Misallocation and Product Choice

Stepan Gordeev[*] Sudhir Singh[†]

August 9, 2023

click for latest version

Abstract

We study the costs of misallocation of inputs between multi-product firms that endogenously choose among heterogeneous products. Misallocation of inputs between firms has been shown to be a significant drag on aggregate productivity, especially in the agricultural sectors of low-income economies. Existing estimates of its costs have relied on models of single-product firms using a single aggregate production function. Using rich farm-crop-level data from India, we estimate product-level production functions and find that they are meaningfully different from one another and from the aggregate one. We build a general equilibrium model of firm-level misallocation in which multi-product firms are able to choose the set and mix of heterogeneous products. Misinterpreting product heterogeneity as evidence of distortions and missing the endogenous product choice response to real distortions biases single-product models to overstate misallocation, while ignoring returns-to-scale heterogeneity and within-firm productivity dispersion biases them to understate it. On net, the single-product model understates the cost of misallocation for India's aggregate agricultural productivity by 28% in our benchmark calibration.

Keywords: misallocation, multi-product firms, agriculture, productivity, distortions

JEL classification: O4, O11, O13, Q1

[*]University of Connecticut, stepan.gordeev@uconn.edu
[†]Michigan State University, singhsud@msu.edu

1 Introduction

Misallocation of production inputs between firms is an important driver of cross-country income differences (Hsieh and Klenow 2009; Restuccia and Rogerson 2017). Various market frictions and distortionary government policies prevent inputs from being allocated to their most productive uses: some productive firms would prefer to hire more labor, capital, or land, while some less productive ones would be happy to supply these inputs. However, both parties find themselves unable to do so due to imperfect markets or institutional obstacles, limiting the aggregate productivity of the entire economy. Misallocation is especially severe within the agricultural sectors of low-income countries: allowing farmers to efficiently trade land can offer an almost three-fold increase in the sector's productivity (Chen, Restuccia, and Santaeulalia-Llopis 2022). Some of the estimated aggregate cost of misallocation can be driven by benign causes like mismeasurement, input adjustment costs, late productivity shocks, and others (Bils, Klenow, and Ruane 2021; Gollin and Udry 2021; Asker, Collard-Wexler, and De Loecker 2014). At the same time, a large portion of this cost can reflect malign causes: poor land property rights enforcement, inefficient land distribution institutions, distortionary government policies, poor market access, and more (Gottlieb and Grobovšek 2019; Chen 2017; Krishnaswamy 2018; Le 2020; Morando 2023). Understanding both the costs and the causes of misallocation is crucial for understanding the obstacles to aggregate development, especially in agriculture, where productivity differences between nations are particularly stark (Gollin, Lagakos, and Waugh 2014).

The misallocation literature has focused on studying single-product firms. The literature on agricultural misallocation has further modeled all firms as producing the same homogeneous agricultural good. Farms in the data, however, have wildly different selection of crops: we find that even the most commonly grown crop in India is produced by only one quarter of all farmers. Moreover, two thirds of Indian farmers grow two or more different products within a year.

In this paper, we study the aggregate costs of misallocation of inputs between multi-product firms that choose among heterogenous products. When products have different production functions, they will be exposed differentially to misallocative market distortions. This causes conventional single-product models to misperceive product heterogeneity as evidence of misallocative frictions, to understate productivity heterogeneity, and to mis-predict firm responses to a counterfactual efficient reallocation of inputs along several dimensions.

We explore these mechanisms in rich farm-crop-level data from India. We document significantly heterogeneous product choice and frequent multi-product behavior among Indian farmers. Next, we estimate crop-specific production functions using instruments derived from shocks to other plots within a farm to address the simultaneity bias. We find that many crops have significantly different input elasticities from one another—and from the traditional aggregate agricultural good production function—suggesting that products in the data are meaningfully heterogeneous for the purpose of quantifying misallocation.

We build a general equilibrium model of heterogeneous firms facing misallocative distortions. The main novelty of the model is that firms (or farms) choose among multiple products (or crops) with heterogenous production functions and can elect to produce any number of products at once. The model provides a mapping from the observable farm-crop-level input and output choices to the unobservable fundamental distortions that are the model's stand-in for all kinds of institutional and market frictions. Once these distortions are recovered from the data, the model provides a quantification of the aggregate cost of misallocation induced by the frictions.

The multi-product model highlights four mechanisms by which conventional single-product models misestimate the cost of misallocation when the data is generated by firms producing heterogeneous products. First, single-product models erroneously perceive heterogeneous product choice as evidence of underlying frictions even when the heterogeneity is driven purely by differential productivity draws: this leads single-product models to *overstate* the cost of misallocation in the data. Second, single-product models ignore the returns-to-scale heterogeneity between products, missing that, if frictions are removed, some firms may be able to scale production up more easily than others: this leads them to *understate* the cost of misallocation. Third, single-product models ignore productivity heterogeneity between products within a single firm, understating the underlying productivity dispersion: this leads them to *understate* the cost of misallocation. Finally, single-product models ignore firms' endogenous product choice response to the frictions they face. In the presence of misallocative distortions and a choice between products with heterogeneous production functions, model firms are able to partly mitigate the effects of frictions they are facing by shifting their product choice away from goods intensive in the relatively distorted input toward goods intensive in the relatively undistorted one. This endogenous response reduces the aggregate cost of misallocation for given fundamental distortions: because single-product models are blind to this response, they *overstate* the cost of misallocation.

We map the firm-product-level model to Indian farm-crop-level data. We conduct quantitative reallocation exercises in both the multi-product and the single-product models to evaluate the net effect of the four mechanisms on the aggregate misallocation within India's agricultural sector. In our preferred calibration, the single-product model *understates* the cost of misallocation of inputs across farms by 28%: accounting for product heterogeneity and multi-product behavior, an efficient reallocation of land, labor, and intermediate inputs across farms promises an almost four-fold increase in agricultural productivity, compared to the three-fold increase estimated by the single-product model. This effect is mainly driven by the returns-to-scale channel, which dominates the channels that push the single-product model in the direction of overstatement. However, the sign and magnitude of the single-product model's net error depends on the exercise being conducted. Only reallocating land between farmers would increase India's agricultural productivity by 30% according to the multi-product model: the single-product model *overstates* this figure by one-fifth.

Estimating the cost of agricultural misallocation in farm-level data is the subject of a growing literature. The foundational result on the severity of misallocation in the agriculture of low-income countries was established by Chen, Restuccia, and Santaeulalia-Llopis (2022). Some papers focus on identifying a particular source of distortions in the data and then using models to quantify the aggregate cost of those distortions (Chen 2017; Gottlieb and Grobovšek 2019; Adamopoulos et al. 2022). Others seek to capture all sorts of market frictions and government distortions generally, and make methodological contributions to estimating their cost (Gollin and Udry 2021; Aragón, Restuccia, and Rud 2022). Our paper contributes to the literature by departing from its assumption of a single homogeneous agricultural product. We estimate crop-specific production functions and build a model of multi-product farms endogenously choosing among those crops. We thus contribute both to the question of measuring distortions in the data and to the question of quantifying the aggregate cost of these distortions.

The effects of government policy or technological changes on country-level crop allocation have been explored before. Blanco and Raurich (2022) show that falling capital costs cause the aggregate allocation to shift toward capital-intensive crops. Le (2020) finds that land use policies in Vietnam generate aggregate misallocation by favoring rice over other crops. Krishnaswamy (2018) documents that Indian price support for staple crops distorts the market in favor of rice and wheat. We contribute to this literature by exploring the interplay between distortions and product choice not at the aggregate level, but at the level of individual farmers. The model we devise then allows us to map the micro-level distortions and farm responses to aggregate productivity: this mapping would apply to any individual friction the literature has studied before.

Multi-product firms have been discussed in the manufacturing misallocation literature by Jaef (2018) and Wang and Yang (2023). In both papers, firms face distortions and can choose the number of products to manufacture among varieties that have the same production function but different productivities. Both papers then calibrate their models to match cross-sectional distribution moments in the data. First, we contribute to this literature by exploring product choice between products not just with different productivities but also with different input elasticities: it's the latter that ends up driving the differential predictions between the single- and the multi-product model in our setup. Second, we contribute by extending Hsieh and Klenow (2009)'s framework of mapping firm-level observables to model-implied firm-level fundamentals to the case of multi-product firms. This allows us to reproduce each observed firm (or farm) with its model analog exactly, without having to calibrate distributions of firms to aggregate moments.

Our contribution on the interaction between misallocation and endogenous product choice is not conceptually restricted to agriculture. However, there are several reasons why agriculture is the perfect setting to begin studying this interaction. First, firm-product-level (or farm-crop-level) inputs and outputs are far more feasible to measure in agriculture than in other sectors. Farmers are better able to estimate the amount of land, labor, and intermediate inputs applied to each crop. Such data is also more commonly collected: we use India's REDS, but multiple other farm-level datasets contain crop-level information. Second, products (crops) are far more homogeneous in agriculture than in other sectors. Outside of very specific sub-sectors, varieties produced by different manufacturers operating in the same sub-sector can be vastly different. In agriculture, however, specimens of the same crop grown on different farms can be reasonably approximated as homogeneous. This facilitates the task of estimating product-level production functions and exploring systematic differences in input intensities between products. For these reasons, we choose to focus our initial exploration of endogenous product choice on agriculture. But the mechanism applies more broadly, and insights gleaned from the agricultural setting of a developing country will apply also to the manufacturing sectors of more developed countries.

The paper proceeds as follows. Section 2 discusses the institutional context of Indian agriculture. Section 3 describes the farm-level data we use. Section 4 documents the heterogeneous product choice among Indian farmers. Section 5 estimates crop-level production functions. Section 6 describes the multi-product firm model of misallocation. Section 7 quantifies misallocation and compares it across models. Section 8 concludes.

2 Institutional Context

Indian agriculture is very diverse and one of the largest in the world in terms of arable lands.[1] Around half of Indian agriculture is rainfed, although the share of the irrigated area has improved over the years (GoI, 2022). The agricultural sector employs 58 percent of households and is a major source of income and livelihood for the majority of rural households. It is worth noting that women make up 30 percent of total cultivators and 41 percent of agriculture laborers. However, women only manage and operate 14 percent of the total operational holding (GoI, 2021).

Land market and tenancy structure. The Indian agricultural system was a zamindari system during the colonial era, where lands were in the hands of a few who had the right to lease out and collect revenue (Banerjee and Iyer 2005). However, after 1947, the government of India brought several land and tenancy-

1. India possesses more than 150 million of hectares of arable land and comes second after the USA.

related reforms that changed the structure of ownership and leasing rights. These reforms varied in scale, timing, and implementation across Indian states (Deininger, Jin, and Nagarajan 2008).

The majority of agricultural land in India is privately owned by households. Smallholder farming is widespread. In 1999-2000, the average farm size was 5.3 acres per household, which declined to 4.4 acres per household in 2007-08. The size of land holding further declined to 2.6 acres per household in 2015-16. This decrease in land holding size can be attributed in part to growth in the number of households and family division (Andrew D Foster and Mark R Rosenzweig 2002). Land is seldom sold and bought: more than 94 percent of the total land is inherited and passed on to the next generation. Only 12 percent of agricultural cultivators participate in the land rental market by renting land for cultivation either in or out.[2] Low participation in the land market hinders productivity as it limits the reallocation of land to productive farmers (Deininger et al. 2003; Bolhuis, Rachapalli, and Restuccia 2021).

Agricultural seasons. There are three major agricultural seasons in India, namely Kharif (monsoon), Rabi (spring/winter), and Zaid (summer/dry). The Kharif season typically spans from May/June to early October, during which crops like paddy (rice), maize, sugarcane, jute, cotton, groundnut, turmeric, and others are cultivated. The Rabi season takes place during winter, with crops being sown between October and December and harvested from March to April. Major crops of the Rabi season are wheat, barley, peas, gram, and mustard. Finally, the Zaid season, the shortest of the three, occurs from March/April to May/June; fruits and vegetables are grown during this season.

3 Data

Our primary data source is the Rural Economics and Demographic Survey (REDS), which was conducted by India's National Council of Applied Economic Research in 1971, 1981, 1999, and 2007-08.[3] The first round was collected in 1971. On each subsequent round, all households (or their descendents) living in the same vil-

2. Source: Rural Economic and Demographic Survey (REDS) 2007-08.
3. Prior to 1981-82, the survey was known as the Additional Rural Income Survey (ARIS).

lage were surveyed again, along with a random selection of new replacement households (Vashishtha 1989).[4] The survey is nationally representative of rural India.

For this analysis, we use REDS 2007-08 (hence REDS07), covering a sample of 8,659 households across 242 villages in 17 major states in 2007/08. REDS07 collects detailed information on agricultural inputs and outputs at the level of plot-season-crops. We restrict our attention to this round because it is the only one in which agricultural data was collected at plot level, which is necessary for our production function identification strategy. Household and village-level modules of the survey were administered during 2007-08. Of the surveyed households, 4,803 cultivated land: we treat each such household as a farm. In total, these farmers cultivated 10,318 plots.

