative impact of the subsequent reallocation under different scenarios.

6 Conclusion

We look at determinants and efficiency of allocation of land to sugar cane using a novel panel database from Western Kenya. The database covers 15,000 accounts for twenty years (1987-2006). Our analysis delivers three main preliminary results. First, when compared to the optimal land allocation, the actual allocation has too much land concentrated in relatively large plots. However, the optimal distribution also exhibits higher variance and a higher degree of inequality, as a consequence of the reallocation toward accounts with high residual productivity. Second, negative rainfall shocks determine a relative reallocation of land toward accounts that are more productive per hectare, either because of higher marginal returns to land or because of higher residual land productivity. Using our estimates, we also find that accounts in the top 10% of the residual productivity distribu-

tion and in the bottom 25% of the plot size distribution expand their cane plot following a rainfall shock, although the effect is statistically significant only for the latter effect. This points at a potential role of rainfall shocks in raising aggregate land productivity via a "cleansing" effect. Future work will assess robustness of these results to a range of alternative sub-samples and to alternative specifications.

There are several topics we plan to focus on in future work. First, we plan to delve into the change in the names of contracted farmers in each account. A correlation between changes in farmers' name, rainfall shock and residual productivity might point at other interesting sources of reallocation within the sector. For instance, the occurrence of shocks might foster or hinder the entry of higher productive managers into low productivity accounts, possibly through rental agreements. We also plan to focus on patterns of trend-adjusted productivity following a change in plot size manager. The focus on farmers will also allow us to disentangle the component of residual productivity coming from land attributes from the one coming from manager's ability. Second, we would like to study how cane profitability varies with distance from the mill and whether exogenous shocks have a differential effect on accounts according to this distance. Third, we are trying to collect historical data on maize prices and yields. Such a database would allow us to study whether food security shocks affect the amount of land allocated to cash crop across the years and how this effect varies across farmers with different productivity.

Should the preliminary evidence of cleansing recessions be confirmed, the results could have important implications for the cost-benefit analysis of policies that aim to help farmers to cope with adverse rainfall shocks. The fact that, at least in the study area, negative (rain) shocks lead to a better allocation of land should be taken into account when elaborating such policies. Shock coping interventions might hinder the cleansing mechanism described above. More broadly, our study points at the importance of reallocation as a source of aggregate productivity. Understanding how to foster the differential expansion of better farmers is an important goal for policy makers. One option might be yield-based pricing which rewards higher productive farmers. Using the available data, we will use price changes in future work in order to try to measure sensitivity of production and yields to change in prices and to thus elaborate suggestions on optimal pricing schedules. Similarly, financial products where loans are conditional on basic productivity measures might prove effective in providing extra resources to farmers with higher potential. We are currently piloting a scheme along these lines in partnership with the same company that provides us with the data used in the above analysis.

References

Ackerberg, D., Benkard, C. Berry, S. and A. Pakes (2007), "Econometric Tools for Analyzing Market Outcomes". In: J. Heckman and E. Leamer (eds.): *Handbook of Econometrics*, vol. 6(1): 4171-4276

Binswanger, H. K. Deninger, and G. Feder (1995), "Power, Distortions, Revolt and Reform In Agricultural Land Relations". In: J. Behrman and T.N. Srinivasan (eds.), *Handbook of Development Economics*, Volume III, Elsevier

Block, S. (2010), "The Decline and Rise of Agricultural Productivity in Sub-Saharan Africa Since 1961", NBER WP 16481, October.

Blundell, R.W. and S.R. Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models", *Journal of Econometrics*, 87, 115- 143.

Blundell, R.W. and S.R. Bond (2000), ""'GMM Estimation with persistent panel data: an application to production functions", *Econometric Reviews*, 19(3), 321- 340.

Conning, J. and C. Udry (2007), "Rural Financial Markets in Developing Countries". In: Evenson R. and P. Pingali (eds.), *Handbook of Agricultural Economics. Vol. 3*, vol. 3, pages 2857-2908, Elsevier, Amsterdam.

Collier, P. and S. Dercon (2009), "African Agriculture in 50 Years: Smallholders in a Rapidly Changing World?", prepared for the *FAO Expert Meeting on How to Feed the World in 2050*.

Duflo, E., M. Kremer, J. Robinson (2011), "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya", *American Economic Review*, forthcoming.

Eastwood, R., M. Lipton, A. Newell, "Farm Size". In Evenson R. and P. Pingali (eds.), *Handbook of Agricultural Economics*, vol. 4, pages 3323-3397, Elsevier, Amsterdam.

Foster, A. and M. Rosenzweig (2010), "Barriers to Farm Profitability in India: Mechanization, Scale and Credit Markets", mimeo

Goldstein, M and C. Udry (2011), "The Profits of Power: Land Rights and Agricultural Investment in Ghana," *Journal of Political Economy*, vol. 116(6), pages 981-1022, December.

