Multidimensional Sorting in the Land Market*

Rafael Araujo† Kátia Nishiyama‡

July 2, 2021

Abstract

In this paper, we study the sorting of farmers and land in a multidimensional setting. We incorporate a discrete choice model, where heterogeneous farmers choose among different crops, inside a central planner's problem, that allocates farmers to land to maximize aggregate expected production. We derive a sorting expression that can be used to build counterfactual allocations of farmers, and discuss when this sorting expression is only necessary or both necessary and sufficient to characterize the optimal matching. To illustrate our theoretical results we estimate our model using pixel level data from Brazil across 15 years. We show that a relocation of farmers could increase the overall agricultural productivity by as much as 34%.
JEL: *Q12, Q18, C78* Keywords: *Multidimensional sorting, Misallocation, Agricultural Productivity, Technology Adoption*

*We are grateful to Francisco Costa, Carlos Eugênio da Costa, Lucas Maestri, Marcelo Sant'Anna, Heron Rios, Arthur Bragança, Alexandre Machado, and seminar participants at FGV EPGE for valuable comments and suggestions. All errors are our own

†FGV EPGE and Climate Policy Initiative (e-mail: carlquist.rafael@gmail.com).

‡FGV EPGE (e-mail: katia.nishiyama@fgv.br).

1 Introduction

Increasing agricultural productivity is paramount to improve food security, a key element of economic development. There is a large literature exploring different approaches to increase productivity in the agricultural sector, such as increasing access to fertilizers (Duflo et al. (2008)); adoption of more productive technologies (Conley and Udry (2010), Suri (2011)); improvement of infrastructure (Fajgelbaum and Redding (2018)); and reducing misallocation (Ayerst et al. (2020), Adamopoulos and Restuccia (2020)). Misallocation can arise from frictions on different input markets, such as, capital, labor, and land. If farmers and land have heterogeneous and complementary productivity, then misallocation of land can be studied in a setting of sorting between farmers and land. One way of enriching our understanding of agricultural productivity is to acknowledge that land is defined by multiple heterogeneous features. Nonetheless, optimal sorting in a multidimensional setting has not been fully characterized in the literature.

In this paper, we combine a discrete choice model of crops with a multidimensional sorting model of farmers and land to study how the misallocation between farmers and land affects aggregate agricultural productivity in a multidimensional setting. Farmers are characterized by a one dimensional space of productivity while land is characterized by a multidimensional vector of characteristics, such as soil productivity and transportation costs. To study the counterfactual where farmers can be relocated, we incorporate the discrete choice model into a central planner's matching problem that wishes to maximize aggregate expected production. We derive a necessary sorting condition as a result of the central planner's problem of allocating farmers to land. This necessary condition can be used to compute a lower bound on the potential gain from relocating farmers. Imposing an additional twist condition, which is a generalization of the Spence-Mirrlees condition (Spence (1978),Mirrlees (1971)), our sorting characterization becomes both necessary and sufficient for the optimal allocation of farmers.

Even though the optimal matching function linking farmers to land can be compli-

cated, we are able to define our sorting conditions with respect to rankings of different variables on both sides of the market, as in Hagedorn et al. (2017). Although land cannot be naturally sorted, due to its multidimensional characterization, the marginal expected land return with respect to farmer's productivity is a one dimensional object that can be sorted. The optimal sorting is then governed by this relation: the more productive the farmer is, the more steep should be the marginal expected return of the land that they are allocated to.

This general result arises when the expected land return – which gives the output function of a matching – is convex with respect to farmer's productivity. This convexity is a property implied by our discrete choice model, since increasing productivity not only increases the production of a crop, but also the probability that a farmer will have access to different crops of higher returns. The convexity of the output function is what guarantees our sorting condition.