4 Heterogeneous Product Choice by Indian Farms

Crop Choice is Heterogeneous. Indian farmers vary greatly in which crops they choose to produce. Figure 1 shows the 10 most commonly produced crops in India. Rice is the most popular crop—yet it is grown only by 26% of farmers. Wheat is a close second, but all other crops are far behind: Indian farmers have little overlap in crops they choose to grow.[5]

Many Farms Grow Multiple Products. Figure 2 displays the share of output produced by farms that grow different numbers of crops. The top bar counts crops that were grown by each farm within the survey year, even if the crops were grown in different seasons: over 80% of output is produced by farmers that grow multiple crops within a year. The bottom bar of Figure 2 treats farm-season combinations as separate entities, thus only counting crops that were all grown in the same season: two-thirds of output is produced by farmers that grow nothing else in the same season, and one-third is produced by multi-crop farmers. Farmers are more specialized within each season than across, although the extent of

4. In 1971, the sampling frame included one district per state in the Intensive Agricultural District Program (IADP) and a random sample of other districts.

5. Appendix Figure A.1 shows that the ranking of crops by their share in aggregate output has a similar shape, although the order of crops does change.

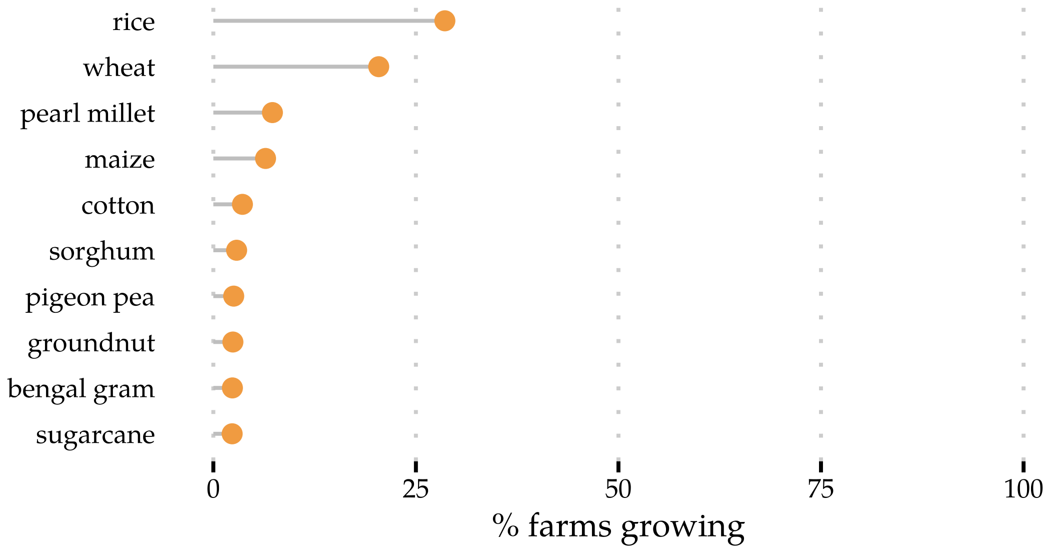


FIGURE 1: Top 10 crops by % of farm-seasons growing them

multi-crop behavior is significant in both cases.

These results have two potential implications for modeling and quantifying misallocation. First, they suggest that production should be split by season when quantifying misallocation, since crops grown in different seasons may be meaningfully different and would affect misallocation estimates if this heterogeneity isn't taken into account. Second, the magnitude of multi-crop production is non-negligible both within and across seasons, suggesting that not only modeling heterogeneous crop choice but also the simultaneous cultivation of multiple crops may be important.

At the same time, the apparent heterogeneity of product choice among Indian farmers may be of little consequence for estimating misallocation if different crops have similar production functions and would thus respond to misallocative frictions the same way. In the next section, we estimate crop-specific production functions to test whether this is the case.

5 ESTIMATING PRODUCT-SPECIFIC PRODUCTION FUNCTIONS

The preceding section documented that Indian farmers engage in heterogeneous multi-product behavior. In this section, we investigate whether these products

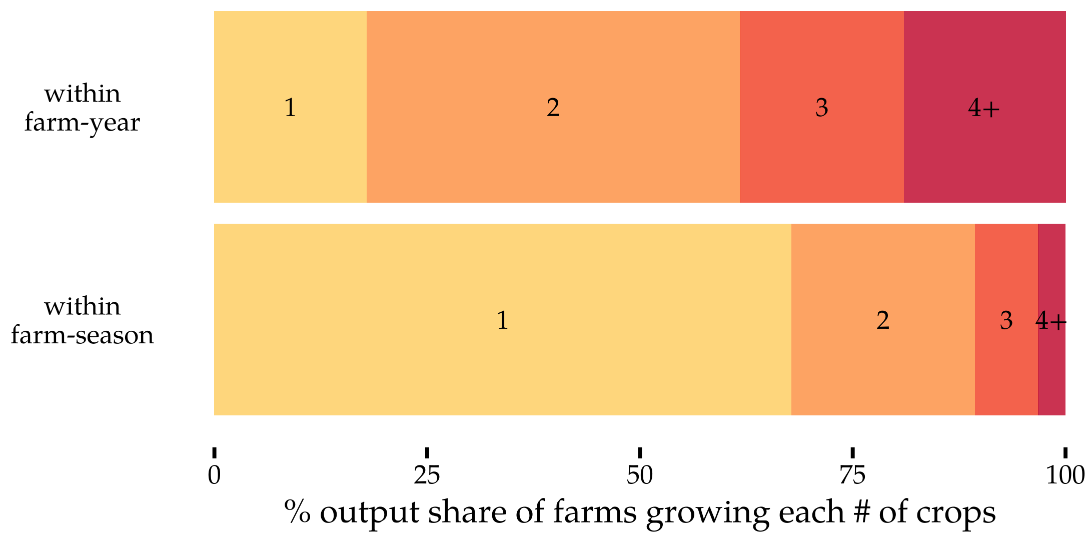


FIGURE 2: Share of output produced by farms growing each number of crops (within a year or within each season alone)

are meaningfully different in their production functions.

5.1 Specification

To estimate input elasticities, we estimate the following crop-specific Cobb-Douglas production function for each farm.

$$y_{f,i,t} = z_{f,i,t} l_{f,i,t}^{\gamma_i} x_{labor,f,i,t}^{\alpha_{labor,i}} x_{inter,f,i,t}^{\alpha_{inter,i}}$$

$y_{f,i,t}$ is the gross physical output of farm f for crop i in season t, weighted with market prices (this weighting is irrelevant at the estimation stage).[6] l is land input with crop-specific elasticity γ_i. The land input is adjusted for quality, as described in Section 5.3 below. x_{labor} is labor input with elasticity $\alpha_{labor,i}$, measured in the number of days worked, including both hired and family labor. x_{inter} is intermediate input with elasticity $\alpha_{inter,i}$, measured as total expenditure on seeds, fertilizer, irrigation, rented machinery and draft animals. $z_{f,i,t}$ is the farm-specific

6. Although value added production functions are common in the misallocation literature, they have a poor theoretical basis and can significantly overstate the productivity dispersion (Gandhi, Navarro, and Rivers 2017). For this reason, we elect to work with a gross output production function.

productivity measure that is not directly observed. After taking logs, we estimate the following regression specification:

$$\log y_{j,i,t} = \gamma_i \log l_{j,i,t}^{\gamma_i} + \alpha_{labor,i} \log x_{labor,j,i,t} + \alpha_{inter,i} \log x_{inter,j,i,t} + \epsilon_{j,i,t} \tag{1}$$

for each crop i on plot j. In our data, we observe more than 50 crops grown all across India. For ease of estimation, we group them into five categories: rice; wheat; other cereals; pulses; vegetables, fruits, and oil seeds. In all that follows, crop-specific production functions refer to these five merged crops.

5.2 Identification

Estimating the production function using OLS as specified in equation (1) is not consistent since plot-specific unobserved productivity is likely to be correlated with observed input choices. Hence, our estimation strategy follows the idea developed in Gollin and Udry (2021). We estimate Equation 1 using the two-stage least square method (2SLS), where instruments are developed following the shadow price approach in Singh, Squire, and Strauss (1986) and Gollin and Udry (2021). The idea is simple: within a farm, shocks to plots $k \neq j$ change the shadow price of inputs on plot j, providing variation in plot-level input prices that can be used as an instrumental variable for input allocations. The identification of the production function rests on two assumptions: first, that observed shocks to farm f's plot k affect the shadow price of inputs on f's plot j; second, that observed shocks affecting input demand on plot k are not correlated with unobserved shocks affecting input demand on plot j, conditional on observed shocks to j. The survey collecs information on a wide array of agricultural, health, and social shocks at the household level. Gollin and Udry (2021) posit that these shocks will have different effects on individual plots depending on the physical features of each plot. Therefore, the household-level (or farm-level) shocks interacted with observed objective plot-level soil features provide the necessary plot-level variation, which serve as instruments in the estimation of the production function.

5.3 Land Input Measure

There are two commonly used ways of measuring the quantity of land input in agricultural survey data similar to India's REDS. The first is to measure it with the physical area of cultivated land. This quantity can be measured very precisely and objectively, but lacks any information on the quality of land. The second way is to measure it with the reported market price of cultivated land, which should capture quality differences between plots. However, land markets in India (and many other developing countries) are extremely under-developed, and so market price reported by farmers is likely to be exceptionally noisy and mismeasured: in the absence of a market, farmers' ability (and need) to estimate the market price of their land is likely to be poor.

Instead of following either of these two approaches, we settle on a compromise solution that attempts to combine the best of both. We measure the land input on each plot with its physical area weighted by a quality index: price per acre predicted by a statistical model using only the observed objective characteristics of the plot as predictor variables.

The objective characteristics measured in the data and included as predictors are various qualities of soil (its type, color, salinity, depth, percolation, and ease of drainage) and objective measures of irrigation access (presence and features of nearby irrigation wells and canals).

The task of constructing the land quality is effectively a regression. Instead of an ordinary least squares regression (OLS), however, we employ random forests (RF): a supervised machine learning algorithm developed by Breiman (2001). The random forest model "grows" a large number of decision trees, with each tree using a random bootstrap sample of the data, and aggregates the predictions of individual trees.[7] Unlike OLS, RF does not require the econometrician to pre-specify the interactions and non-linearities expected in the data: it searches for them on the fly and adapts to what it finds in the data. Furthermore, random forests tend to outperform OLS and many other machine learning techniques in out-of-sample predictive performance, reducing the danger of overfitting the noise in reported market price (Varian 2014; Mullainathan and Spiess 2017).

7. See Biau and Scornet (2016), Schonlau and Zou (2020), or Ziegler and König (2014) for excellent reviews.

To evaluate both methods, we estimate an RF regression and an OLS regression using a randomly selected training sample from our data (2/3 of observations) with the remaining 1/3 reserved for the test sample. Both methods regress log price per acre on observed physical characteristics of the plot. The OLS regression includes a complete set of two-way interactions. The RF's parameters m_{try} and *nodesize* were tuned with k-fold cross-validation (with $k = 3$), although tuning provides only a small improvement in predictive performance relative to conventional defaults.[8]

Once the two methods have been estimated, we evaluate their performance on the reserved testing sample: results are presented in Table 1. The random forest has both a lower mean squared error and a higher R^2 than OLS . Due to RF's better out-of-sample performance, we elect to use it to construct the land quality index: predicted price per acre computed for the whole sample.

Table 1: Out-of-sample performance of OLS and RF in predicting land price

	Ordinary Least Squares	Random Forest
Mean Squared Error	0.61	0.49
R^2	0.39	0.51

Note. The table presents two metrics of out-of-sample performance of Ordinary Least Squares (OLS) and Random Forest (RF) regressions in predicting log price per acre of a plot with observed physical characteristics of the plot. Two-thirds of plot-level observations were used for estimation and one-third was reserved as a test sample to compute the presented out-of-sample performance metrics.

Because the random forest captures over half of observed variation in reported price using nothing but observed objective plot features, it likely captures much of the underlying variation in land quality. Because we used predicted prices and took several steps to minimize overfitting (bootstrapping within RF, k-fold

8. Each tree within a random forest is constructed by repeatedly finding the predictor variable and its cutoff value to split the data on into two branches that maximizes the goodness of fit of the branching. m_{try} sets the number of candidate predictors randomly drawn to be considered for branching at each node when constructing each tree. *nodesize* sets the cutoff number of observations per node at which further splitting stops. The third free parameter in a random forest is the number of trees: we set it high enough that further increases do not improve predictive performance. See any of the RF reviews cited above for details on their operation and the role of these parameters. See Bischl et al. (2021) for an overview of k-fold cross-validation and other tuning techniques.

cross-validation, and train/test data split), this captured variation is driven by population-level relationships between plot characteristics and market price, rather than the noise in any individual farmer's measurement.

5.4 Production Functions Are Heterogeneous Across Products

Complete results from production function estimation are shown in Appendix Table A.1. Figure 3 visually summarizes the estimated production function elasticities, including 95% confidence intervals. The first panel displays the input elasticities when a single production function is estimated on all crops pooled together: the commonly used approach. In this case, the results are in line with the literature's findings: land elasticity γ is the highest at 0.43, intermediates elasticity α_{inter} is slightly lower at 0.35, labor elasticity α_{labor} is 0.19. The sum of the three elasticities is 0.97, insignificantly different from 1, indicating that the production function exhibits slightly decreasing or roughly constant returns to scale.