Griliches, Z, and J. Mairesse (1998), "Production Functions: The Search for Identification", NBER WP 5067.

Grosh, M. (1994), "Contract Fanning in Africa: an Application of the New Institutional Economics", *Journal of African Economies*, vol3(2):231-261.

Hsieh, C. and P. Klenow (2009), "Misallocation and Manufacturing TFP in China and India", *The Quarterly Journal of Economics*,, vol. 124(4), pages 1403-1448, November.

Kenya Sugar Board (2005), *Year Book of Sugar Statistics*

Kenya Sugar Board (2011), *Kenya Sugar Industry Strategic Plan, 2010-2014*.

Mundlak, Y. (2001), "Production and Supply". In: B. Gardner and G. Rausser (eds.), *Handbook of Agricultural Economics*, vol. 1, pages 3-85, Elsevier, Amsterdam.

Olley, S. and A. Pakes (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, Vol. 64, No. 6, November, pp. 1263-1297.

Pagan, A. (1984), "Econometric Issues in the Analysis of Regressions with Generated Regressors", *International Economic Review*, vol. 25(1), pages 221-47, February.

Rosenzweig, M. and H. Binswanger (1993), "Wealth, Weather Risk and the Composition and Profitability of Agricultural Investments", *The Economic Journal*, vol. 103, No. 416,

January, pp. 56-78
Syverson, C. (2011). "What Determines Productivity", *Journal of Economic Literature*, vol. 49(2): 32665.
World Bank (2009), *World Development Report 2008. Agriculture for Development.*

