Input-Driven Production Technology Heterogeneity and the Allocation of Inputs

Stepan Gordeev[*] Sudhir Singh[†]

October 13, 2023

click for latest version

Preliminary and incomplete

Abstract

We study whether heterogeneity in the intrinsic features of production inputs may generate heterogeneity in production technologies optimally chosen by firms, leading existing estimates of misallocation to overstate its costs. Existing estimates of the severity of the misallocation of inputs across firms rely on assuming homogeneous production technology and interpreting deviations from that technology as evidence of misallocative distortions. We use a state-of-the-art clustering algorithm for ordinal data to group Indian agricultural plots into land types by the intrinsic physical features of each plot. We find that production functions are significantly heterogeneous across land types, which we confirm with placebo-like randomization inference. Some types of land are better suited to land-intensive technology, others to labor-intensive technology, etc. We build a model in which heterogeneous farmers face distortions and choose the type of land to rent. We use the model to quantify the cost of misallocation for India's aggregate agricultural productivity and compare it to conventional models that assume homogeneous production technology.

Keywords: misallocation, technology heterogeneity, input heterogeneity, agriculture, productivity

JEL classification: O4, O11, O13, Q1

[*] University of Connecticut, stepan.gordeev@uconn.edu

[†] Michigan State University, singhsud@msu.edu

1 Introduction

Market frictions and distortionary government policies can prevent the production inputs present in the economy from being allocated to their most efficient use: productive firms may fail to hire inputs to expand or be forced to use suboptimal mixes of inputs (Restuccia and Rogerson 2017). The resulting misallocation contributes to cross-country income differences: it is especially severe in low-income country land markets, drastically depressing their aggregate agricultural productivity (Chen, Restuccia, and Santaeulalia-Llopis 2022). The identification of firm- or farm-level frictions and the quantification of their aggregate output cost relies on measuring the dispersion in marginal products across producers and inputs Hsieh and Klenow (2009). The measurement of heterogeneity in marginal products, however, is sensitive to the assumed production structure. If all producers in the economy use the same production function, then any observed heterogeneity in their input choices reflects misallocative frictions. But if different producers optimally choose different production methods, failing to account for this heterogeneity would lead estimates of misallocation to be overstated.

In this paper, we exploit the heterogeneity in the intrinsic features of a prominent production input (land) and show that farmers operating different types of land use different production approaches. This generates optimal heterogeneity in input choices that is mistaken for evidence of misallocative frictions by standard models.

We use rich farm-plot-level data from India that includes information on the intrinsic physical features of each cultivated plot along several dimensions—soil type, color, salinity, and others—capturing the qualitative differences between land inputs used by different farmers. To reduce the dimensionality of plot differences and facilitate production function estimation, we apply clustering analysis from the field of machine learning to group plots into a handful of discrete types (Han, Kamber, and Pei 2012). The available plot features are categorical but ordinal: commonly used clustering techniques are ill-suited to such data. We apply HD-NDW, a state-of-the-art clustering algorithm developed for ordinal data by Zhang and Cheung (2022). The algorithm groups plots into five land types such that plots are similar within type and dissimilar across types.

We estimate agricultural production functions separately for each land type identified by the clustering algorithm. We find that farmers operating land of different types use different production functions with significantly heterogeneous input elasticities and returns to scale. We conduct a placebo exercise inspired by randomization inference and find that this heterogeneity is not driven simply by slicing the data into types: alternative plot groupings are exceedingly unlikely to produce production functions this heterogeneous.

To study the impact of input-driven production technology heterogeneity on quantifying misallocation, we build a model populated with heterogeneous farmers facing general misallocative distortions, akin to Hsieh and Klenow (2009) and Chen, Restuccia, and Santaeulalia-Llopis (2022). The models of this class allow input and output market distortions to be identified through observed marginal revenue products of individual farms, and for the aggregate output cost of these distortions to be quantified. First, we compare the frictions and the cost of misallocation between a standard model that assumes a single homogeneous production function and a model that allows for production function heterogeneity between land types. Next, we extend the model with an endogenous choice of land types by farmers to understand the effect of market frictions on plot choice by farmers.