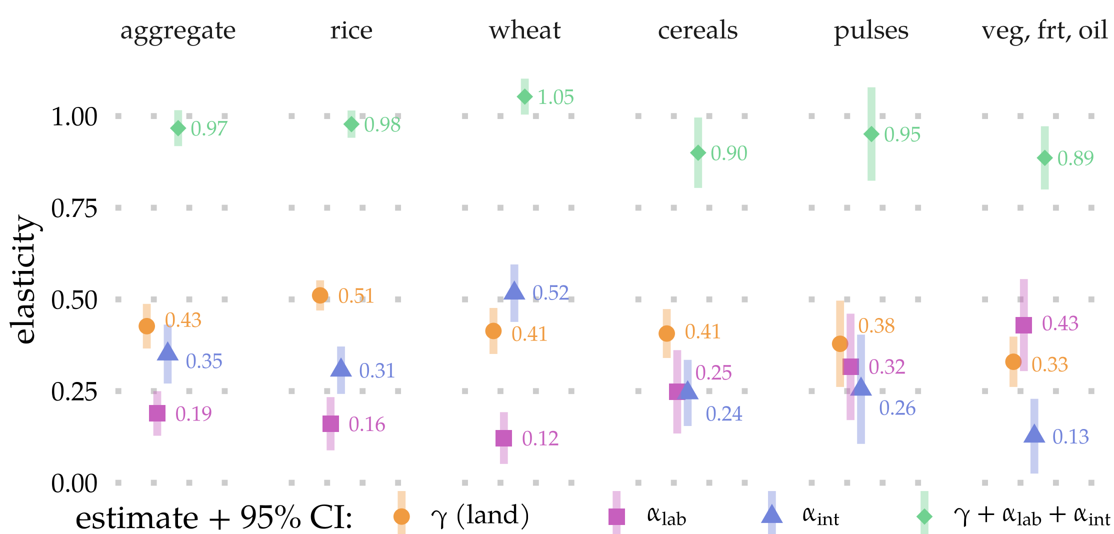


Figure 3: Estimated production function input elasticities

As the following panels indicate, however, crop-specific production functions are heterogeneous across crops. Rice is significanly more land-intensive than the aggregate production function would indicate. Wheat is more intermediates-intensive than rice or the aggregate function, and its returns to scale are increasing. Cereals other than wheat are more in line with the aggregate elasticities, but have significantly decreasing returns to scale. For pulses, all three inputs are similarly important. Vegetables, fruits, and oils are more labor-intensive than any other crop, and exhibit decreasing returns to scale.

Pairwise comparisons of a given input's elasticity across crops indicate that many pairs of crops have significantly different elasticities from one another as well as from the aggregate estimates. Production functions are heterogeneous across products, and a single aggregate production function provides only a poor approximation of any one individual crop. To quantify the relevance of this heterogeneity for quantifying the cost of misallocation, we turn to the model.

6 Model

The model pursues three objectives. The first is to provide a framework for thinking about the production decisions of multi-product farmers in the presence of misallocative distortions. The second is to provide a mapping from observable farm-level input and output choices to unobserved distortions these farms must be facing. The third objective—and the ultimate goal of the paper—is to quantify the aggregate cost of misallocation induced by distortions and to tease out the role of product choice in explaining this cost.

To achieve these objectives, we build a model of farm-level misallocation in which multi-product farmers make production decisions over heterogeneous products. It builds upon the single-product firm model of Hsieh and Klenow (2009). Because their model was applied to data on manufacturing firms, it assumed monopolistic competition and a constant returns to scale production function. Because agricultural markets are populated with small producers growing fairly homogeneous crops, we assume perfect competition instead, in line with the simpler single-product agricultural misallocation framework of Chen, Restuccia, and Santaeulalia-Llopis (2022). Instead of assuming constant or decreasing returns to scale for all crops, we rely on our estimated returns to scale for each crop.

6.1 Farm Problem

The economy is populated by F heterogeneous farmers making production decisions over N heterogeneous crops.

Consider the problem of a profit-maximizing farm f. The farm produces quantity $y_{f,i}$ of each crop i and sells it at market price p_i. The farm has idiosyncratic productivity $z_{f,i}$ in each crop. It chooses how much of each input to hire or allocate to each crop.

Inputs. Inputs $g = 1, \dots, G$ are flexible: the farm can hire each at market rate r_g. The flexible input elasticity $\alpha_{g,i}$ is input-crop-specific. In our calibration, there are two flexible inputs ($G = 2$): labor and intermediates. The farm chooses quantity $x_{f,g,i}$ of each input g to allocate to each crop i.

Land input l is available to the farm in fixed supply L_f: it needs to decide how to split this endowment between available crops, assigning $l_{f,i}$ to each. The crop-specific land elasticities are γ_i. There are three reasons for modeling land as a fixed input. The first is that it is a reasonable approximation of the state of the land market in India. Land is rarely purchased, sold, or rented: it is usually simply inherited. The second is that having one input be in fixed supply generates interdependent crop production: a change in parameters of crop i will lead to a change in the shadow cost of the fixed input (land) and thereby impact the input and output decisions the farm makes on crops $j \neq i$.[9] The third is that an input in fixed supply justifies the production function identification assumptions we made in Section 5: land l and its shadow cost λ_f provides the link through which observable shocks to one plot can affect the shadow cost of inputs on other plots.

Inputs are combined with Cobb-Douglas production technology to produce each crop i: $y_{f,i} = z_{f,i} l_{f,i}^{\gamma_i} \Pi_g \left(x_{f,g,i}^{\alpha_{g,i}} \right)$.

Distortions. Misallocative frictions in this class of models are captured by distortion terms $\tau_{f,g}$. These are idiosyncratic to the farm and operate as taxes or subsidies. If farm f's $\tau_{f,g}$ is higher for input g than for input h, the farm is facing stronger frictions in input g's market, and will hire a relatively higher quantity of h

9. The importance of modeling interdependent crop production with a fixed input rather than any alternative mechanisms is discussed in Just, Zilberman, and Hochman (1983) and Shumway, Pope, and Nash (1984).

instead. If farm f's frictions $\tau_{f,g}$ for all inputs are higher than those of farm e, then farm f will hire less of every input and will operate at an inefficiently small scale compared to farm e. This simple framework has been used in the misallocation literature to capture all kinds of market frictions and government distortions that firms may be suffering from. Any variation in input ratios or firm scales that is unexplained by the observed production functions and productivities can be replicated by these distortion terms.

Furthermore, once we add multi-product firms into the model, we have to add another layer of distortions. $\tau_{f,g,i}$ and $\tau_{f,l,i}$ are farm-input-crop-specific distortions that can fit unexplained variation in input ratios between crops *of a single farm*.

Finally, while land input l does not have an explicit distortion term like $\tau_{f,l}$ in the model, land is still de facto distorted since its quantity is fixed. Any variation in the shadow cost of land λ_f between farms is effectively driven by implicit τ-like frictions that prevent the farmers from operating an efficient amount of land. As long as land is not distributed in precisely the right way to equalize λ_f between farmers, it is distorted, with the dispersion of λ_f reflecting the severity of distortions.

Additional concavity. When crop-specific production functions have decreasing returns to scale, model farms choose to grow multiple crops at once. However, most of the crop-specific production functions estimated in Section 5.4 exhibit returns to scale that are close to constant, generating little multi-cropping in the model unless frictions are extreme: production concavity alone struggles to explain the extent of multi-product behavior in the data. To allow for a range of other mechanisms that may be incentivizing farmers to grow multiple crops, we raise the crop-level revenue within the farmer's objective function to some power $\eta \leq 1$. When $\eta < 1$, this generates additional concavity in crop-level revenue. Introducing this parameter allows the model to parsimoniously capture mechanisms such as risk (farmers multi-crop to hedge against uncertain revenue yields in each individual crop), diet diversification by subsistence farmers (subsistence farmers grow multiple crops to satisfy their love of variety), and market power (farmers with some market power in their local market face downward sloping demand for each crop, generating revenue that is concave in inputs). Appendix B shows that the latter two mechanisms, if modeled explicitly, generate a farm's problem that is exactly equivalent to the $\eta < 1$ case. Another reason to allow for $\eta < 1$ is that it takes the burden of generating the observed distribution of farm-crop revenues away from τ frictions, leading to more conservative misallocation estimates, as Section 6.4 discusses below.

Fixed cost. For each crop i that the farm chooses to produce ($y_{f,i} > 0$), it pays a fixed cost ω. Without this feature, farmers would produce all possible crops, with some of them grown in negligible quantities: introducing the fixed cost makes such marginal crops unprofitable and leads the farm to drop them entirely. We introduce the fixed cost to reproduce the observed distribution of the number of crops grown per farm, discussed in Section 4.

Problem. The complete problem of the farm is:

$$\max_{\{y_{f,i},l_{f,i}\}_i,\{x_{f,g,i}\}_{g,i}} \sum_{i=1}^{N} \left(p_i y_{f,i}\right)^{\eta} - \sum_{g=1}^{G} r_g \tau_{f,g} \sum_{i=1}^{N} \tau_{f,g,i} x_{f,g,i} - \sum_{i=1}^{N} \omega \cdot 1\left[y_{f,i} > 0\right] \tag{2}$$

s.t.

$$y_{f,i} = z_{f,i} l_{f,i}^{\gamma_i} \Pi_g \left(x_{f,g,i}^{\alpha_{g,i}}\right) \tag{3}$$

$$\sum_{i=1}^{N} l_{f,i} \tau_{f,l,i} = L_f \qquad (\lambda_f) \tag{4}$$

The solution and the computational algorithm are discussed in Appendix C.

6.2 General Equilibrium

Consumer. To close the model, we introduce a representative consumer who has constant elasticity of substitution preferences over N crops with elasticity of substitution σ and crop-specific taste weights φ_i. The consumer purchases quantity C_i of each crop from farms at price p_i. The consumer owns flexible inputs and rents out the endowment X_g^{agg} of each input g to the farms at rental rate r_g. The consumer also owns all farms and receives their profits as dividend Π.

The consumer's problem is:

$$\max_{\{C_i\}_{i=1}^N} \left(\sum_i \varphi_i C_i^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \tag{5}$$

s.t.

$$\sum_i p_i C_i = \sum_g r_g X_g^{agg} + \Pi \tag{6}$$

For conciseness, we adopt the following formulation of the dividend Π:

$$\Pi = \sum_f \left[\sum_{i=1}^N p_i y_{f,i} - \sum_{g=1}^G r_g \sum_{i=1}^N x_{f,g,i} \right]$$

In this formulation, farmers act *as if* they faced distortions $\tau_{f,g}$ and $\tau_{f,g,i}$, fixed cost ω, and additional concavity η (since these show up in farmers' maximization problems), but these are ultimately non-monetary frictions and obstacles and so do not show up in the dividends Π that farmers send to the consumer. An exactly equivalent formulation would be to interpret distortions, fixed costs, and additional concavity as monetary taxes or subsidies administered by the consumer: in this case, they would be included in Π and would likewise show up (with an opposite sign) in the consumer's budget constraint as government revenue or expense [10].

Market Clearing. Crop prices p_i and input prices r_g need to be such that all crops and all inputs markets clear:

$$C_i = \sum_f y_{f,i} \quad \forall i \tag{7}$$

$$\sum_f \sum_i x_{f,g,i} = X_g^{agg} \quad \forall g \tag{8}$$

10. Extending this logic, distortion terms τ can be interpreted as a mixture of non-monetary frictions and monetary government taxes/subsidies. In this case, a fraction (representing the tax/subsidy portion) of each τ would show up in Π and the consumer's budget constraint. The choice between all these interpretations is completely arbitrary as they collapse to identical equilibrium conditions.

Land needs to be distributed between farms in such a way that land used in cultivation equals the total amount of land available:

$$\sum_f \sum_i l_{f,i} = L^{agg} \tag{9}$$

6.3 Mapping Observables to Fundamentals

For the model to be useful for quantifying the cost of misallocation in the data, unobserved distortions τ need to be estimable from observed characteristics. In the class of misallocation models following Hsieh and Klenow (2009), the fundamental distortions faced by each firm can be mapped to marginal revenue products of inputs chosen by the firm. With the assumed production structure, unobserved marginal revenue products can be extracted from observed inputs and outputs.

This result extends to our multi-product model as well, for both flexible inputs g and the land input l:

$$r_g \tau_{f,g} \tau_{f,g,i} = \alpha_{g,i} \eta \frac{\left(p_i y_{f,i}\right)^{\eta}}{x_{f,g,i}} \tag{10}$$

$$\lambda_f \tau_{f,l,i} = \gamma_i \eta \frac{\left(p_i y_{f,i}\right)^{\eta}}{l_{f,i}} \tag{11}$$

While only the total cost of the input can be extracted from the data in this way, the split of the total cost $r_g \tau_{f,g} \tau_{f,g,i}$ into its three components is arbitrary from the point of view of the farm. All that matters for quantifying misallocation is the *dispersion* in these costs, which is driven entirely by the dispersion in distortion τ terms. The average *level* of each distortion is then split from the level of market rental rates r_g by the general equilibrium market clearing conditions. Furthermore, goods taste weights φ_i are picked to ensure that the current allocation is consistent with a general equilibrium: these details are discussed in Appendix C.2.