Figure 1: Number of obs per Harvest Year

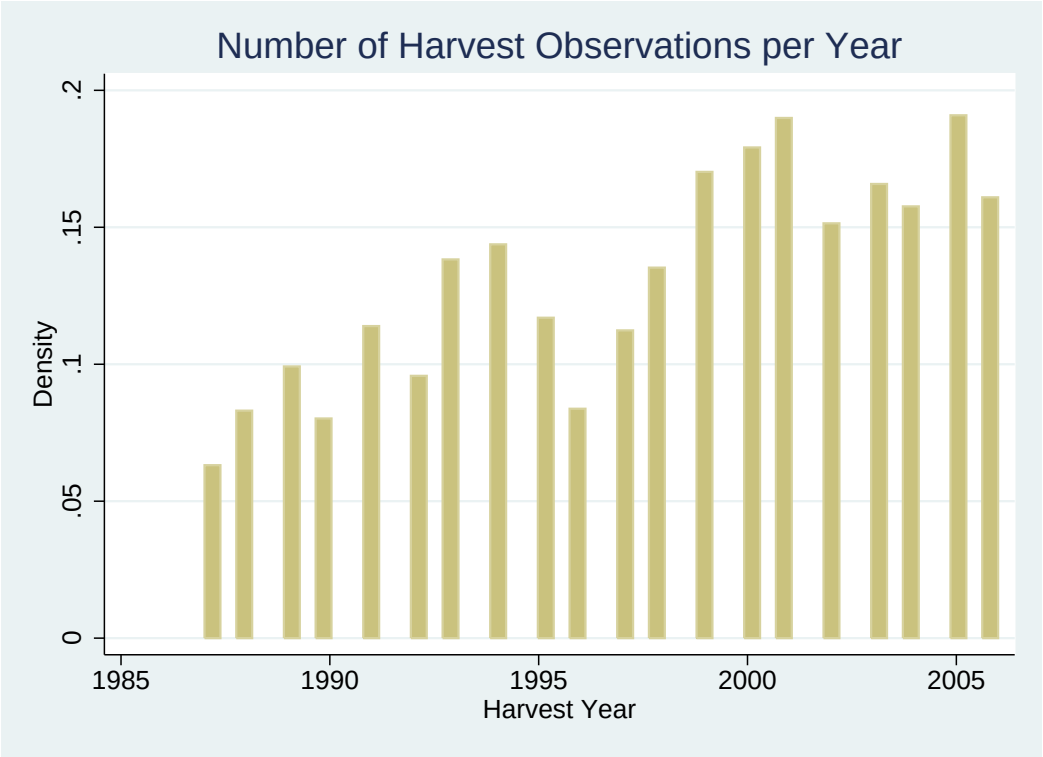


Figure 2: Entry and Exit by Harvest Year

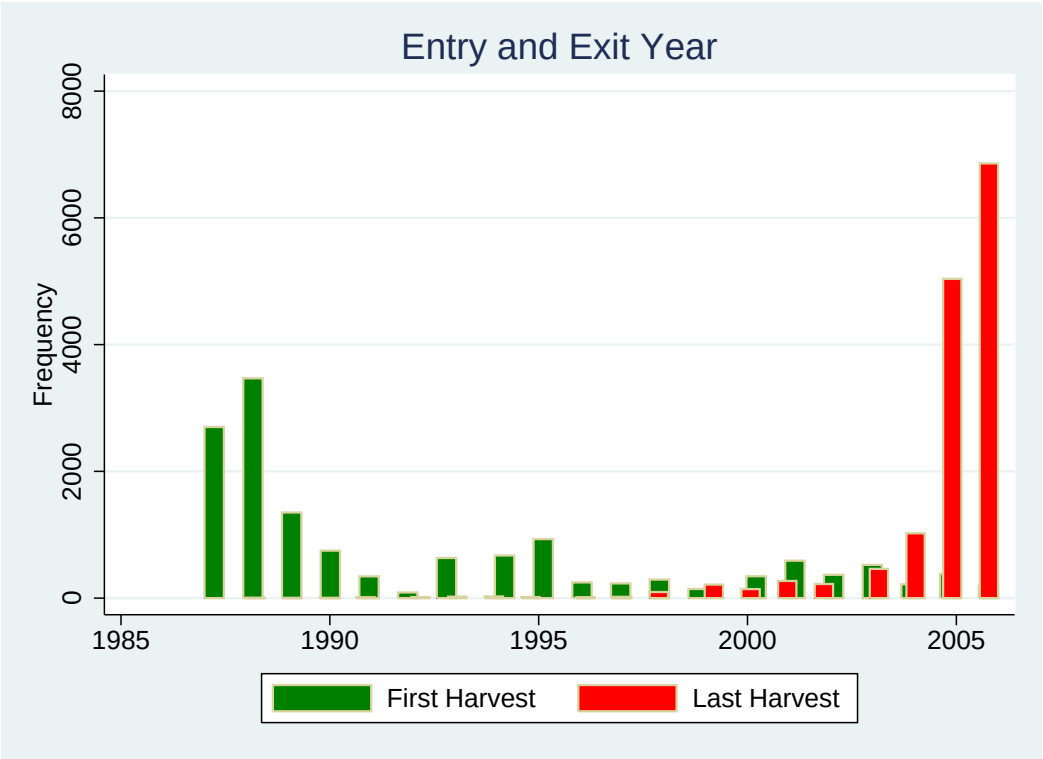


Table 1: Harvest Level Summary Statistics

Variable	Mean	Std. Dev.	N
Total cane: tonnes	64.757	50.473	112319
Plot Size	0.877	0.517	112319
Yield (tons/ha)	74.818	35.776	112319
Profit per Hectare (deflated)	33.193	23.077	111073
Plant Cycle	0.351	0.477	112319
Harvest Cycle Length (months)	20.741	5.096	112319
Transport zone (1-8)	4.953	1.33	111214
Avg rain 10 months, stn 1	152.174	32.037	112319
Rain 10th pct,stn1,10 months	0.13	0.336	112319
Plant*(Rain 10th pct,stn1,10 months)	0.155	0.362	39383
Rain 10th pct,stn 1-2,10 months	0.053	0.224	112319
Plant*(Rain 10th pct,stn 1-2,10 months)	0.069	0.254	39383