The impact of endogenous production technology choice on estimating productivity differences across countries has been explored within manufacturing by Eberhardt and Teal (2020) and within agriculture by Mundlak, Butzer, and Larson (2012). At the micro-level, Li and Sasaki (2017) and Kasahara, Schrimpf, and Suzuki (2023) develop methods of identifying production functions with heterogeneous elasticities. Our exercise instead exploits a single potential driver of heterogeneous production technology choice—the intrinsic features of land—and shows that it is robustly associated with heterogeneous input elasticities. In the misallocation literature, the most related paper is Gordeev and Singh (2023), which estimates production functions at crop level and finds that accounting for product heterogeneity is important for quantifying the cost of misallocation. In contrast, the present paper explores production function heterogeneity driven by immutable land features (rather than endogenous product choice).

The paper proceeds as follows. Section 2 discusses the farm-level data we use and the construction of land types. Section 3 estimates type-specific production functions. Section 4 describes the model and quantifies misallocation. Section 5 concludes.

2 Data

We use the 2007-08 round of the Rural Economic and Demographic Survey (REDS), conducted by India's National Council of Applied Economic Research.[1] It contains detailed information on household characteristics and economic activity of 8,659 households across 242 Indian villages, sampled to be nationally representative of rural India. Of these, 4,803 cultivated land: we restrict our sample to such farm-operating households and treat them as farms.

2.1 Inputs

We focus on three agricultural inputs observed at farm-plot level: land, labor, and intermediate inputs. Labor is measured with the number of days worked by family members and hired workers. Intermediate inputs comprise the total expenditure on seeds, fertilizer, irrigation, rented machinery, and draft animals.

The land input deserves special treatment. It is commonly measured either with the area cultivated or with the reported market price. The former ignores quality differences while the latter is likely to be exceedingly noisy due to how undeveloped the land markets are throughout most of India. Therefore, we adopt a compromise solution of Gordeev and Singh (2023): we estimate a random forest regression (Breiman 2001) to predict the reported price of each plot based on the observed objective features of the plot. We then use the predicted market price of each plot as the quality index. This method effectively captures most variation in reported price but is de-noised and avoids overfitting: see Gordeev and Singh (2023) for details.

2.2 Land Types

Land characteristics. Observable soil characteristics of a plot of land can be viewed as largely intrinsic and immutable features of the land input operated by each farmer. Crucially, there is very little sorting of Indian farmers between plots: 94% of land owners inherited the land, and only 12% of land cultivators

[1] While the survey was collected in five rounds since 1971, only the 2007-08 round includes the plot-level data needed for our analysis.

participate in the rental market.[2] While soil characteristics are likely to be related to plot productivity, our land quality index explicitly captures any population-level relationship between these characteristics and land price.

The REDS survey collected information on six physical characteristics of each plot of land, each assessed on a discrete scale: *top soil depth* (up to 1 ft, 1-3 ft, more than 3 ft), *soil color* (red, black, gray, yellow, brownish black, offwhite), *soil type* (sand, loam, light clay, heavy clay, gravel, latrite), *soil salinity* (nil, moderate, high), *rate of percolation after one round of irrigation* (fast, medium, slow), and *ease of drainage in case of heavy rainfall* (easy, moderate, difficult).

HD-NDW clustering algorithm. To facilitate the empirical and quantitative analysis of whether production technology responds to heterogeneous land characteristics, we seek to reduce their dimensionality by grouping plots into several discrete land types. Each plot in the data can be viewed as a point in the six-dimensional space of measured soil features. Partitioning multi-dimensional spaces into a handful of "clusters" such that points in each cluster are close to one another but distant from points in other clusters is the objective of a well-developed field of cluster analysis within machine learning (Han, Kamber, and Pei 2012).

A number of clustering algorithms are commonly used. Some are designed for continuous data, like *k*-means—others for categorical data, like *k*-modes. Plot characteristics in REDS, however, present an intermediate case of ordinal categorical data: each of the six variables has a clear order, but the distance between possible values is not well-defined.[3] Applying algorithms designed for continuous data would require imposing arbitrary distance metrics between categories. Applying algorithms designed for unordered categorical data would ignore the order. HD-NDW overcomes this tradeoff: it is a novel clustering method developed by Zhang and Cheung (2022) specifically for ordinal categorical data. It is based on iteratively constructing a distance metric between levels of a variable and across variables. HD-NDW outperforms other algorithms in cases where the correct clustering is known and can be used for validation, justifying the use of

[2] Source: Rural Economic and Demographic Survey (REDS) 2007-08.