The productivity of the farm in each grown crop $z_{f,i}$ can be extracted from the data directly using the assumed production function, exploiting the availability of physical input and output quantities in the data:

$$z_{f,i} = \frac{y_{f,i}}{l_{f,i}^{\gamma_i} \Pi_g \left(x_{f,g,i}^{\alpha_{g,i}} \right)} \tag{12}$$

Seasons. As discussed in Section 4, viewing farms' production at the annual level may be deceptive as farmers tend to produce different products in different seasons. Bundling production across seasons may lead to nonsensical counterfactual predictions in the model like the model farm reallocating three times its total amount of land (summed across three seasons) to a single crop that's only ever grown in the most productive of the three seasons. Therefore, we map each model farm f not to an empirically observed farm, but to a farm-season combination. This ensures that different seasons have distinct productivity distributions. Section 7.1 will discuss the treatment of seasons at the aggregate level.

6.4 Calibration

Additional concavity parameter η. Farm-crop revenue can be written as:

$$p_i y_{f,i} = \Bigg(\underbrace{\left(\frac{1}{\lambda_f \tau_{f,l,i}} \right)^{\gamma_i} \Pi_g \left(\frac{1}{\tau_{f,g} \tau_{f,g,i}} \right)^{\alpha_{g,i}}}_{\text{composite distortion, } dist_{f,i}} \Bigg)^{\frac{1}{1-\eta(\sum_g \alpha_{g,i} + \gamma_i)}}$$
$$\cdot \underbrace{\left(p_i z_{f,i} \gamma_i^{\gamma_i} \eta^{\sum_g \alpha_{g,i} + \gamma_i} \Pi \left(\frac{\alpha_{g,i}}{r_g} \right)^{\alpha_{g,i}} \right)^{\frac{1}{1-\eta(\sum_g \alpha_{g,i} + \gamma_i)}}}_{\text{“objective” factors}} \tag{13}$$

where the composite distortion term summarizes the effects of distortions on revenue, and the "objective factors" term captures effects driven by productivity, production parameters, and prices.

As Section 6.3 above discussed, the model can rationalize any observed heterogeneity across and within farms using τ distortions. However, the variance of distortions needed to achieve this (summarized in the variance of the $\log dist_{f,i}$ term) depends on the chosen additional concavity parameter η, as Figure 4 illustrates. When η is low, the degree of concavity in farm-crop revenue is high: in the frictionless optimum, farm-crop revenues are distributed comparatively uniformly, implying a uniform size distribution across farms and significant crop mixing within farms. The current observed allocation is highly dispersed in comparison, and the model rationalizes it with extreme frictions, producing a high variance of $dist_{f,i}$. When η is high, on the other hand, there is little concavity: in the frictionless optimum, farm-crop revenues are dispersed, implying high dispersion in farm sizes and significant crop specialization within farms. The current observed allocation is highly uniform in comparison, and the model rationalizes it with extreme frictions yet again. Somewhere in between is an intermediate value of η that allows the model to reproduce the observed level of dispersion in farm-crop revenues with the most conservative possible distribution of distortions. [11] Upon estimation, the value that minimizes the variance of $\log dist_{f,i}$ in Equation 13 is $\eta = 0.63$, and this is the value we elect for our headline calibration.

Misallocation estimates in general are sensitive to the degree of concavity in the model, whether that concavity comes from monopolistic competition (as in Hsieh and Klenow 2009) or decreasing returns to scale in production (as in Chen, Restuccia, and Santaeulalia-Llopis 2022). By introducing a flexible parameter η, we are able to explicitly explore the dependence of our estimates on this parameter (as sensitivity analysis will do in Section 7.2) and to intentionally pick a conservative estimate for the headline calibration.

Elasticity of substitution σ. Representative consumer's demand for product i is

$$C_i = \left(\frac{\varphi_i}{p_i}\right)^{\sigma} \frac{\sum_g r_g X_g^{agg} + \Pi}{\sum_j \varphi_j^{\sigma} p_j^{1-\sigma}} \tag{14}$$

The log of product i's share in consumption expenditure can then be written as

$$\log\left(\frac{p_i C_i}{\sum_j p_j C_j}\right) = -\log\left(\sum_j \varphi_j^{\sigma} p_j^{1-\sigma}\right) + (1-\sigma)\log p_i + \sigma \log \varphi_i \tag{15}$$

Adapting the expression for household-level consumption data with households indexed by h and denoting crop i's share in h's expenditure with $s_{h,i}$, the

11. Distortions are demeaned within season.

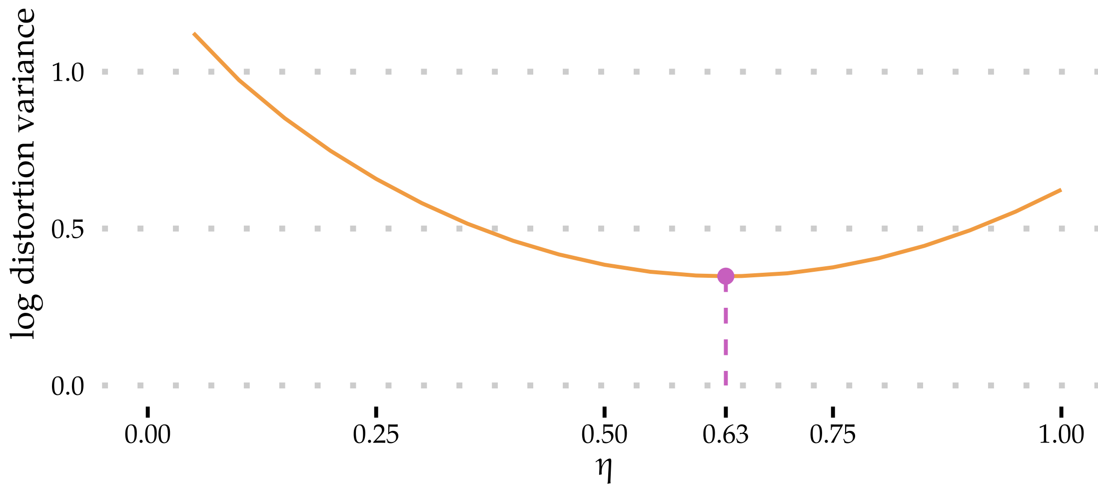


FIGURE 4: Variance of log extracted composite distortion term vs chosen concavity parameter η

expression can be written as a regression:

$$\log s_{h,i} = \beta_0 + \beta_1 \log p_{h,i} + \gamma_i + \varepsilon_{h,i} \tag{16}$$

Then the elasticity of substitution can be extracted from the coefficient on log price: $\sigma = 1 - \beta_1$. Furthermore, product taste weights can be extracted (up to a normalization) from the estimated product fixed effects: $\varphi_i = \exp\left(\frac{\gamma_i}{\sigma}\right)$.

Estimating Equation 16 as-is raises endogeneity concerns due to factors that can affect both prices and expenditure shares. To address these, we use an instrumental variable approach, instrumenting for crop prices using the interaction of village-level measures of elevation and ruggedness with rainfall, as well as the availability of pucca roads (all-weather roads). The identifying assumption is that ruggedness and slope interacted with rain affect the productivity and yield of the crops and hence affect the market price of crops, but do not influence the consumption share of crops independently. Similarly, the availability of roads is likely to affect the shadow prices of inputs and crops. The F statistic for the first stage is 230 which establishes the relevance of the instrument.

Results of estimation are reported in Appendix Table A.2. The estimation yields $\beta_1 = -0.7$, implying $\sigma = 1.7$. This is in between the estimates obtained in a similar setting by Kebede (2020) for Ethiopia, 1.3, and by Sotelo (2020) for Peru, 2.4.

6.5 Product Choice and Misallocation: Mechanisms

Allowing for product heterogeneity, endogenous product choice, and multi-product behavior in a model of firm-level misallocation impacts the estimates of misallocation obtained in the model. There are four major mechanisms through which a conventional single-product model misestimates the severity of misallocation in the data generated by farms growing heterogeneous crops. This section explores these mechanisms qualitatively.

6.5.1 Single-product models misinterpret product heterogeneity as frictions

Even in the absence of any fundamental frictions, the fact that different farms choose to produce different products will lead the single-product model to mistake heterogeneous crop choice for evidence of frictions.

Suppose that farm 1 draws a high productivity $z_{1,rice}$ in rice and farm 2 draws a high productivity $z_{2,veg}$ in vegetables, causing each farm to specialize in their respective high-productivity crop (so that $y_1 = y_{1,rice}$ and $y_2 = y_{2,veg}$). In the absence of frictions, marginal revenue products will be equalized across farms, accounting for the fact that rice is more land-intensive but vegetables are more labor-intensive:

$$\frac{\alpha_{g,rice}\eta(p_{rice}y_{1,rice})^{\eta}}{x_{1,g}} = \frac{\alpha_{g,veg}\eta(p_{veg}y_{2,veg})^{\eta}}{x_{2,g}} \quad \forall g$$

However, the marginal products evaluated by the single-product model that assumes a single aggregate production function for both crops will generally not happen to be equal:

$$\frac{\alpha_{g,agg}\eta(p_{agg}y_1)^{\eta}}{x_{1,g}} \neq \frac{\alpha_{g,agg}\eta(p_{agg}y_2)^{\eta}}{x_{2,g}} \quad \forall g$$

This apparent heterogeneity in single-product marginal revenue products will lead the single-product model to impute frictions where there are none and thus to *overstate* the true cost of misallocation.

Figure 5 illustrates this mechanism for a set of farms that suffer from no distortions ($\tau_{f,g} = 1\ \forall f, g$) but have different relative productivities in rice vs vegetables. Farmers who are relatively productive in rice specialize in this land-intensive crop, while farmers who are relatively productive in vegetables specialize in this labor-intensive crop. The single-product model explains the heterogeneous input choices with heterogeneity in friction terms, such as the pictured single-product model's estimated $\hat{\tau}_{f,labor}$.

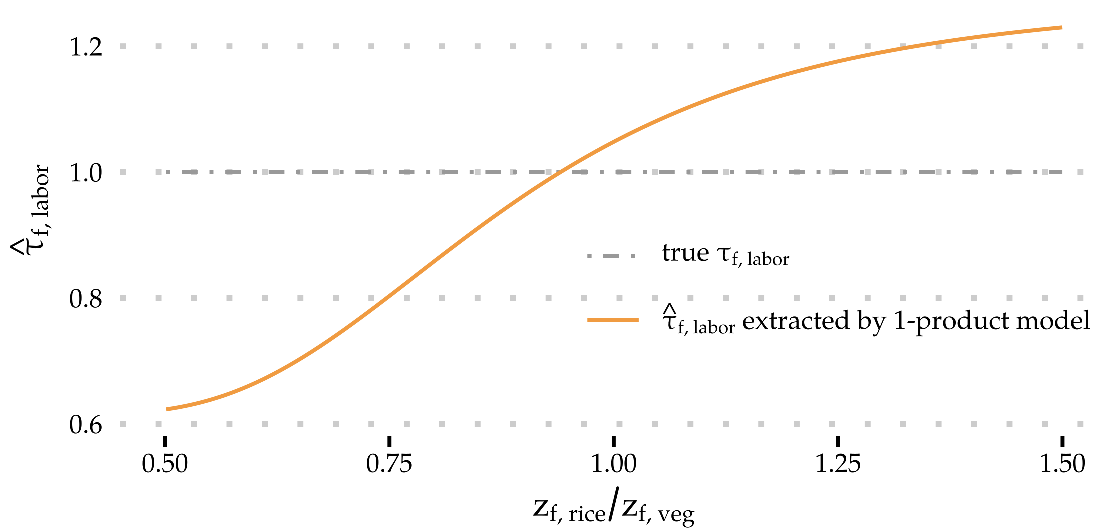


FIGURE 5: Distortion extracted by 1-crop model from multi-crop model data when productivities are heterogeneous: model illustration

6.5.2 Single-product models miss returns-to-scale heterogeneity

Estimates of misallocation depend on the returns to scale in the production function. When returns to scale are small, the most productive firms aren't able to grow their output by much even when the distortionary frictions are removed, reducing the gains from reallocation. Section 7.2 will illustrate the sensitivity of misallocation estimates to the degree of concavity in the model, of which decreasing returns to scale is one potential source. Single-product models impose the same level of returns to scale on the whole economy. But as Section 5.4 showed, we estimate returns to scale to be different across products: some, like "vegetables, fruits, and oil", have significantly decreasing returns to scale. Others, like rice, have returns to scale that are slightly decreasing but close to constant. The estimates for wheat, on the other hand, suggest increasing returns to scale. A single-product model will miss the fact that, in the event of a reallocation, farmers growing rice and wheat may be able to scale their production up more easily than vegetable farmers. This channel can lead single-product models to *understate* the true cost of misallocation.

6.5.3 Single-product models understate TFP dispersion

By ignoring heterogeneity across products within a single farm, single-crop models miss productivity differences between crops that farmers may exploit once frictions are removed.

Suppose that a given farm draws a low productivity in rice, $z_{1,rice} = z_L$, and a high productivity in vegetables, $z_{2,veg} = z_H$. If the farm faces severe frictions, they may destort the farm's product choice so that instead of specializing in vegetables, the farm also grows a significant quantity of the less productive rice. If the farm is treated as producing a homogeneous agricultural good, its farm-level TFP will be some average $z_L < z_M < z_H$. Applying the multi-product model to the farm would show that if frictions were removed, the farm would specialize in vegetables, raising its farm-level productivity from z_M to z_H. Applying a single-product model, however, would suggest that the farm-level productivity would remain at z_M even if frictions were removed, since the single-product model is blind to productivity differences between crops within a single farm. Therefore, the single-product model would understate the underlying productivity differences within the economy, and would *understate* the true cost of misallocation.