Figure 3: Plot Sizes time series

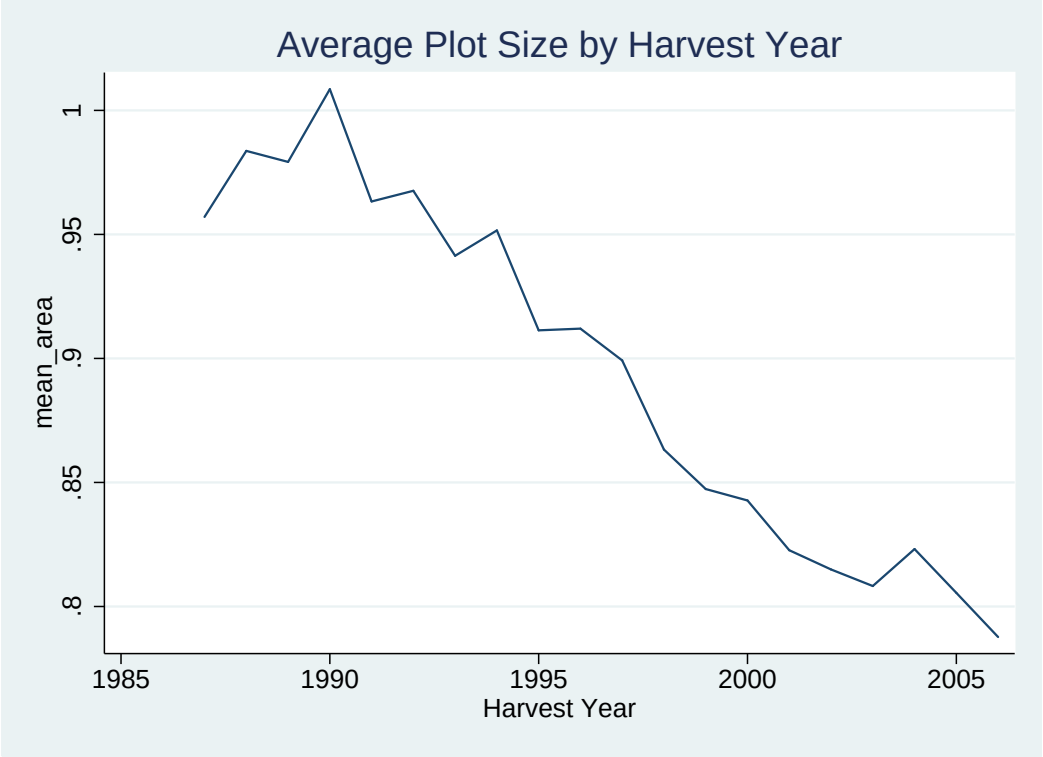


Figure 4: Yields (tons/ha) time series

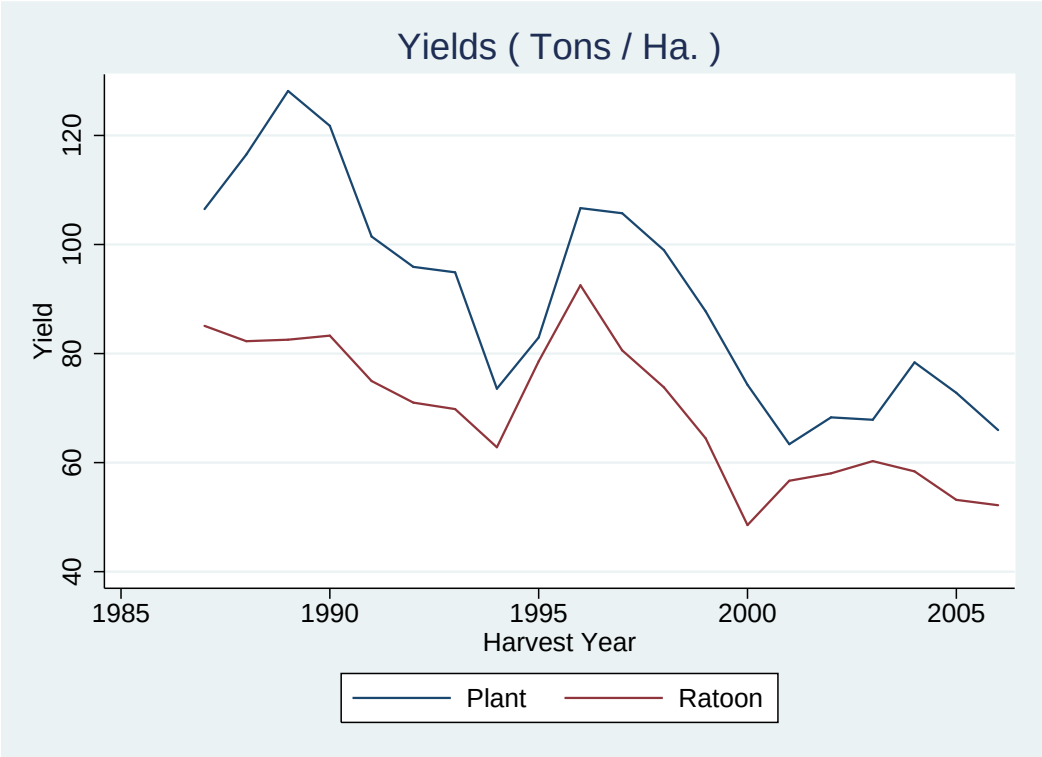