To motivate our theoretical results, we estimate the crop choice model by maximum likelihood using highly disaggregated data from the most important agricultural producer state in Brazil, the state of *Mato Grosso*. We observe yearly land use decisions at the pixel level from 2003 up to 2017 in an area bigger than 900,000 km^2. We merge this land use data with data from the FAO GAEZ project - which gives land potential productivity for different crops at the pixel level - and pixel level transportation cost of different crops to international markets. This disagregated land use data coupled with the discrete choice model allows us to recover a residual heterogeneity which we interpret as being farmer's heterogeneous productivity. In this particular application, the twist condition is not empirically satisfied, but our sorting condition can be used to build a lower bound of the benefit of relocating farmers.

The results point in the direction that there is a significant mismatch of farmers and land. In our counterfactual results, we show that a better allocation of farmers to land would be equivalent to increasing the overall productivity by as much as 34%. This shows

a poor allocation of farmers to land, with much room to improvement.

Our paper contributes to four main topics in economics. First, we contribute to the recent works on multidimensional sorting that have been motivated by marriage and labor markets applications (Chiappori et al. (2012), Chiappori et al. (2016), Chiappori et al. (2017), Low (2014), Chiappori et al. (2020))[1]. By using a central planner's problem to define the optimal allocation, we find a sorting relation between farmers' productivity and the marginal expected land return. This result is in line with the model developed in Chiappori et al. (2017), where the authors consider a continuous multidimensional sorting problem. Nonetheless, we depart from their results in important ways. We show that in a discrete setting, when the output function is convex with respect to the productivity on the one dimensional side of the market, their results can be simplified and extended. Chiappori et al. (2017)'s results show that the existence of positive assortative matching and the characterization of the optimal allocation depend not only on the output function but also on the distribution of characteristics from both sides of the market, in an *edge of the knife* condition. In our setting this is no longer the case, as the distribution of characteristics is free to take any form and our sorting condition is always present.

Additionally, we show that imposition of the *twist condition* implies a *nested* structure found in Chiappori et al. (2016), that guarantees the optimality of the sorting. As noted in Chiappori et al. (2020), this need not to be the case in general, even for multi-to-one matching models.

Second, we contribute to the literature of misallocation (Restuccia and Rogerson (2008) and Hsieh and Klenow (2009)), more specifically misallocation in the agricultural sector (Bolhuis et al. (2020), Adamopoulos et al. (2017), Gollin and Udry (2021), Shenoy (2017), Gottlieb and Grobovšek (2019), Adamopoulos and Restuccia (2020), Restuccia and Santaeulalia-Llopis (2017)) and frictions in land markets (Chen et al. (2017) and Chari et al.

[1]Notice that a multi-to-one or m-to-n sorting is different from a many-to-many or m-to-m sorting as in (Lindenlaub (2017)), where the author considers firms that employ two types of skill and workers that have two types of skill. In that setting, sorting can be defined separately for each type.

(2017)). Dispersion of marginal productivity of inputs across farms is the usual measure of misallocation, since in an optimal equilibrium resources should be allocated to zero this dispersion. Nonetheless, the measurement of this dispersion is conditional on the crop choice, which in turn is conditional on the realized sorting between farmers and land. Our paper shows how discrete choice models can help introducing sorting to this literature, by uncovering the role that multidimensional sorting may have on misallocation.

Third, we contribute to the literature of technology adoption in the agricultural sector, as in Suri (2011), Griliches (1957), Griliches (1980), Conley and Udry (2010), Duflo et al. (2008), Duflo et al. (2011), and Pellegrina et al. (2019). The existence of a poor matching of farmers and land can be an important factor in explaining the non-adoption of apparently profitable technologies. In Suri (2011), for example, the author concludes that the low adoption rate of hybrid maize can be explained by heterogeneity of benefits and costs of the technology. Nonetheless, this result is again conditional on the actual matching of farmers and land, and therefore a different sorting can shift the distribution of heterogeneity, changing the adoption rate. Thus, the heterogeneity of costs and benefits can be, at least in part, the result of land market frictions that leads to the observable sorting.