6.5.4 Single-product models miss the endogenous response of farms to frictions

Farms respond to frictions by changing product choice. The existence of choice between heterogeneous products changes how farmers respond to distortions. Figure 6 shows how crop-specific outputs change as the labor market distortion is altered for a farm that has access to two crops: rice (relatively land-intensive, see Section 5.4) and the combined "vegetables, fruits, and oilseeds" crop (relatively labor-intensive). When the labor distortion $\tau_{f,labor}$ is below average (the farm's labor is effectively subsidized), the farm chooses to specialize heavily in vegetables, fruits, and oils: this combined crop allows the farm to take advantage of the relatively cheaper labor input. As the distortion term grows and the subsidy diminishes, the farm reduces its output as it is unable to hire as much labor. At a certain point, the subsidy is so minor that the farm starts growing both crops at the same time. As the subsidy turns into a tax and labor becomes unattractive relative to land, the farm changes its specialization to rice: the relatively less labor-intensive crop.

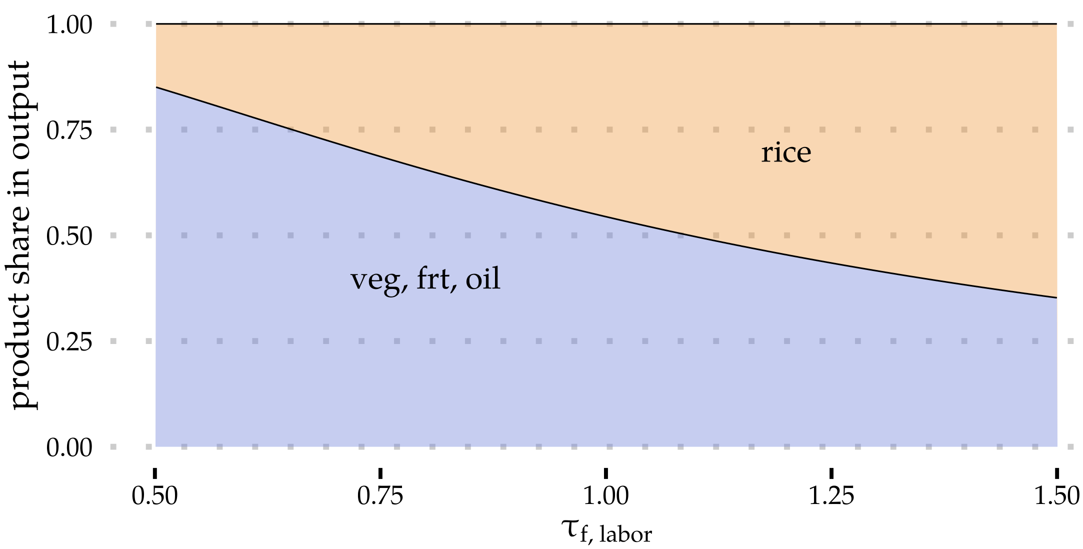


FIGURE 6: Product choice vs labor distortion: model illustration

Single-product models overstate the severity of distortions. This endogenous adjustment of farm's product choice to the distortions it faces allows the farm to lower its exposure to market frictions. Because traditional single-product models miss this margin of adjustment, they overstate the underlying distortions. A modest tax-like labor distortion would cause the farm to lower its $\frac{\text{labor}}{\text{output}}$ ratio both in the single- and the multi-product models of misallocation. But in the multi-product model, the distortion further leads the farm to shift part of its production to less labor-intensive crops, lowering the overall $\frac{\text{labor}}{\text{output}}$ ratio even more. This endogenous product choice response to a modest underlying friction would be mistaken by the single-product farm as evidence of severe underlying frictions. Figure 7 plots the labor distortion estimated by the single-crop model, $\hat{\tau}_{f,labor}$, when applied to observed inputs and outputs generated by the simulated multi-crop model. If the single-crop model could extract the fundamental friction correctly, its estimate would lie along the 45° line. Instead, the single-crop model significantly overstates the severity of both tax-like and subsidy-like distortions, and therefore *overstates* the cost of misallocation.

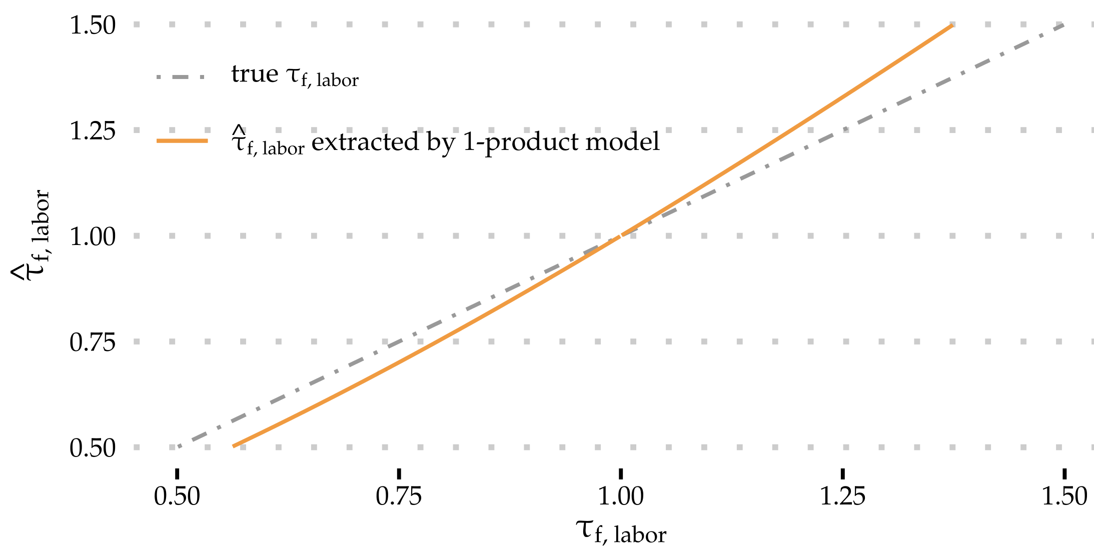


FIGURE 7: Distortion extracted by 1-crop model from multi-crop model data when distortions are heterogeneous: model illustration

7 QUANTITATIVE RESULTS

The preceding section showed that the way underlying distortions affect farm behavior and the way these distortions are extracted from observational data are both affected by the presence of heterogeneous products and multi-product farms. The present section quantifies the importance of these differences for the aggregate cost of misallocation in the economy.

To estimate the aggregate cost of misallocation, we use the model to conduct counterfactual reallocation exercises. Each exercise equalizes distortions between farms, computes their optimal counterfactual outputs, and aggregates the output to obtain a counterfactual reallocation output gain. The % difference in output between the counterfactual exercise and the currently observed one constitutes the headline measure of the aggregate cost of misallocation: it captures the magnitude of output being lost due to the market frictions currently present in the economy.

Note that the aggregate input quantities are not changed in the reallocation exercises: only the distortions are. Therefore, any aggregate output gain from reallocation is due purely to the increase in aggregate TFP. Furthermore, because the idiosyncratic farm-crop-level TFPs are not changed in the counterfactual either, the increase in aggregate TFP is driven not by any increase in farm-level TFPs but rather by an improved allocation of existing inputs to the more productive farms.

7.1 Benchmark Reallocation Exercise

In our benchmark reallocation exercise, we equalize all farm-level distortions while keeping the crop set of each farm restricted to what it is observed growing in the data. Furthermore, we leave farm-crop-level distortions intact to preserve idiosyncratic product choice motives.

Equalizing distortions. We impose equality of farm-input-level distortions: $\tau_{f,g} = \bar{\tau}_g \; \forall g$, $\lambda_f = \bar{\lambda}$. This equalization effectively makes input markets frictionless across farms. The levels of average distortions are fixed by the general equilibrium conditions. We leave the farm-input-crop-level distortions $\tau_{f,g,i}$, $\tau_{f,l,i}$ at their current levels to allow the model to capture idiosyncratic product choice motives, as will be discussed below.[12]

Keeping crop sets fixed. In this exercise, farmers will change the total quantity purchased of each input as well as the allocation of inputs between the crops they grow. However, farmers are not allowed to change the set of crops they are

12. Aragón, Restuccia, and Rud (2022) argue that estimates of agricultural misallocation are more reliable when the unit of analysis is a farm rather than a plot. We rely on plot-level data for production function identification, but in our misallocation exercise, we use variation across plots within a farm only when they produce different crops. Furthermore, leaving farm-input-crop-level distortions intact ensures that any remaining unexplained variation across plots is not interpreted as a source of misallocation, effectively making the exercise capture misallocation between farms and crops, but not between plots.

growing: if a particular farm is observed growing only rice and groundnut in the data, it will not be allowed to start growing oilseeds in the counterfactual. This restriction permits us to only use productivities and distortions that are extracted directly from the data: we don't need to make any assumptions on what each farm's counterfactual productivity or farm-crop distortion in an unobserved crop would be. Lifting this restriction requires calibrating and simulating the underlying unobserved distributions and will be done in Section 7.4.

Preserving idiosyncratic product choice motives. The model of a profit-maximizing farm is a simplification. Real-world farmers may choose to grow a particular crop for reasons other than pure profit maximization based on observed total factor productivity and input costs. For instance, semi-subsistence farmers may prefer a crop that is less profitable if sold on the market but serves their family's dietary needs better. The model is blind to these motives and would explain this observed preference of a non-profit-maximizing crop with a combination of an inefficient crop set (which the model would make the farm move away from if given a choice in a complete counterfactual reallocation) and farm-input-crop-level distortions $\tau_{f,g,i}$ and $\tau_{f,l,i}$ making inputs relatively cheaper for the subsistence crop (which would be equalized in a complete reallocation exercise). Therefore, allowing farmers to change crop sets and equalize farm-input-crop-level distortions may result in the model overstating the true costs of misallocation as it would overstate the degree to which farmers are likely to switch to profit-maximizing crops. Avoiding this is the main reason we elect not to equalize farm-input-crop-level distortions and the second reason we keep crop sets fixed in the headline counterfactual exercises. Even though the model does not explicitly include subsistence and other idiosyncratic product choice motives, making these two restrictions makes the counterfactual reallocation exercise only capture the output gain from efficient input markets allowing farmers to more efficiently grow whatever they currently prefer, without making unrealistic predictions on how they could "improve" their existing product preferences.

Seasons. The treatment of seasons in the model is irrelevant for its ability to reproduce the observed allocation, but becomes important for counterfactual exercises. If model farms f, mapped to farm-seasons in the data, are all bundled together into the same economy, a naive counterfactual reallocation may feature double-booking of inputs: land and labor inputs from less productive seasons may be reallocated to the more productive one, implying that a single physical piece of land may be allocated to three different uses in a single season and to no uses in the other two seasons. To preclude this nonsensical possibility, we split the model economy into three sub-economies by season, and conduct counterfactual reallocations only within each season: this ensures that no inputs are reallocated across seasons. The aggregate counterfactual output in each exercise is then a summation of three seasonal outputs.

Comparing multi-product and single-product models. To understand the effect of the product choice channel on aggregate productivity, we compare the estimated reallocation gain between the multi-product farm model and a traditional single-product farm model. The single-product farm model uses inputs and outputs measured at the farm-level rather than the farm-crop-level. Consequently, the single-product model uses a single production function with estimated "aggregate" elasticities from Figure 3.

7.2 Cost of Misallocation

Benchmark reallocation. The "benchmark" row of Table 2 shows the results of the benchmark counterfactual reallocation exercise. According to the multi-product model, allowing input markets to operate efficiently would raise the aggregate output of India's agricultural sector by 294%.[13] Repeating the same exercise in the single-product model promises a still formidable but smaller gain of 212%. For our preferred calibration and exercise, the single-product model *understates* the cost of misallocation by 28% (or 82 percentage points of current output).

Alternative reallocations. Our benchmark reallocation removes estimated frictions in all three modeled input markets: land, labor, and intermediates. The agricultural misallocation literature, however, has mainly focused on the costs of frictions in the land market. In the alternative "land only" reallocation, we remove

13. Enormous costs of misallocation in the agricultural sectors of developing countries have been estimated previously, most notably in Chen, Restuccia, and Santaeulalia-Llopis (2022).

TABLE 2: Reallocation exercises

	Gain	Gain (1 product)	Error (1 product)
land only	30%	36 %	20%
benchmark	**294%**	**212 %**	**-28%**
complete	311%	212 %	-32%

Note. Column "Gain" displays the % agricultural output gain from each counterfactual exercise, estimated using the multi-product model. Column "Gain (1 product)" displays the same but using the single-crop model. Column "Error (1 product)" displays the relative difference between the two: positive values indicate an overestimation of reallocation gain by the single-product model. "Land only" exercise equalizes land frictions between farms. "Benchmark" exercise equalizes all frictions between farms. "Complete" exercise additionally equalizes farm-crop frictions.

land frictions between farms (equalize λ_f) but keep their labor and intermediates frictions in place ($\tau_{f,g}$ stay at current levels). The expected reallocation gain in this case is far more modest but still sizeable: 30% of current agricultural output. The single-product model estimates a slightly higher reallocation gain of 36%: in contrast to the benchmark exercise, the single-product model *overstates* the potential gain in this incomplete reallocation by 20% (or 5 p.p.). Note also that Bolhuis, Rachapalli, and Restuccia (2021) arrived at a very similar gain of 38% from efficiently reallocating land in India, using a conventional single-product model.