Figure 5: Partial Linear Regression LogOuput vs. Mean Rain 10 months

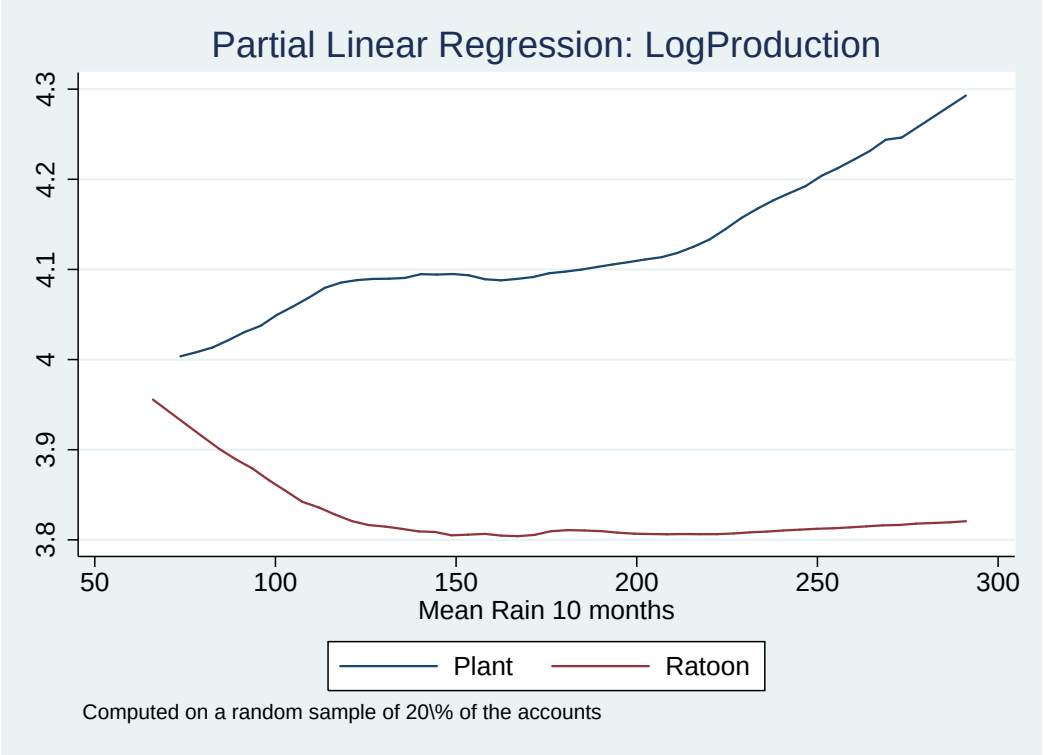


Figure 6: Optimal vs. Actual Plot Size

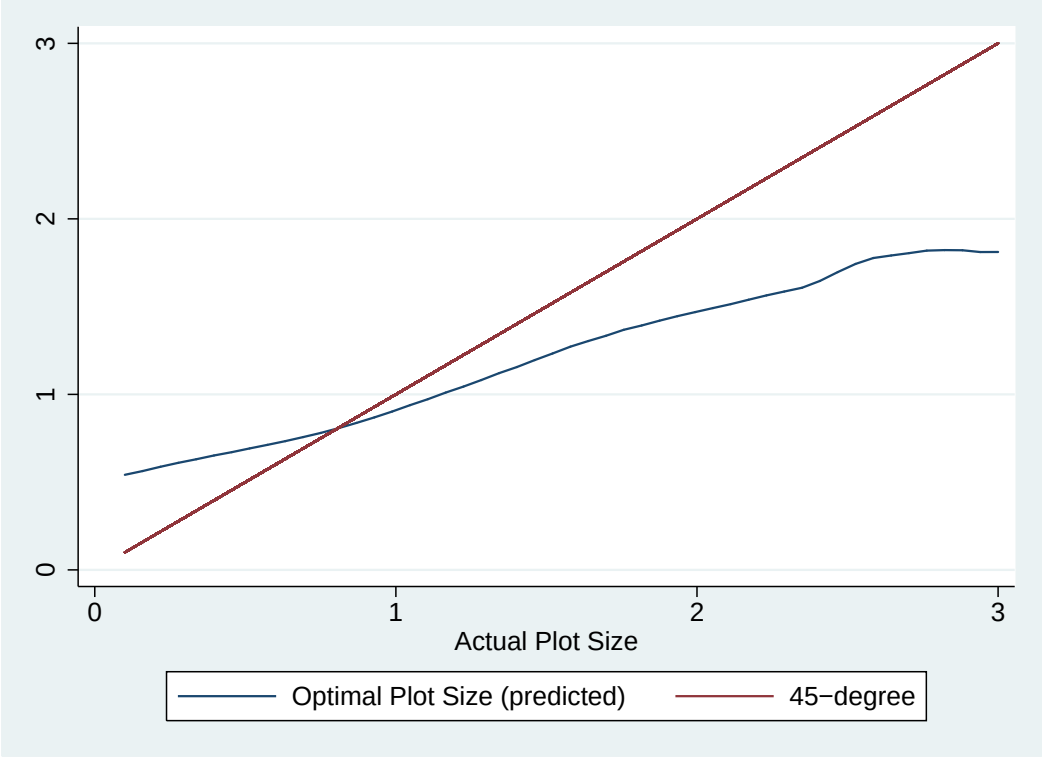