[3] The "soil type" feature can also be viewed as ordinal as it is ultimately determined by the size of particles forming the soil.

HD-NDW for its purpose: clustering data in which the correct clusters are not known and need to be constructed.

The ND-NDW algorithm groups arbitrary ordinal data into a pre-specified number of clusters k. To pick the optimal number of clusters, we use the average silhouette coefficient metric: it captures the degree to which the average distance between a point and other points in its cluster is low (clusters are compact) but the average distance between a point and other points in different clusters is high (clusters are separated) (Han, Kamber, and Pei 2012). A high value indicates a high quality of clustering. We compare $k = 3, \dots, 10$. The average silhouette coefficient is maximized when $k = 5$, so we use five clusters in the construction of land types.

Land type clusters. Table 1 lists the land type clusters constructed by the HD-NDW algorithm, summarizing each by the mode of each of the used land features. In the next section, we test whether the production technologies used by operators of different land types are similar.

land cluster	top soil depth	soil color	soil type	soil salinity	rate of percolation	ease of drainage	# plots
I	<1 ft	yellow	sand	nil	fast	easy	1746
II	1-3 ft	yellow	light clay	moderate	fast	easy	1024
III	<1 ft	yellow	loam	nil	medium	easy	1761
IV	1-3 ft	red	light clay	moderate	medium	moderate	3485
V	1-3 ft	yellow	light clay	nil	medium	moderate	1999

Table 1: Land type clusters described with the mode of each feature

3 Production Functions

3.1 Estimation

We estimate the Cobb-Douglas production function at the level of plot i belonging to land type cluster c, in season t, allowing input elasticities to vary at the land type cluster level c.

$$y_{c,i,t} = z_{c,i,t} l_{c,i,t}^{\alpha_c} n_{c,i,t}^{\beta_c} m_{c,i,t}^{\gamma_c}$$

$y_{c,i,t}$ is the physical output produced of plot i in season t. Production utilizes three inputs: quality-adjusted land l (with land type-specific elasticity α_c), labor n measured in days worked (elasticity β_c), intermediate inputs m measured as total expenditure on seeds, fertilizer, irrigation, rented machinery, and draft animals. $z_{c,i,t}$ is the total factor productivity.

Taking logs, the regression specification is:

$$\log y_{c,i,t} = \alpha_c \log l_{c,i,t} + \beta_c \log n_{c,i,t} + \gamma_c \log m_{c,i,t} + \epsilon_{c,i,t} \tag{1}$$

Estimating Equation 1 as-is suffers from a well-known simultaneity bias: endogenously chosen observed input allocations are likely to be correlated with the unobserved error term. To overcome this issue, we follow the method developed for plot-level agricultural data by Gollin and Udry (2021). It is based on the idea that observed productivity shocks at plot j should not affect the unobserved productivity of plot i but should affect the shadow cost of inputs of plot i, as long as it is operated by the same farmer. Such exogenous variation in the shadow cost of inputs across plots offers an instrumental variable for plot-level input choices. The second insight of their method is that plot-level shocks can be constructed by interacting observed farm-level shocks with observed immutable plot-level characteristics. We use an array of agricultural, health, and social shocks measured at the household level and interact them with the same soil characteristics that were used for quality-adjusted land input and land type cluster construction in Section 2. We estimate Equation 1 using two-stage least squares (2SLS) and these instruments.

3.2 Production Functions are Heterogeneous Across Land Types

Figure 1 visualizes the estimated input elasticities.[4] The first panel displays the elasticities of a single aggregate production function. The remaining panels display land type-specific estimates.

[4] The complete regression table is shown in Appendix Table A.1.