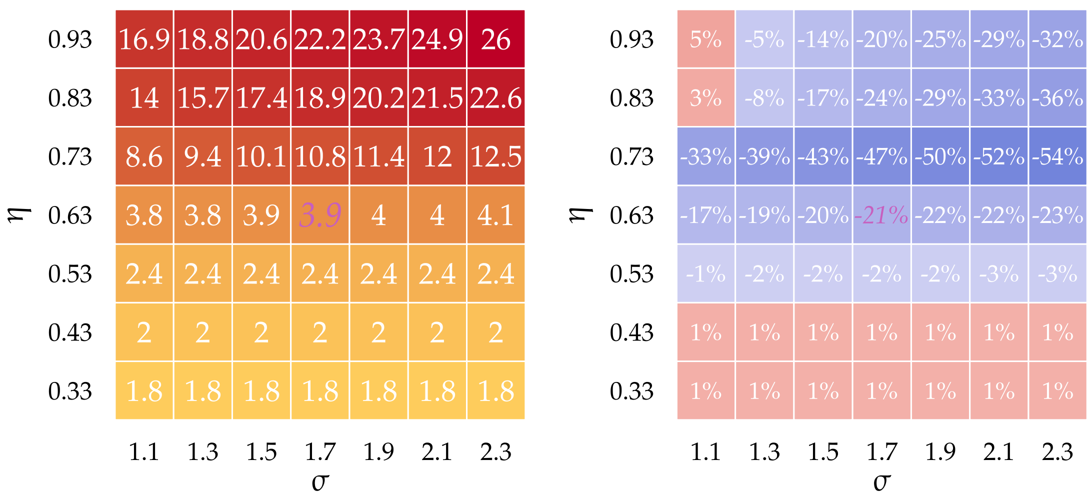
The "land only" results are instructive also because of the gap between them and the benchmark exercise: efficiently reallocating all three inputs is much more effective than reallocating land alone, implying that labor and input market frictions impose a significant drag on Indian agricultural productivity.[14]

The benchmark reallocation removes frictions between farms, but keeps farm-crop-level distortions intact in order to preserve any idiosyncratic product choice motives that the model does not explicitly capture. However, some of the farm-crop-level $\tau_{f,g,i}$ distortions are likely to reflect malign frictions rather than benign unmodeled motives. To place an upper bound on their importance, we conduct an alternative "complete" reallocation which equalizes these farm-crop-level frictions. Their contribution is relatively minor, increasing the reallocation gain from the 294% of the benchmark exercise to 311%. Because the single-product model

14. See Andrew D. Foster and Mark R. Rosenzweig (2022) for an exploration of labor market frictions in India.

is completely blind to the within-farm variation picked up by these frictions, its underestimation becomes even more severe, going from 28% to 32%.

Sensitivity of misallocation estimates. The multi-product model has two free parameters: additional concavity parameter η and elasticity of substitution between crops σ. Figure 8 displays the sensitivity of the benchmark reallocation exercise outcomes to alternative parameter values, centered on the benchmark calibration.



(A) Reallocation gain (B) Single-product model error

Note. Panel (A) displays estimated reallocation gain using the multi-product model at each set of parameters σ, η, expressed as a ratio (1 = no gain). Panel (B) displays the error in the single-product model's reallocation gain ratio relative to the one displayed in Panel (A): positive values denote overstatement of misallocation by the single-product model. The benchmark calibration is highlighted in the center.

Figure 8: Reallocation gain and single-product model error for alternative parameterizations

Panel (A) of Figure 8 shows how the reallocation gain predicted by the multi-product model (in ratios) changes with the two parameters. Our preferred calibration, discussed in Section 6.4, and its 3.9-fold (or 294%) reallocation gain is highlighted. The figure illustrates how sensitive misallocation estimates are to the degree of concavity in the model. The smaller is the additional concavity (higher

η), the greater is the reallocation gain predicted by the model. Productive farms grow in the efficient reallocation, but the lower the concavity, the greater is their optimal size, driving up the benefits of reallocating inputs to the most productive farms. The same effect is created by increasing the elasticity of substitution σ, although its magnitude is negligible at lower values of η. The figure tells a cautionary tale against basing any analysis of misallocation on a single parameterization: models of misallocation used in the literature are highly sensitive to the extent of concavity, whether it is driven by market power, decreasing returns to in production, or some other channel. This is also the reason why in Section 6.4 we elected not to calibrate η to some moment in the data but instead to pick the value that implies the most conservative underlying distribution of frictions.

Panel (B) of Figure 8 shows the response in the single-product model's error to the changing parameter values. While at the benchmark calibration the single-product model understates the reallocation gain by 21% (or 82 p.p.), neither the magnitude nor the sign of its error are set in stone. At some alternative calibrations, the single-product model *overstates* the reallocation gain instead, as the contribution of understating mechanisms diminishes.

These results imply that accounting for product heterogeneity and product choice in estimating misallocation is not a matter of applying a simple correction to single-product estimates: the sign and the magnitude of the needed correction depend on the exercise and calibration used by the researchers.

7.3 Mechanism Decomposition

Section 6.5 discussed several prominent mechanisms that make the single-product model's estimates of misallocation diverge from those of the multi-product model. In this section, we decompose the difference between the two models identified above into contributions by each of these mechanisms.

Exercises. The mechanisms are conceptual (rather than driven by a single paramter) and highly interacting, and so do not allow a clean decomposition of the single-product model's error into additive components. However, we can get a sense of the contribution of each mechanism through a series of exercises.

To evaluate the extent to which the single-product model misinterprets product heterogeneity as frictions (Section 6.5.1), we take the counterfactual allocation produced by the multi-product model once frictions are removed and use the single-product model to extract frictions from it as if it was real data. Then we conduct an additional counterfactual reallocation, removing these "fake" frictions from the single-product model. Because the fake frictions reflect only the fact that the single-product model misinterprets production function heterogeneity as evidence of frictions even when there are none, the additional reallocation gain obtained in this exercise provides an estimate of this mechanism's contribution to the single-product model's overall error.

We isolate the contribution of the single-product model's blindness to returns-to-scale heterogeneity (Section 6.5.2) by shutting this heterogeneity down. We conduct an auxiliary counterfactual reallocation in the multi-product model after equalizing the returns to scale in every product's production function to those in the aggregate production function used by the single-product model. Elasticities of all three inputs are rescaled accordingly, to preserve their relative quantities. The difference in reallocation gain between this auxiliary exercise and the benchmark exercise indicates the contribution of returns-to-scale heterogeneity to the single-product model's error.

The endogenous product choice mechanism (Section 6.5.4) can be shut down by precluding farms from changing their product sets and the relative allocation of inputs across crops. The former is already prohibited in the benchmark allocation (see Section 7.1). The latter can be shut down by solving an auxiliary model in which farms can change the total quantity of each input hired but not the allocation of this quantity between crops: product shares are fixed at exogenously given levels. This model is described in Appendix D. To get an estimate of the endogenous product choice response to frictions currently present in the economy, we re-introduce currently observed frictions into the counterfactual "efficient" economy, but don't let farmers change their product choice from the counterfactual ratios. Comparing the predicted output in this case to the observed actual output isolates the contribution of the endogenous mechanism.

The TFP dispersion mechanism (Section 6.5.3) can be shut down by applying the single-product model to current data in which all farm-season-crop combinations are treated as separate farm-seasons, effectively turning it from a farm-level model into a plot-level model. In this case, the single-product model still applies the same production function to all farms but it sees a more fine-grained distribution of productivities. However, we do not conduct this exercise for the benchmark reallocation because it would lead the single-product model to treat farm-crop-level distortions $\tau_{f,g,i}$ (which are not reallocated in the benchmark exercise) as farm-level distortions $\tau_{f,g}$ (which would be reallocated), leading to an unfair comparison of reallocation gains.

Results. Figure 9 presents the contribution of each of the three mechanisms (excluding the TFP disperion mechanism since it doesn't apply to the benchmark exercise) to the single-product model's misallocation error in the benchmark exercise, in percentage points. The contributions are visualized sequentially and additively, but note that the sum of the three mechanisms does not add up to the 82 percentage point error exactly, since the three identified mechanisms are neither exhaustive nor independent of one another. Still, the decomposition provides a useful overview of their relative importances.

The fact that the single-product model can treat efficient product heterogeneity as evidence of non-existing frictions leads the model to overstate misallocation by 27 p.p. The single-product model more than offsets this overstatement, however, by ignoring the heterogeneous returns to scale between products, which leads it to understate misallocation by 93 p.p. Finally, the endogenous product choice mechanism makes a minor 0.5 p.p. contribution to the model's understatement. Note that in this exercise, the endogenous product choice mechanism being identified works only through the relative quantities of crops grown by the farm rather than through the choice of *which* products it grows, since the latter is fixed at its current level (this limitation will be addressed in Section 7.4 below).[15]

7.4 Extensive Product Margin

Allowing farmers to change not only their crop mixes, but also their crop sets requires making additional assumptions as the farm's idiosyncratic productivities $z_{f,i}$ and input-crop-level distortions $\tau_{f,g,i}$, $\tau_{f,l,i}$ are not known for farm-crop combinations not currently observed. Therefore, we have to assume parametric

15. Appendix Figure A.2 repeats the decomposition for the "complete" reallocation and includes the TFP dispersion mechanism. The TFP dispersion mechanism contributes a 24 p.p. understatement.

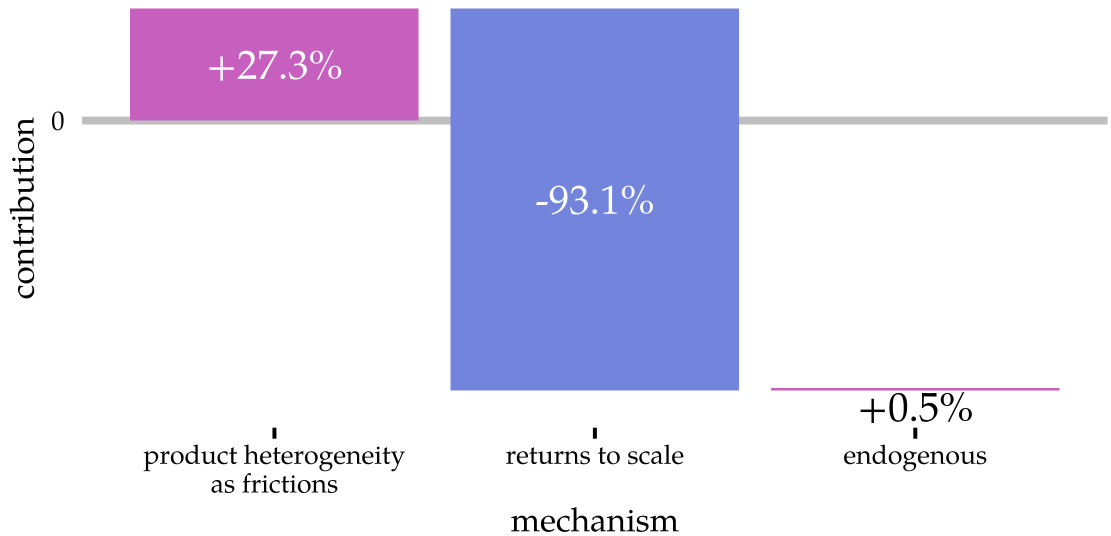


FIGURE 9: Mechanism contributions to the single-product model's misallocation error, benchmark reallocation

forms for the unconditional distributions of farm-level productivities and farm-input-crop-level distortions and then calibrate the parameters of these *unconditional* distributions to match the empirically observed *conditional* distributions.

Thus, this exercise is based on simulating the economy in a calibrated model rather than directly representing each observed farm with its model analog, as exercise 1 was able to do.

coming soon!

8 CONCLUSION

Misallocation of inputs between firms imposes a significant drag on the productivity of developing countries. The cost of misallocation in agriculture is especially large and contributes to explaining the enormous agricultural productivity gap between developed and low-income countries.

Estimates of misallocation in non-agricultural sectors have been limited by the treatment of all firms as producing only a single good or multiple homogeneous goods. Estimates of agricultural misallocation have been further limited by assuming that all farms produce the same agricultural product. We use Indian farm-level data to estimate production functions for individual crops and find that their input elasticities and returns to scale are meaningfully heterogeneous for the purposes of estimating misallocation.

We build a model of multi-product firms choosing among products with heterogeneous production functions. We find that conventional single-product models mis-estimate the cost of misallocation in several ways, some of which push the single-product models to overstate misallocation while others push them to understate it. The net error of a single-product model depends on the exercise of interest and chosen calibration. This implies that accounting for product heterogeneity and multi-product behavior is not a matter of applying a simple correction to a single-product model's output.

We conducted our analysis on Indian farm-level data, since agricultural micro-level data provides several advantages in exploring the interplay of product heterogeneity and misallocation. However, the model developed in the paper—and the mechanisms it illustrates—would apply equally to non-agricultural firms. Future research can extend this analysis from agriculture to non-agriculture by applying the model to firm-product-level datasets.

Another future direction to consider the role of dynamics in the interaction of product choice and misallocation.[16]. Crop rotation can lead to input choices that appear inefficient in the short run but preserve long-run productivity of land. Moreover, the heterogeneous suitability of different crops to different seasons can lead—when input rental markets are underdeveloped—to input choices that appear inefficient in one season but are there to have the input in place for the next season.