Figure 7: Inequality in Actual and Optimal Land distribution

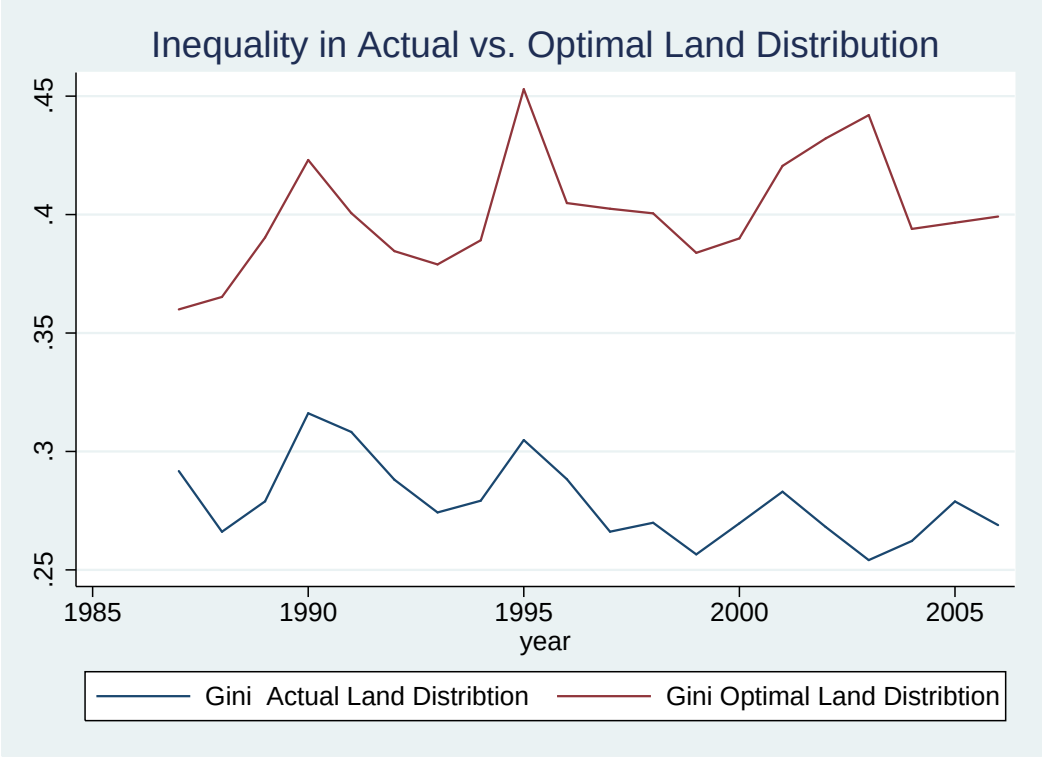


Figure 8: Optimal vs. Actual Plot Size (One Location 1990)