Finally, our paper is related to the literature of environmental conservation and deforestation (Alix-Garcia et al. (2015), Jayachandran et al. (2017), Burgess et al. (2012)), particularly deforestation in the Amazon (Assunção et al. (2013), Assunção et al. (2015), Assunção et al. (2019), Araujo et al. (2020), Burgess et al. (2019), Stabile et al. (2020), Souza-Rodrigues (2019), Soares-Filho et al. (2006), Nepstad et al. (2014), Laurance et al. (2001), and Cochrane and Schulze (1998)). Our study area, the state of Mato Grosso, is a fast expanding agricultural frontier in the transition between the Cerrado biome to the Amazon biome. The link between agricultural productivity and conservation has been studied with different frameworks (Koch et al. (2019), Abman and Carney (2020a), Abman and Carney (2020b), Assunçao et al. (2017)). A better understanding of the determinants of agricultural productivity is key to conservation efforts.

The paper proceeds as follows. In section (2) we describe the discrete choice model where farmers choose among crops. We then describe how to incorporate this model inside a central planner's matching problem. We derive optimality conditions and discuss the counterfactual. In section (3) we discuss the estimation strategy. In section (4) we present the data set. In section (5) we show the results of the estimation and in (6) the results of the counterfactual. In section (7) we conclude.

2 Model

In this section, we formulate a discrete choice model in which every year a profit maximizing farmer chooses how to use each plot of land. The farmer can choose among a range of different crops and double-crops. Through the discrete choice model we obtain the output of a matching between farmer and land, which we incorporate into a central planner's matching problem. We derive our sorting condition and discuss when this condition is enough to characterize the optimal matching. Finally, we show how to run the counterfactual.

2.1 Discrete Choice Model

Environment

There are N farmers indexed by i with productivity $x_i \in X$. A farmer owns a plot of land (ℓ) that belongs to a finite and discrete set L. There is a set of available crops K. Each land ℓ is characterized by a vector of potential productivity for each crop denoted by $\{z_{\ell,k}\} \in \mathbb{R}^K$, a transportation cost to international markets (ports) for each crop $\{\tau_{\ell,k}\}$, and remaining factors such as precipitation, temperature, and slope of the terrain.

We differentiate between available crops $k \in K$ and the production choice of the farmer $\kappa \in \mathcal{K}$, where $\mathcal{K}$ is a set of subsets containing up to two elements of K. We do this to allow a farmer to choose to employ a double-crop system, where two elements of K can be

produced in the same agricultural year. For example, a double-crop system of soybeans followed by a mid-season corn. Our model is static, but we index the farmer's decision by each period t with $t = \{0, 1, ..., \infty\}$. Each choice $\kappa \in \mathcal{K}$ has a cost $c_{\kappa,t}$ of adoption. The selling price in international markets of the production of crop $k \in K$ in each period is given by $p_{k,t}$.

Choosing crops

At the beginning of each period t a farmer i matched with land $\ell \in L$ faces a discrete choice problem. Each choice $\kappa \in \mathcal{K}$ has a return given by:

$$\pi_{i,\ell,\kappa,t} = \alpha x_i \left[\sum_{k \in \kappa} z_{\ell,k}(p_{k,t} - \tau_{\ell,k}) \right] + \beta_\kappa X_{\ell,t} - c_{\kappa,t} + \epsilon_{i,\ell,\kappa,t} \tag{1}$$