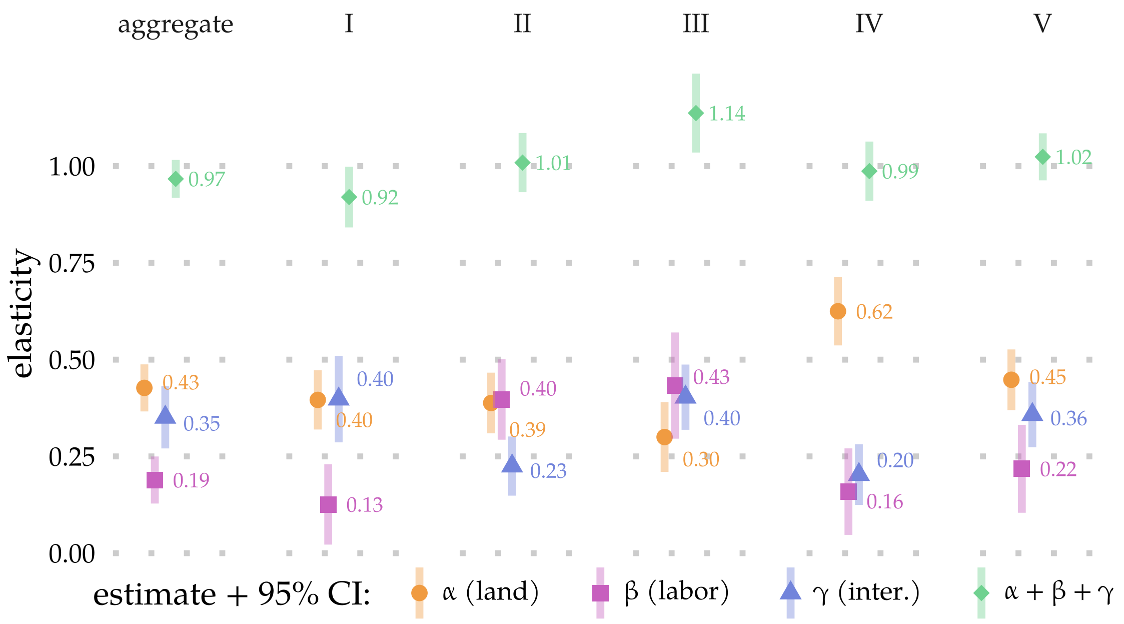

Figure 1: Production function input elasticities, by land type

The aggregate production function has expected input elasticities. Land is the most important input, followed by intermediates and then labor. The three elasticities add up to 0.97, suggesting that returns to scale are insignificantly different from constant. However, different land types have significantly different production functions. Type I exhibits significantly decreasing returns to scale. Types II and III are significantly more labor-intensive. Type III, furthermore, exhibits increasing returns to scale. Type IV's land intensity far exceeds the aggregate one. Only Type V is well approximated by the aggregate production function.

Placebo Test. These results suggest that different land types are indeed associated with different production functions. But can these estimated differences be produced by random noise resulting from the fact that type-specific production functions partition the data into thinner slices? To explore this possibility, we conduct a test inspired by randomization-based inference.

First, we conduct pairwise equality tests for all possible coefficient pairs between land types. For each input, there are $\binom{5}{2} = 10$ pairs of type-specific elasticities of this input. In total, for three inputs, there are $3 \times 10 = 30$ such pairs and thus 30 pairwise equality tests we can conduct. We find that in 15 of these, the equality between the two types' coefficients is rejected at the 10% level.

Next, we repeat this procedure for 1,000 randomly permuted land type assignments: the number of observations in each land type is preserved, but the association between plots and land types is randomized. Each such random permutation constitutes a "placebo" land type assignment. Computing the number of equality tests rejected at each random permutation allows us to construct a null distribution of this metric, presented in Figure 2.