16. Dynamic choices have been shown to be important for misallocation using single-product models by, for instance, Asker, Collard-Wexler, and De Loecker (2014) and Kehrig and Vincent (2019).

References

Adamopoulos, Tasso, Loren Brandt, Jessica Leight, and Diego Restuccia. 2022. "Misallocation, Selection, and Productivity: A Quantitative Analysis With Panel Data From China." *Econometrica* 90 (3): 1261–1282.

Aragón, Fernando, Diego Restuccia, and Juan Pablo Rud. 2022. *Assessing Misallocation in Agriculture: Plots versus Farms.* Working paper w29749. Cambridge, MA: National Bureau of Economic Research, February.

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. "Dynamic Inputs and Resource (Mis)Allocation." *Journal of Political Economy* 122 (5): 1013–1063.

Banerjee, Abhijit, and Lakshmi Iyer. 2005. "History, institutions, and economic performance: The legacy of colonial land tenure systems in India." *American economic review* 95 (4): 1190–1213.

Biau, Gérard, and Erwan Scornet. 2016. "A Random Forest Guided Tour." *TEST* 25 (2): 197–227.

Bils, Mark, Peter J. Klenow, and Cian Ruane. 2021. "Misallocation or Mismeasurement?" *Journal of Monetary Economics* 124:S39–S56.

Bischl, Bernd, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, et al. 2021. "Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges." Preprint, November 24, 2021. Accessed March 25, 2023. arXiv: 2107.05847 [cs, stat]. http://arxiv.org/abs/2107.05847.

Blanco, Cesar, and Xavier Raurich. 2022. "Agricultural Composition and Labor Productivity." *Journal of Development Economics* 158:102934.

Bolhuis, Marijn, Swapnika Rachapalli, and Diego Restuccia. 2021. *Misallocation in Indian Agriculture.* Working Paper 709. University of Toronto Department of Economics, October.

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

Chen, Chaoran. 2017. "Untitled Land, Occupational Choice, and Agricultural Productivity." *American Economic Journal: Macroeconomics* 9, no. 4 (1, 2017): 91–121.

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

Deininger, Klaus, Songqing Jin, and Hari K Nagarajan. 2008. "Efficiency and equity impacts of rural land rental restrictions: Evidence from India." *European Economic Review* 52 (5): 892–918.

Deininger, Klaus W, et al. 2003. *Land policies for growth and poverty reduction.* World Bank Publications.

Foster, Andrew D, and Mark R Rosenzweig. 2002. "Household division and rural economic growth." *The Review of Economic Studies* 69 (4): 839–869.

———. 2022. "Are There Too Many Farms in the World? Labor Market Transaction Costs, Machine Capacities, and Optimal Farm Size." *Journal of Political Economy* 130, no. 3 (1, 2022): 636–680.

Gandhi, Amit, Salvador Navarro, and David Rivers. 2017. *How Heterogeneous Is Productivity? A Comparison of Gross Output and Value Added.* Working Paper.

Gollin, Douglas, David Lagakos, and Michael E. Waugh. 2014. "Agricultural Productivity Differences across Countries." *American Economic Review* 104, no. 5 (1, 2014): 165–170.

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

Gottlieb, Charles, and Jan Grobovšek. 2019. "Communal Land and Agricultural Productivity." *Journal of Development Economics* 138:135–152.

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.

Jaef, Roberto N. Fattal. 2018. "Entry and Exit, Multiproduct Firms, and Allocative Distortions." *American Economic Journal: Macroeconomics* 10, no. 2 (1, 2018): 86–112.

Just, Richard E., David Zilberman, and Eithan Hochman. 1983. "Estimation of Multicrop Production Functions." *American Journal of Agricultural Economics* 65 (4): 770–780.

Kebede, Hundanol A. 2020. *The gains from market integration: The welfare effects of new rural roads in Ethiopia.* Technical report. Mimeo, University of Virginia.

Kehrig, Matthias, and Nicolas Vincent. 2019. *Good Dispersion, Bad Dispersion.* Working paper w25923. Cambridge, MA: National Bureau of Economic Research, June.

Krishnaswamy, Nandita. 2018. *At What Price? Price Supports, Agricultural Productivity, and Misallocation.* Working paper.

Le, Kien. 2020. "Land Use Restrictions, Misallocation in Agriculture, and Aggregate Productivity in Vietnam." *Journal of Development Economics* 145:102465.

Morando, Bruno. 2023. "Subsistence Farming and Factor Misallocation: Evidence from Ugandan Agriculture." *The World Bank Economic Review* (6, 2023).

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2 (1, 2017): 87–106.

Restuccia, Diego, and Richard Rogerson. 2017. "The Causes and Costs of Misallocation." *The Journal of Economic Perspectives* 31 (3): 151–174.

Schonlau, Matthias, and Rosie Yuyan Zou. 2020. "The Random Forest Algorithm for Statistical Learning." *The Stata Journal: Promoting communications on statistics and Stata* 20 (1): 3–29.

Shumway, C. Richard, Rulon D. Pope, and Elizabeth K. Nash. 1984. "Allocatable Fixed Inputs and Jointness in Agricultural Production: Implications for Economic Modeling." *American Journal of Agricultural Economics* 66 (1): 72–78.

Singh, Inderjit, Lyn Squire, and John Strauss. 1986. "A survey of agricultural household models: Recent findings and policy implications." *The World Bank Economic Review* 1 (1): 149–179.

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

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28, no. 2 (1, 2014): 3–28.

Vashishtha, Prem S. 1989. "Changes in Structure of Investment of Rural Households: 1970-71-1981-2." *Journal of Income and Wealth* 10 (2): 21–45.

Wang, Wenya, and Ei Yang. 2023. "Multi-Product Firms and Misallocation." *Journal of Development Economics* 163:103102.

Ziegler, Andreas, and Inke R. König. 2014. "Mining Data with Random Forests: Current Options for Real-World Applications: Mining Data with Random Forests." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4 (1): 55–63.

Appendix

A Additional Figures and Tables

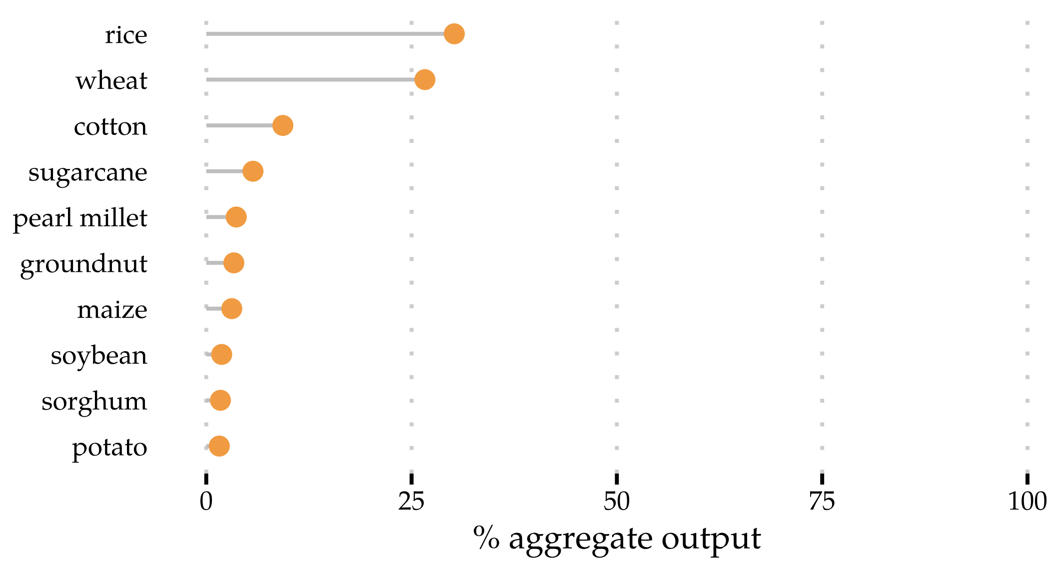


Figure A.1: Top 10 crops by % of aggregate output

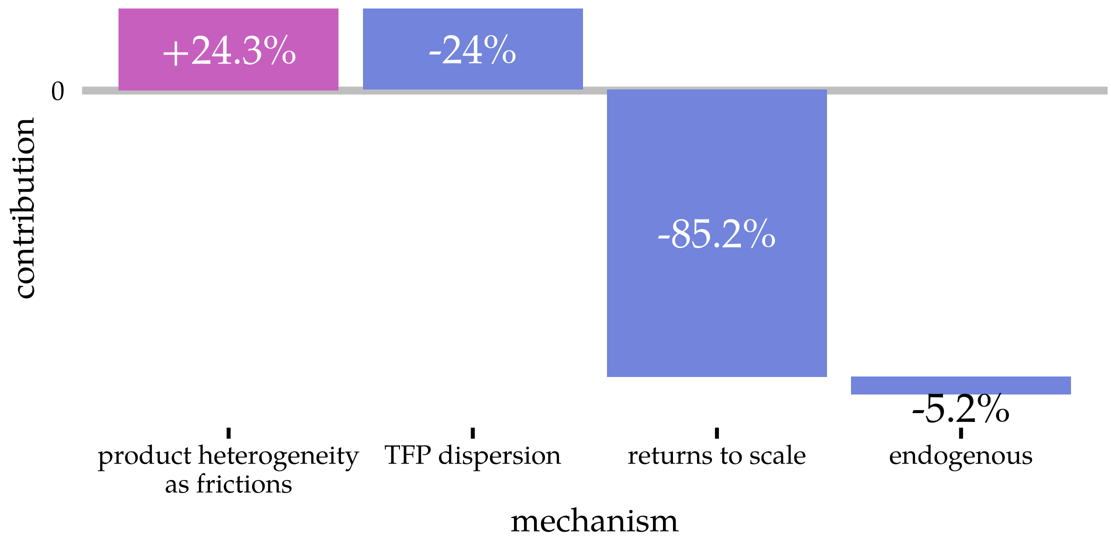


Figure A.2: Mechanism contributions to the single-product model's misallocation error, complete reallocation

TABLE A.1: Production function estimates

	(1) All crops	(2) Rice	(3) Wheat	(4) Other Cereals	(5) Pulses	(6) Veg, Frt, Oil
Land	0.427	0.511	0.414	0.407	0.379	0.330
	(0.031)	(0.021)	(0.032)	(0.034)	(0.060)	(0.035)
Labor	0.189	0.161	0.122	0.248	0.316	0.430
	(0.031)	(0.037)	(0.036)	(0.058)	(0.074)	(0.064)
Intermediates	0.351	0.307	0.517	0.245	0.255	0.127
	(0.041)	(0.033)	(0.040)	(0.046)	(0.076)	(0.052)
Observations	14,705	4,807	3,566	2,779	1,128	2,338
R^2	0.624	0.742	0.713	0.590	0.417	0.572
Village FEs	Y	Y	Y	Y	Y	Y
Season FEs	Y	Y	Y	Y	Y	Y
			First Stage: F statistics			
Land	77.0	62.0	40.3	37.8	15.7	19.3
Labor	49.3	34.7	17.7	25.2	12.9	14.8
Intermediates	35.8	31.7	21.5	19.9	8.9	11.8
K-Paap Wald F statistic	51.1	40.4	16.0	30.8	12.4	12.7

Note. Each column represents a separate regression. We use 2SLS to estimate the coefficients, using instruments described in Section 5.2. Robust standard errors are presented in parentheses. The "Other Cereals" group include barley, maize, jawar, bajra, ragi, and millets. The "Veg, frt, oil" group includes vegetables, fruits, oilseeds, and fiber crops. In panel B of the table, we report the F-stat from the first stage regression for each input and each regression separately. We also report the joint F-stat at the bottom line of the table.

TABLE A.2: Estimate of σ

	(1) $\log s_{h,i}$	(2) $\log p_{h,i}$
σ	1.699	
$\log p_{h,i}$	-0.699	
	(0.067)	
Elevation × rain		-0.004
		(0.000)
Ruggedness × rain		0.016
		(0.001)
Pucca roads availability		-0.023
		(0.001)
Observations	40,833	40,833
Kleibergen-Paap F stat	230.9	

Note. The table presents the estimation of Equation 16 using 2SLS. Column 1 presents second stage results and column 2 presents first stage results. Elevation, ruggedness, rain, and availability of pucca roads are village-level measures. "Pucca roads" are all-weather roads made of concrete and tar. Standard errors are clustered at the household level. The "σ" row shows the value of σ implied by β_1, the coefficient on $\log p_{h,i}$.

B Additional Concavity Mechanisms

Additional concavity governed by parameter η, described in Section 6.1, is a reduced-form stand-in for multiple mechanisms that could incentivize farmers to produce multiple crops at once.

In this section, we explore two such mechanisms explicitly. Both generate first order conditions that are functionally equivalent to those in the original problem. At the same time, different micro-foundations of η may carry with them different natural ways of aggregating crops and clearing markets at the aggregate level, generating slightly different general equilibrium implications.

B.1 Subsistence

Suppose that instead of operating a profit-maximizing farm, the household is operating a subsistence farm and consuming its output. The farmer is minimizing costs of providing utility constant elasticity of substitution utility U to the family:

$$U = \left(\sum_i \varphi_i y_{f,i}^{\eta} \right)^{\frac{1}{\eta}}$$

In this interpretation, the farmer has an incentive to mix crops in order to provide a diverse diet for the family. The resulting first order conditions are equivalent to those in Section 6.1.