Where: (1) the parameter α normalizes the variance of the idiosyncratic error[2]; (2) x_i denotes the farmer's productivity which, without loss of generality, we assume to be $x_i \in [0, 1]$; (3) β_κ is a vector of parameters that governs how the controls ($X_{\ell,t}$) - such as precipitation and temperature - affect the cost of adoption. These characteristics should affect the cost of planting and harvesting, so that we allow them to affect the return of choice κ. These land characteristics should also influence crop yields. Nonetheless, this effect should be fully captured by the potential yield measure ($z_{\ell,k}$). Note that we allow for the components of $X_{\ell,t}$ to affect each choice κ differently, through the vector of parameters β_κ; (4) $c_{\kappa,t}$ is a parameter that captures the cost of adopting the crop system κ; (5) The idiosyncratic shock $\epsilon_{i,\ell,\kappa,t}$, is drawn from a Frechet distribution (F_ϵ) i.i.d across farmers, choice, and time. This is a standard assumption in the discrete choice literature, that allows us to derive a closed form solution to the probability of a farmer making a choice. Each farmer i will choose a $\kappa \in \mathcal{K}$ in order to maximize profits

[2]As we explain in more detail later, the idiosyncratic shock is assumed to have a Generalized Extreme Distribution. It is standard in the literature of Discrete Choice Models to normalize the variance of the idiosyncratic shock to one. Therefore the parameter α is inversely proportional to the standard error of the shock. This normalization is innocuous to the model, since in a discrete choice model the level of the utility/profit does not matter for the choice decision. What matter is the difference between different decisions.

$$\max_{\kappa \in \mathcal{K}} \{\pi_{i,\ell,\kappa,t}\} \tag{2}$$

Given the distribution of idiosyncratic shocks we are able to compute the probability of choosing $\kappa \in \mathcal{K}$[3]:

$$P_{i,\ell,\kappa,t} = \frac{\exp\left(\alpha x_i \left[\sum_{k \in \kappa} z_{\ell,k}(p_{k,t} - \tau_{r,k})\right] + \beta_\kappa X_{\ell,t} - c_{\kappa,t}\right)}{\sum_\eta \exp\left(\alpha x_i \left[\sum_{j \in \eta} z_{\ell,j}(p_{j,t} - \tau_{r,j})\right] + \beta_\eta X_{\ell,t} - c_{\eta,t}\right)} \tag{3}$$

We can also compute the expected return from a matching of a farmer x_i with a land ℓ prior to the realization of the idiosyncratic error. This expected return represents the output of the matching (x_i, ℓ) in period t:

$$\begin{aligned} \pi_t(x_i, \ell) &= \mathbb{E}_\epsilon \left[\max_\kappa \{\pi_{i,\ell,\kappa,t}\}\right] \\ &= \ln\left(\sum_\kappa \exp\left(\alpha x_i \left[\sum_{k \in \kappa} z_{\ell,k}(p_{k,t} - \tau_{r,k})\right] + \beta_\kappa X_{\ell,t} - c_{\kappa,t}\right)\right) + \gamma \end{aligned} \tag{4}$$

Where γ is the Euler's constant.

2.2 The central planner's problem

The central planner allocates farmers - which is equivalent to allocating x_i - across available land in order to maximize the aggregate expected production. From the discrete choice model we obtain Expression 4, that gives the expected output of a matching between land ℓ with a farmer with productivity x_i in period t. Taking expectation with respect to crop prices and cost of technology adoption, both varying over the years, we define for each land a function that describes the expected return from land over the years for each possible matching:

$$\pi^e_\ell(x) = \mathbb{E}_t[\pi_t(x, \ell)] = \int_\Omega \pi_t(x, \ell) dP \tag{5}$$

[3]See Train (2009) for a derivation of Expressions 3 and 4.

where dP is the probability measure of prices and costs with support in Ω. Expression 5 is the function that the central planner considers, since it gives the expected output of the matching between land ℓ and a farmer with productivity x. We then define the following central planner's problem:

$$\max_{\{x_i\}} \sum_{\ell} \pi_{\ell}^{e}(x_i) \tag{6}$$

That is, the central planner allocates farmers (identified by its productivity x_i) to plots of land in order to maximize total expected production. A brute force approach to solve the central planner's problem would be computationally unfeasible, since we would need to check every possible allocation which amounts to $N!$ combinations. To the best of our knowledge, the only references about solving a similar problem are Chiappori et al. (2017) and Chiappori et al. (2016). The authors consider a continuous version of Expression 6, which they are able to solve uniquely, under some conditions, using optimal transportation theory (Monge (1781), Kantorovich (1942)).