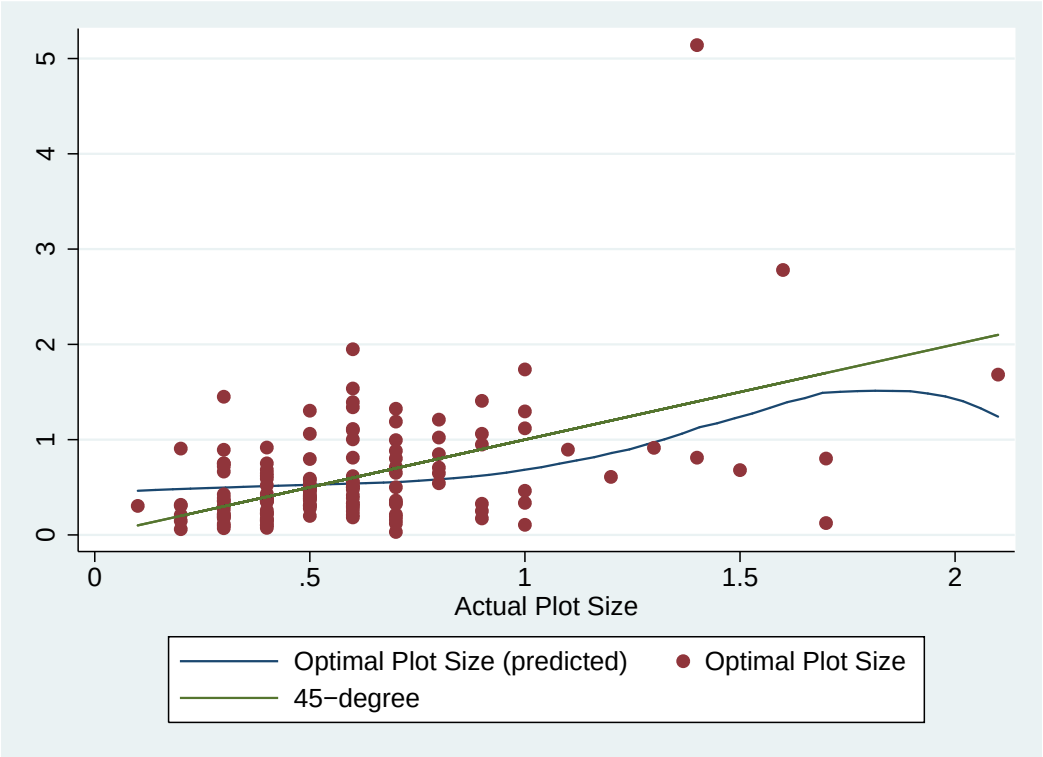


Figure 9: Simulation: Total Production

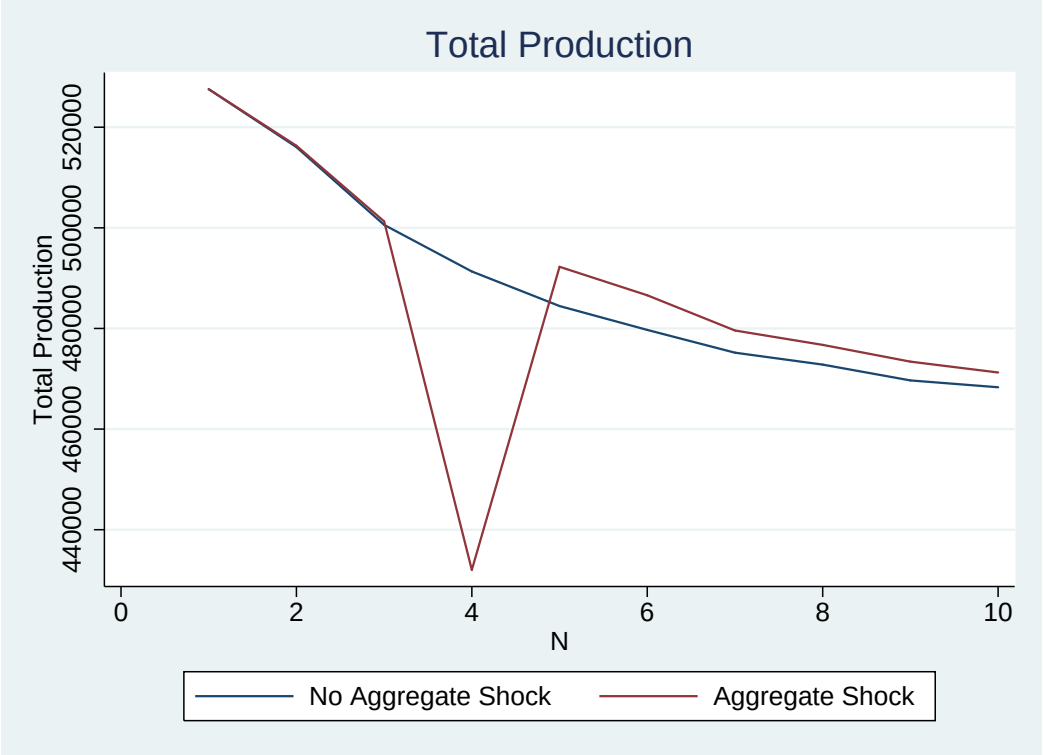


Figure 10: Simulation: Yields (Tons/ha)