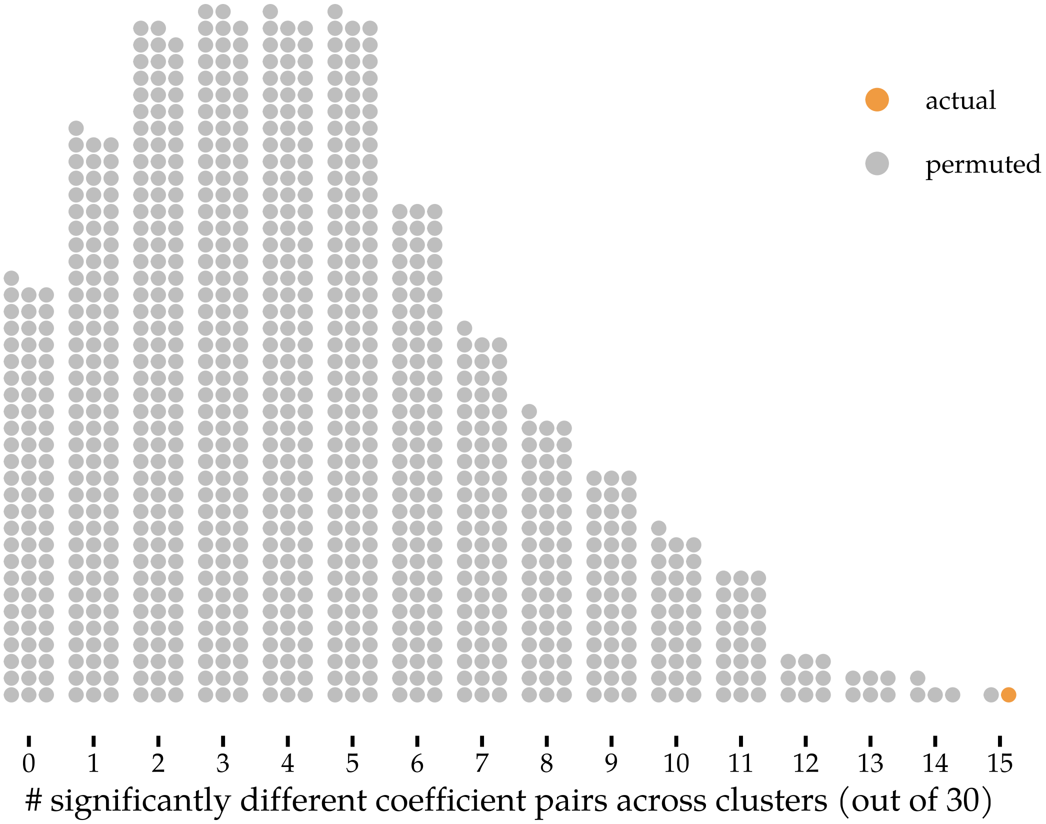

FIGURE 2: Pairwise coefficient equality tests for randomly permuted cluster assignments

Of the 1,000 random permutations, only one matched the number of rejected pairwise equality tests in the true land type assignment (15), and none exceeded it. This result suggests that random slices of the data are exceedingly unlikely to produce production functions that are as heterogeneous as they are between land types constructed by the clustering algorithm.

4 Model

Having determined that Indian farmers operating land of different types use different production functions, we turn to the model to explore the significance of this finding for quantifying misallocation. The model builds on the conventional frameworks of firm- or farm-level misallocation like Hsieh and Klenow (2009), Restuccia and Rogerson (2008), and Chen, Restuccia, and Santaeulalia-Llopis (2022). The principal novelty is that land in the economy is split into several types with a unique production function associated with each. The most related model is the multi-product farm model of Gordeev and Singh (2023).

The model economy is populated with F heterogeneous farmers that rent inputs and produce the agricultural good.

4.1 Farm Problem

Each farm is indexed with f. The farm produces quantity y_f of the agricultural good, sells it at market price p, and seeks to maximize the resulting profits.

The farm chooses the quantity of each input to hire: land l_f, labor n_f, and intermediate inputs m_f. The market rental rates are r_l, r_n, r_m.

The farm chooses one land type c out of 5 available types and rents land of that type. We abstract from farms combining plots across different land types for simplicity: 89% of farms operate plots that all belong to the same land type cluster.

The farm uses a Cobb-Douglass production function $y_f = z_f l_f^{\alpha_c} n_f^{\beta_c} m_f^{\gamma_c}$. Its input elasticities α_c (land), β_c (labor), and γ_c (intermediates) depend on the land type c chosen by the farm. The total factor productivity z_f is farm-specific.