In the one-to-one literature the Spence-Mirrlees (or supermodularity) condition usually guarantees that sorting is efficient (Chade et al. (2017)). Thus, an important first step is whether we can obtain an sorting relation as a result of optimality in a multi-to-one setting. In a standard one-to-one matching environment, e.g., marriage market and labor market, the usual result is a positive assortative matching relation between the one-dimensional characteristic of each side of the market. For example, worker's and firm's productivity. But, in our case of multi-to-one matching there is not a natural definition of sorting. Consequently, we need to start by defining what sorting means in this environment. As it turns out, in a multi-to-one matching, the sorting itself can be the characterization of the optimal matching.

This is exactly the case in Chiappori et al. (2017)'s approach, where a nested criterion defines a sorting relation, which is used to build the optimal matching. Nonetheless, for our application, the problem with Chiappori et al. (2017)'s approach is that the existence

of a positive assortative matching and the existence of a characterization of the optimal matching is very restrictive, depending on conditions that the authors characterize as being sharp. Precisely, it depends on specific conditions that constraint not only the output function (Expression 5 in our model), but also the distribution of characteristics on both sides of the market – this contrast with the one-to-one dimensional matching, in which the supermodularity condition is enough.

We then build on some particularities of our model to study the optimal matching. As we will show, our definition of sorting will be closely related to the definition in Chiappori et al. (2017). Nonetheless, the sorting conditions will depend only on the properties of the surplus function and will not be sharp conditions.

Necessary conditions for optimality

Our model presents the important property that the output function of a matching (x, ℓ) is described by a strictly convex function, $\pi^e_\ell(x)$. As this property will be important in what will follow, we state it here for further reference in Lemma 1.

Lemma 1. *Consider the farmer discrete choice model described above. Expected output of land is continuous, increasing and strictly convex with respect to farmers productivity*

$$\pi^{e\prime}_\ell(x) > 0 \quad \textit{and} \quad \pi^{e\prime\prime}_\ell(x) > 0.$$

Proof. See Appendix. ■

We connect the central planner's problem with sorting of farmers and land through two propositions. In both propositions we will make use of the *cross-difference* function, introduced by McCann (2012), that gives us the gain from relocating two farmers. Let x and x' be farmers matched with plots of land ℓ and ℓ', respectively. In our setting, the

cross-difference function is defined on $(X \times L)^2$ by:

$$\delta(x, \ell, x', \ell') = [\pi^e_{\ell'}(x) + \pi^e_{\ell}(x')] - [\pi^e_{\ell}(x) + \pi^e_{\ell'}(x')]. \tag{7}$$

Expression 7 gives the surplus from a relocation between farmers x and x'. Therefore, the planner would like to change farmers across lands whenever $\delta(x, \ell, x', \ell') > 0$.

Define a matching as a map connecting farmers to land $M : X \to L$. Our first proposition gives a necessary condition for optimality.

Proposition 1. *Let $\pi^e_\ell(x)$ be the output function of a matching between land $\ell \in L$ and a farmer with productivity $x \in X$. Consider that $\pi^e_\ell(\cdot)$ is continuous, increasing and strictly convex. If the matching M^* is optimal then for all $x, x' \in X$ such that $x > x'$ and $M^*(x) = \ell$, $M^*(x') = \ell'$, we have:*

$$\pi^{e\prime}_\ell(x) > \pi^{e\prime}_{\ell'}(x')$$

Proof. Optimality implies that for any match we have $\delta(x, \ell, x', \ell') \leq 0$:

$$[\pi^e_{\ell'}(x) + \pi^e_{\ell}(x')] - [\pi^e_{\ell}(x) + \pi^e_{\ell'}(x')] \leq 0 \Leftrightarrow \pi^e_\ell(x) - \pi^e_\ell(x') \geq \pi^e_{\ell'}(x) - \pi^e_{\ell'}(x')$$

$$\Leftrightarrow \frac{\pi^e_\ell(x) - \pi^e_\ell(x')}{x - x'} \geq \frac{\pi^e_{\ell'}(x) - \pi^e_{\ell'}(x')}{x - x'}$$