B.2 Market Power

Suppose that instead of selling the harvest at a perfectly competitive market, the farmer has some market power in a monopolistically competitive market. This can be represented by introducing an intermediate crop aggregator operating within each crop i's market, who buys farms' varieties $y_{f,i}$ of this crop at $p_{f,i}$ and sells the combined i product to the consumer at price P_i. The aggregator uses the following production function:

$$Y_i = \left(\sum_f y_{f,i}^{\eta} \right)^{\frac{1}{\eta}}$$

This generates downward-sloping demand for farm f's variety of crop i, creating an incentive for the farmer to mix crops (to avoid excessive price drops in any one crop). The resulting first order conditions are equivalent to those in Section 6.1

C Model Solution and Computational Algorithm

C.1 Solving the Problem of a Single Farm

For N products, there are $2^N - 1$ non-empty product sets that each farm could be producing. Let I_f be the set of products in a single such combination, s.t. $y_{f,i} > 0\ \forall i \in I_f$ and $y_{f,i} = 0\ \forall i \notin I_f$. If $\omega > 0$, the profit-maximizing product set might have $y_{f,i} = 0$ for some crops. Each combination has to be solved separately, and then their profits compared to pick the best one.

For each product set I_f, the following non-linear equation needs to be solved for the shadow cost of land λ_f:

$$\sum_{i \in I_f} \lambda_f^{\frac{\eta \sum_g \alpha_{g,i} - 1}{1 - \eta \sum_g \alpha_{g,i} - \eta\gamma_i}} \left((p_i z_{f,i})^\eta \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta\alpha_{g,i}} \left(\frac{\gamma_i}{\tau_{l,i}} \right)^{1 - \eta \sum_g \alpha_{g,i}} \right)^{\frac{1}{1 - \eta \sum_g \alpha_{g,i} - \eta\gamma_i}} \tau_{l,i} = L_f \tag{C.1}$$

Each term within the sum is decreasing in λ_f. Using this fact, the following bounds can be derived on λ_f that solves the condition, simplifying the numerical cost. For each crop, define:

$$A_i = \left((p_i z_{f,i})^\eta \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta\alpha_{g,i}} \left(\frac{\gamma_i}{\tau_{l,i}} \right)^{1 - \eta \sum_g \alpha_{g,i}} \right)^{\frac{1}{1 - \eta \sum_g \alpha_{g,i} - \eta\gamma_i}} \tau_{l,i} \tag{C.2}$$

Then the upper and lower bounds on λ_f are

$$\lambda_f^{UB} = \max_i \left(\frac{L_f}{|I_f| \cdot A_i} \right)^{\frac{1 - \eta \sum_g \alpha_{g,i} - \eta\gamma_i}{\eta \sum_g \alpha_{g,i} - 1}} \tag{C.3}$$

$$\lambda_f^{LB} = \min_i \left(\frac{L_f}{|I_f| \cdot A_i} \right)^{\frac{1 - \eta \sum_g \alpha_{g,i} - \eta\gamma_i}{\eta \sum_g \alpha_{g,i} - 1}} \tag{C.4}$$

Once the λ_f for this product set is known, input choices *for products within I_f* can be obtained from:

$$x_{f,h,i} = \frac{\alpha_{h,i}}{r_h \tau_{f,h} \tau_{f,h,i}} \left((p_i z_{f,i})^{\eta} \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta \alpha_{g,i}} \left(\frac{\gamma_i}{\lambda_f \tau_{l,i}} \right)^{\eta \gamma_i} \right)^{\frac{1}{1-\sum_g \eta \alpha_{g,i} - \eta \gamma_i}} \tag{C.5}$$

$$l_{f,i} = \left((p_i z_{f,i})^{\eta} \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta \alpha_{g,i}} \left(\frac{\gamma_i}{\lambda_f \tau_{l,i}} \right)^{1-\eta \sum_g \alpha_{g,i}} \right)^{\frac{1}{1-\eta \sum_g \alpha_{g,i} - \eta \gamma_i}} \tag{C.6}$$

Set $x_{f,h,i} = l_{f,i} = 0 \ \forall i \notin I_f$.

With input choices computed, the profit for this product set can be evaluated using Equation 2. Iterate over $2^N - 1$ possible product sets and choose the product set yielding the highest profit.

C.2 Solving for the General Equilibrium

Merge consumer's product demands, farms' input demands, farms' optimal output, and market clearing conditions to obtain market clearing conditions in terms of fundamentals, shadow costs of land λ_f, goods prices p_i, inputs prices x_g, and farms' product sets I_f (letting the set of farms producing each product to be denoted by F_i):

$$\sum_f \sum_{i \in I_f} \left((p_i z_{f,i})^{\eta} \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta \alpha_{g,i}} \cdot \left(\frac{\gamma_i}{\lambda_f \tau_{f,l,i}} \right)^{1-\eta \sum_g \alpha_{g,i}} \right)^{\frac{1}{1-\eta \sum_g \alpha_{g,i} - \eta \gamma_i}} = L^{agg} \tag{C.7}$$

$$\sum_f \sum_{i \in I_f} \frac{\alpha_{h,i}}{r_h \tau_{f,h} \tau_{f,h,i}} \left((p_i z_{f,i})^{\eta} \cdot \eta \cdot \Pi_g \left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\eta \alpha_{g,i}} \cdot \left(\frac{\gamma_i}{\lambda_f \tau_{l,i}} \right)^{\eta \gamma_i} \right)^{\frac{1}{1-\eta \sum_g \alpha_{g,i} - \eta \gamma_i}} = X_h^{agg} \tag{C.8}$$

$$p_i = \Bigg(\varphi_i^{-\sigma} \frac{\sum_j \varphi_j^{\sigma} p_j^{1-\sigma}}{\sum_j p_j \sum_{f \in F_j} y_{f,j}} \\ \cdot \sum_{f \in F_i} \left(\eta^{\sum_g \alpha_{g,i} + \gamma_i} z_{f,i} \Pi_g \left[\left(\frac{\alpha_{g,i}}{r_g \tau_{f,g} \tau_{f,g,i}} \right)^{\alpha_{g,i}} \right] \left(\frac{\gamma_i}{\lambda_f \tau_{f,l,i}} \right)^{\gamma_i} \right)^{\frac{1}{1-\eta \sum_g \alpha_{g,i} - \eta\gamma_i}} \Bigg)^{\frac{1-\eta \sum_g \alpha_{g,i} - \eta\gamma_i}{\left(1-\eta \sum_g \alpha_{g,i} - \eta\gamma_i\right)(1-\sigma) - 1}} \tag{C.9}$$

The p_i condition can be simplified in two ways, to abstract from its dependence on all other goods' prices p_j. First, the price level can be normalized: $\sum_j \varphi_j^{\sigma} p_j^{1-\sigma} = 1$. Second, the nominal output can be substituted with a variable Y_{nom}, adding its identity to the list of market-clearing conditions:

$$Y_{nom} = \sum_j p_j \sum_{f \in F_j} y_{f,j} \tag{C.10}$$

Splitting distortions and ensuring the current allocation is consistent with general equilibrium. Once frictions have been extracted from the data (Section 6.3), additional steps need to be taken to ensure that the model representation of the current allocation is in general equilibrium.

Splitting $r_g \tau_{f,g} \tau_{f,g,i} = mrpg_{f,i}$ into r_g and $\tau_{f,g}$ is arbitrary. To ease interpretation of frictions, we impose that $\tau_{f,g} \tau_{f,g,i}$ only distort the farm-level input demands and do not affect the aggregate input demand X_g^{agg}. This implies the following market price of g:

$$r_g = \frac{\sum_i \eta \alpha_{g,i} \sum_f (p_i y_{f,i})^{\eta}}{X_g^{agg}} \tag{C.11}$$

Extending the logic, we impose that $\tau_{f,g,i}$ only distort the within-farm allocation of inputs between crops and do not affect the farm-level input demand, implying the following:

$$\tau_{f,g,i} = \frac{\eta \alpha_{g,i} (p_i y_{f,i})^{\eta}}{\sum_j \eta \alpha_{g,j} (p_j y_{f,j})^{\eta}} \frac{\sum_j x_{f,g,j}}{x_{f,g,i}} \tag{C.12}$$

Finally, the unobserved taste weights φ_i need to be consistent with the observed allocation being in general equilbrium. To simplify notation, we map the price-weighted outputs in the data to real output $y_{f,i}$ in the model, implying that all prices in the current model allocation are: $\{p_i\} = 1$. Any such rescaling does not affect the real behavior of the model, neither in the current allocation nor in the counterfactual reallocations, since units of output in different products are arbitrary. This rescaling simplifies extracting the taste weights consistent with the current allocation being in general equilibrium to:

$$\varphi_i = \left(\frac{\sum_f y_{f,i}}{\sum_j \sum_f y_{f,j}} \right)^{\frac{1}{\sigma}} \tag{C.13}$$

Note that this procedure makes extracting taste weights from estimated fixed effects in Equation 16 unnecessary. The estimated taste weights are still needed, however, when the model is simulated to allow for flexible product sets (see Section 7.4).

When Product Sets are Fixed. Although the current allocation extracted from the data can be easily made to be consistent with general equilibrium in the model, market-clearing prices have to be solved numerically for any counterfactual reallocation. When product sets are fixed at the currently observed level, the problem of finding market-clearing prices using Equations C.7-C.9 is simplified by the fact that farmers cannot switch their product sets in response to changing prices.

If at least land is being efficiently reallocated between farms (λ_f is equalized between farms), the market-clearing prices can be solved without having to resolve each individual farm's problem.[17] This is the case in the benchmark reallocation exercise discussed in Section 7.1.

In this case, farms' product sets I_f are known from the current allocation. Since land is being reallocated, $\lambda_f = \bar{\lambda}\ \forall i$. This allows the land market condition C.7, the flexible input markets conditions C.8 and the input market identity C.10 to be solved as a system with $G+2$ dimensions in $G+2$ unknowns: $\bar{\lambda}$, $\{r_g\}_g$, and Y_{nom}. On each guess of these variables, goods prices consistent with the guess can be computed from C.9.

17. This approach works whether or not any of the flexible input frictions $\tau_{f,g}$, $\tau_{f,g,i}$ are being equalized.

When Product Sets are Allowed to Change. When farms are allowed to change their product set I_f from its current level, the need to solve all individual farms' problems on each price guess for new product sets can no longer be avoided.

By analogy with flexible inputs (Section 6.3), let each farm's land shadow cost deviation be fixed at $\hat{\lambda}_f = \lambda_f / \bar{\lambda}$.

The brute-force approach to solving for market-clearing prices is to solve conditions C.7, C.8, C.9 as a system with $G + N + 1$ dimensions in the same number of unknowns: $\bar{\lambda}, \{x_g\}_g, \{p_i\}_i$. Within each price guess:

1. Solve individual farms' problems. For each farm:

    (a) For each potential product set I_f, find the land endowment L_f consistent with $\lambda_f = \bar{\lambda} \cdot \hat{\lambda}$ using Equation C.1.

    (b) Solve the farm's problem for the optimal product set I_f, following Appendix C.1.

2. Evaluate conditions C.7, C.8, C.9.

The brute-force algorithm is guaranteed to find market-clearing prices, but is exceptionally computationally costly due to its high-dimensionality and the need to re-solve all farms' problems on each price guess. However, there is a much faster sequential heuristic algorithm:

1. Pick an initial guess of $\bar{\lambda}, \{x_g\}_g, \{p_i\}_i$.

2. Solve individual farms' problems as in the brute-force algorithm.

3. Solve the system of market conditions as in the fixed crop set algorithm for new price guesses: this assumes that that product sets do not respond to changing prices.

4. Solve individual farms' problems as in the brute-force algorithm with price guesses obtained from the fixed crop set GE system.

5. Evaluate convergence in product sets. If all farms' optimal product sets obtained in step 4 are the same as in step 2, stop. If not, return to step 2.

This algorithm exploits the low-dimensionality and low computational cost of solving for general equilibrium when product sets are fixed. By sequentially updating the farms' product sets, this algorithm converges to the same answer as the brute-force approach in just a few iterations.

D Fixed Input Shares Model

An auxiliary model in which the relative allocation of inputs across crops within a farm is fixed is useful for shutting down the effect of the endogenous product choice mechanism (Sections 6.5.4, 7.3). In this model, the farmer can choose the overall quantity of each input hired, but the share of the input g allocated to each crop i is fixed at exogenously given $s_{f,g,i}$. The set of produced crops is likewise fixed at I_f. The setup of the model is otherwise similar to that in Section 6.1:

$$\max_{\{y_{f,i}\}_i, L_f^d, \{X_{f,g}\}_g} \sum_{i \in I_f} \left(p_i y_{f,i}\right)^{\eta} - \sum_{g=1}^{G} r_g \tau_{f,g} \sum_{i \in I_f} \tau_{f,g,i} s_{f,g,i} X_{f,g} \tag{D.1}$$

s.t.

$$y_{f,i} = z_{f,i} \left(s_{f,l,i} L_f^d\right)^{\gamma_i} \Pi_g \left[\left(s_{f,g,i} X_{f,g}\right)^{\alpha_{g,i}}\right] \tag{D.2}$$

$$\sum_{i \in I_f} s_{f,l,i} \tau_{f,l,i} L_f^d = L_f^s \qquad (\lambda_f) \tag{D.3}$$