Market frictions and distortionary government policies are represented in the model with distortion terms that vary at the input-farm level: $\tau_{f,l}$, $\tau_{f,n}$, $\tau_{f,m}$. These act as a tax or a subsidy on the cost of each input and have been used in the misallocation literature to flexibly capture a broad array of factors distorting the allocation of inputs between firms or farms. While only the input costs have explicit friction terms, this setup can also capture output distortions: a revenue tax/subsidy is equivalent to a common component in all three input distortions.

The complete problem of the farm is:

$$\max_{c}\left\{\max_{l_f,n_f,m_f}\left\{pz_f l_f^{\alpha_c} n_f^{\beta_c} m_f^{\gamma_c} - r_l\tau_{f,l}l_f - r_n\tau_{f,n}n_f - r_m\tau_{f,m}m_f\right\}\right\} \tag{2}$$

4.2 General Equilibrium

A representative consumer of the agricultural good purchases the outputs of individual farms. The consumer rents its endowments of labor N, intermediate inputs M, and type-specific land endowments $\{L_c\}_c$ to them. The consumer owns all farms and receives their profits as dividends Π.

The consumer's problem is:

$$\max\{\log C\} \tag{3}$$

s.t.

$$pC = r_l L + r_n N + r_m M + \Pi \tag{4}$$

where

$$\Pi = \sum_f \left[py_f - r_l l_f - r_n n_f - r_m m_f\right] \tag{5}$$

All input and output markets clear. Note that the endowment of each land type is fixed: the market for land of each type must clear individually. Denote the set of farms that choose type c with F_c.

$$C = \sum_f y_f \tag{6}$$

$$\sum_{f\in F_c} l_f = L_c \quad \forall c \tag{7}$$

$$\sum_f n_f = N \tag{8}$$

$$\sum_f m_f = M \tag{9}$$

4.3 Extracting Distortions

In the class of models of misallocation following Hsieh and Klenow (2009), to which our model belongs, the fundamental distortion terms can be extracted for any farm in the data from its observed marginal products:

$$r_l \tau_{f,l} = \alpha_c \frac{py_f}{l_f} = mrpl_f \tag{10}$$

$$r_n \tau_{f,n} = \beta_c \frac{py_f}{n_f} = mrpn_f \tag{11}$$

$$r_m \tau_{f,m} = \gamma_c \frac{py_f}{m_f} = mrpm_f \tag{12}$$

The way the total cost of each input is split into its market price r_x and the distortion term $\tau_{f,x}$ does not matter for the farm's choices. Only the dispersion in these costs will affect the aggregate allocation: and that dispersion is determined purely by the heterogeneity in $\tau_{f,x}$.

The farm-specific productivity z_f is implied by the assumed production function and is directly observable due to the availability of physical input and output measures in the data we use:

$$z_f = \frac{y_f}{l_f^{\alpha_c} n_f^{\beta_c} m_f^{\gamma_c}} \tag{13}$$

Extracting the distortions and physical productivity from the data in this way allows the model to represent every observed farm with its model equivalent, reproducing all observed heterogeneity in input and output choices between farms.

4.4 Quantifying Misallocation

By limiting the ability of markets to allocate inputs to the most productive farms, τ distortions depress the aggregate productivity of the agricultural sector. Quantifying the aggregate cost of misallocation boils down to conducting a counterfactual reallocation exercise in which τ distortions are equalized between farms and the counterfactual output is compared to the currently observed one.

Because the appropriate production function is different between land types, observed input choice heterogeneity between farmers operating different land types may be optimal. Failing to account for production function heterogeneity would lead conventional models of misallocation to overstate the magnitude of distortions in the economy and their aggregate cost. To quantify the importance of this effect, we conduct two sets of counterfactual reallocation exercises. First, we compare the conventional model in which all farms use the same land type with the same “aggregate” production function to the heterogeneous production model in which farms use type-specific production functions but cannot move from the currently observed land type. This comparison highlights the way that misallocative frictions can be overstated when technology heterogeneity is not accounted for. Second, we allow for endogenous land type selection in the model and repeat the exercise. This comparison highlights the way that farmers may be able to move to a different land type once market frictions are eased.

4.5 Quantitative Results

Coming Soon!