Then, as $\pi^e_\ell(\cdot)$ and $\pi^e_{\ell'}$ are continuous, there exist $x_1, x_2 \in (x', x)$ such that

$$\pi^{e\prime}_\ell(x_1) = \frac{\pi^e_\ell(x) - \pi^e_\ell(x')}{x - x'} \geq \frac{\pi^e_{\ell'}(x) - \pi^e_{\ell'}(x')}{x - x'} = \pi^{e\prime}_{\ell'}(x_2)$$

As $x > x_1$ and $x_2 > x'$, by convexity

$$\pi^{e\prime}_\ell(x) > \pi^{e\prime}_\ell(x_1) \geq \pi^{e\prime}_\ell(x_2) > \pi^{e\prime}_{\ell'}(x')$$

■

This proposition gives us a sorting relation between x_i and $\pi_\ell^{e\prime}(x_i)$ that must be satisfied in the optimum. This result is in order with Chiappori et al. (2016), in the sense that a more productive farmer will be matched with a land in which he gives a higher marginal benefit. If we were to consider a land market, the land ℓ would be more willing to pay (or to give a higher share of the output from the match) for a more productive farmer x to be installed there, as the land benefits more from an increase in productivity.

Notice that, with Proposition 1, the convexity of the output function gives us a necessary condition for the optimum and that this condition does not depend on the distribution of characteristics of land and farmers. With this proposition we start to move back to a one dimensional setting, where sorting can be defined as the relation between two variables. Although the cardinal relation of x_i and $\pi_\ell^{e\prime}(x_i)$ can be very complicated, the ordinal relation should be one to one in the optimal matching. For now, the optimal allocation of farmers to land implies perfect sorting between x_i and $\pi_\ell^{e\prime}(x_i)$, however, it is not a sufficient condition. Even if an estimation of this sorting relation were close to one, it would not mean that the allocation is close to the optimum.

Sufficient conditions to constrained optimality

We adapt to our setting the definition of Chiappori et al. (2016) of the *twist* condition and show that this condition is sufficient to characterize the optimal allocation.

Definition 1. *The function $\pi_\ell^e(\cdot)$ satisfies the twist condition if for any $\ell, \ell' \in L$ such that $\ell \neq \ell'$, we have:*

$$\pi_\ell^{e\prime}(x) \neq \pi_{\ell'}^{e\prime}(x)$$

for all $x \in X$.

This condition is equivalent to the *injectivity* of land – $\ell \in L$ – and the derivative of the output function – $\pi_\ell^{e\prime}(x)$ – for each $x \in X$. This means that the derivatives of the land expected return do not cross each other. In our model, the *twist* condition implies that

each land can be uniquely characterized by its marginal returns to an increase in farmers productivity.

To reach our sufficient condition, we start with the following lemma and proposition, where we guarantee optimality in subsets of $X \times L$.

Lemma 2. *Consider any subset $X' \times L' \subseteq X \times L$. Let x_m be the most productive farmer in X'. If there is a land $\ell_m \in L'$ such that $\pi^{e\prime}_{\ell_m}(x) > \pi^{e\prime}_{\ell}(x)$ for all $(x, \ell) \in X' \times L'$, then x_m should be matched with ℓ_m in order to maximize the expected return in this subset of lands.*

Proof. We just need to check that once x_m is matched with ℓ_m, any relocation in $X' \times L'$ would lead to a loss in expected return, that is $\delta(x_m, \ell_m, x, \ell) < 0$.