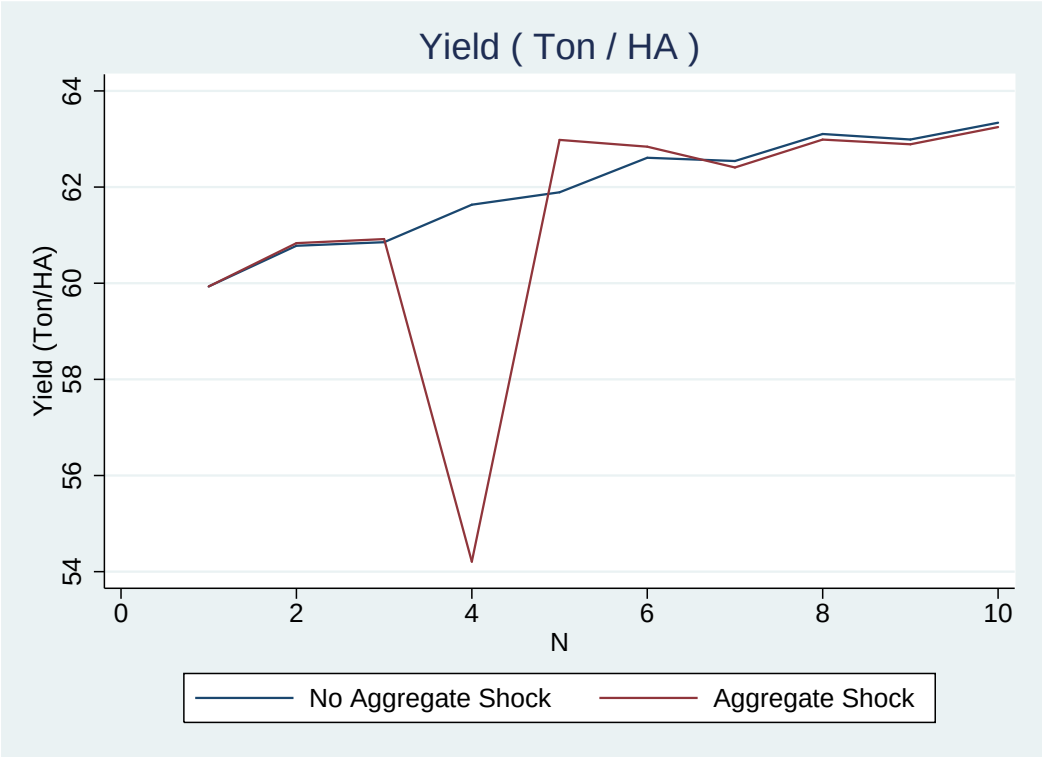


Table 2: Production Function Regression

	(1) OLS	(2) AccountFE	(3) ReplantFE	(4) BlundellBond
Rain 10th pct,stn 1-2,10 months	0.007	-0.010	-0.020	-0.036
	(0.023)	(0.020)	(0.018)	(.022)
Plant*(Rain 10th pct,stn 1-2,10 months)	-0.146***	-0.073*	-0.106***	-0.145***
	(0.049)	(0.039)	(0.035)	(0.043)
Plant Cycle	0.289***	0.308***	0.180***	0.358***
	(0.016)	(0.013)	(0.017)	(0.012)
Log Plot Size	0.851***	0.572***	0.167***	0.614***
	(0.009)	(0.020)	(0.022)	0.082
R^2	0.464	0.284	0.283	N/A
ρ				0.232
Sublocation FE	Yes	No	No	No
Mean Dependent Var	3.904	3.904	3.904	3.904
Rain Vars p-value	.006	.032	.001	.003
Observations	112347	111753	103164	19376

Standard errors clustered at field and quarter level

* p<0.1, ** p<0.05, *** p<0.01

Regression includes Harvest Year FE

Rain shock =1 if mean rain in first 10 months of harvest cycle below 10th percential in both closest rainfall stations

Table 3: Unobserved Productivity Estimates: Summary Statistics

Variable	**Mean**	**Std. Dev.**	**N**
OLS	0	0.567	39396
Account FE	-0.022	0.381	39396
Replant FE	0.167	0.653	25365
BlundellBond	-0.031	0.407	38752

Table 4: Correlation Across Unobserved Productivity Estimates

Variables	OLS	Account FE	Replant FE	BlundellBond
OLS	1.00			
Account FE	0.48	1.00		
Replant FE	0.19	0.70	1.00	
BlundellBond	0.64	0.92	0.59	1.00

Table 5: LogPlotSize in subsequent replant

	(1) OLS	(2) OLS	(3) Account FE	(4) Account FE	(5) Replant FE	(6) Replant FE	(7) BlundellBond	(8) BlundellBond
Rain Shock	-0.050**	-0.073**	-0.042*	-0.069**	-0.037*	-0.067*	-0.048**	-0.076**
	(0.025)	(0.030)	(0.023)	(0.032)	(0.022)	(0.037)	(0.023)	(0.032)
Log Plot Size	0.745***	0.755***	0.697***	0.710***	0.654***	0.665***	0.710***	0.722***
	(0.011)	(0.008)	(0.014)	(0.009)	(0.015)	(0.011)	(0.014)	(0.009)
Shock*Log Plot Size		-0.129**		-0.164**		-0.137*		-0.166**
		(0.051)		(0.073)		(0.075)		(0.070)
Unobserved Productivity	0.062***	0.055***	0.177***	0.167***	0.167***	0.165***	0.146***	0.136***
	(0.013)	(0.011)	(0.012)	(0.010)	(0.009)	(0.009)	(0.014)	(0.011)
Shock*Unobserved Prod.		0.107*		0.095		0.011		0.099
		(0.055)		(0.069)		(0.032)		(0.064)
R^2	0.637	0.640	0.644	0.646	0.666	0.667	0.641	0.643
Mean Dependent Var	-.284	-.284	-.284	-.284	-.284	-.284	-.285	-.285
Observations	24963	24963	24963	24963	24963	24963	24862	24862

Standard errors clustered at field and quarter level

* p<0.1, ** p<0.05, *** p<0.01

Controls include Harvest Year and Sublocation FE

Only Plant Cycle Observations

The column headers refer to the estimation strategy used to estimate the production function and the unobserved productivity measure.