5 Conclusion

Estimates of misallocation rely on mapping the observed heterogeneity in input choices between producers to unobserved heterogeneity in fundamental frictions. This mapping relies on the optimal production technology being the same for all producers. We explore a particular source of potential heterogeneity in chosen production technologies: the intrinsic characteristics of land operated by Indian farmers. We group farm plots into land types using a clustering algorithm and find that different land types have significantly heterogeneous production functions. We quantify the effect of this heterogeneity on estimated misallocation in India’s agricultural sector.

Our study focuses on a single plausibly exogenous driver of heterogeneity in optimal input choices. Thus, it can only place a lower bound on the effect that heterogeneous production technologies can have on existing estimates of misallocation that assume a homogeneous production function. Further work could generalize these findings by integrating the general methods of identifying heterogeneous production function elasticities (like Li and Sasaki (2017) and Kasahara, Schrimpf, and Suzuki (2023)) into the studies of misallocation of inputs.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Chen, Chaoran, Diego Restuccia, and Raul Santaeulalia-Llopis. 2022. *Land Misallocation and Productivity.* Working paper w23128. Cambridge, MA: National Bureau of Economic Research, March.

Eberhardt, Markus, and Francis Teal. 2020. "The Magnitude of the Task Ahead: Macro Implications of Heterogeneous Technology." *Review of Income and Wealth* 66 (2): 334–360.

Gollin, Douglas, and Christopher Udry. 2021. "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture." *Journal of Political Economy* 129 (1).

Gordeev, Stepan, and Sudhir Singh. 2023. *Misallocation and Product Choice.* Working paper.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. "Cluster Analysis." In *Data Mining,* 443–495. Elsevier.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *The Quarterly Journal of Economics* 124 (4): 1403–1448.

Kasahara, Hiroyuki, Paul Schrimpf, and Michio Suzuki. 2023. "Identification and Estimation of Production Function with Unobserved Heterogeneity." Preprint, May 19, 2023. Accessed October 5, 2023. arXiv: 2305.12067 [econ]. http://arxiv.org/abs/2305.12067.

Li, Tong, and Yuya Sasaki. 2017. "Constructive Identification of Heterogeneous Elasticities in the Cobb-Douglas Production Function." Preprint, November 27, 2017. Accessed September 22, 2022. arXiv: 1711.10031 [econ]. http://arxiv.org/abs/1711.10031.

Mundlak, Yair, Rita Butzer, and Donald F. Larson. 2012. "Heterogeneous Technology and Panel Data: The Case of the Agricultural Production Function." *Journal of Development Economics* 99 (1): 139–149.

Restuccia, Diego, and Richard Rogerson. 2008. “Policy Distortions and Aggregate Productivity with Heterogeneous Establishments.” *Review of Economic Dynamics* 11 (4): 707–720.

———. 2017. “The Causes and Costs of Misallocation.” *The Journal of Economic Perspectives* 31 (3): 151–174.

Zhang, Yiqun, and Yiu-ming Cheung. 2022. “Learnable Weighting of Intra-attribute Distances for Categorical Data Clustering with Nominal and Ordinal Attributes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (7): 3560–3576.

Appendix

A Additional Figures and Tables

Table A.1: Production function estimates by land type

	aggregate	I	II	III	IV	V
land	0.427	0.396	0.388	0.300	0.625	0.448
	(0.031)	(0.039)	(0.040)	(0.046)	(0.045)	(0.040)
labor	0.189	0.126	0.397	0.433	0.159	0.218
	(0.031)	(0.053)	(0.053)	(0.070)	(0.057)	(0.058)
Int. Inputs	0.351	0.398	0.225	0.403	0.203	0.358
	(0.041)	(0.057)	(0.039)	(0.043)	(0.040)	(0.043)
Observations	14,705	2,605	1,608	2,710	4,649	3,133
R-squared	0.624	0.649	0.684	0.595	0.600	0.664
Season FEs	Y	Y	Y	Y	Y	Y
Village FEs	Y	Y	Y	Y	Y	Y

Note. The table presents the estimation of Equation 1 using 2SLS and instruments described in Section 3.1. The “aggregate” column pools all agricultural plots together. Columns I-V restrict the sample to each land type cluster.