$$\begin{aligned}
\delta(x_m, \ell_m, x, \ell) =& [\pi^{e}_{\ell_m}(x) + \pi^{e}_{\ell}(x_m)] - [\pi^{e}_{\ell_m}(x_m) + \pi^{e}_{\ell}(x)] \\
=& [\pi^{e}_{\ell}(x_m) - \pi^{e}_{\ell}(x)] - [\pi^{e}_{\ell_m}(x_m) - \pi^{e}_{\ell_m}(x)] \\
=& \int_{x}^{x_m} \pi^{e\prime}_{\ell}(s)ds - \int_{x}^{x_m} \pi^{e\prime}_{\ell_m}(s)ds = \int_{x}^{x_m} (\pi^{e\prime}_{\ell}(s) - \pi^{e\prime}_{\ell_m}(s))ds < 0
\end{aligned}$$

■

The lemma above tells us that if we have a plot of land in which the marginal benefit from increasing farmer's productivity is the highest for every farmer in X', then the most productive farmer in this subset should be matched with this plot of land. This plot of land is uniquely determined by its derivative. If we are able to build subsets of land in which land is uniquely identified by their derivatives, then land may be sorted according to their derivatives. A characterization of optimal conditions in this subset is given in Proposition 2.

Proposition 2. *Suppose there exists a non-empty set $L' \subseteq L$ such that for every $\ell, \ell' \in L'$ the functions $\pi^{e\prime}_{\ell}(x) \neq \pi^{e\prime}_{\ell'}(x)$ for all $x \in X$, that is, the derivatives do not cross (the twist condition is satisfied). Take any set of farmers $X' \subseteq X$, such that the cardinality of X' equals the cardinality of L'. Then, the matching M in the subset $X' \times L'$ is optimal if and only if for all $x, x' \in X'$ such*

that $x > x'$ *and* $M(x) = \ell$, $M(x') = \ell'$, *we have:*

$$\pi^{e}_{\ell'}(x) > \pi^{e\prime}_{\ell'}(x')$$

Proof. The proof follows from Lemma 2. In the subset L' we can always manage to sort land according to their derivatives. From Lemma 2, the most productive farmer should be matched with the land with the highest derivative. Then, we consider the subset minus this matching and we apply the lemma iteratively until all the farmers are matched with a land. The resulting matching is the only one that satisfy the condition that the most productive farmer matched with the land which he brings the highest return. As any relocation from this match would fail to satisfy this condition, our result is necessary and sufficient in any subset of lands in which the derivatives do not cross. ■

Proposition 2 tells us that under the hypothesis that the derivatives of the surplus function do not cross each other, our sorting condition is necessary and sufficient. If it were the case that $L' = L$, then our proposition shows that the existence of positive assortative matching is neither a sharp condition, nor depends on the distributions of both sides of the market. Furthermore, the characterization of the optimal allocation is entirely defined by the sorting relation. Different from Chiappori et al. (2017) that uses optimal transportation theory to solve their matching problem, here we are able to solve a similar problem with basic calculus, given the convexity of the output function.

In the next subsection we show how we can use Propositions 1 and 2 to build a candidate for the optimal allocation of farmers. This candidate can be used to build a lower bound on the benefit of better allocating farmers when only the necessary condition is present, that is, when the *twist* condition is not satisfied. When the *twist* condition is satisfied the candidate is the only candidate and therefore is the optimal one.

2.3 Counterfactual

An algorithm to build a counterfactual allocation. We define the candidate allocation $M^c(x)$ as the allocation generated by the following algorithm.

1. Initialize the set of available farms $\bar{L} = L$.
2. For each $x_i \in X$, starting from the most productive to the least one
   (a) Compute $\pi_\ell^{e\prime}(x_i)$ for all $\ell \in \bar{L}$
   (b) Create a matching of x_i with land ℓ that gives the highest computed derivative
   (c) Eliminate the matched farm from $\bar{L}$.

This algorithm matches a farmer with the highest possible derivative, given that all farmers that are more productive have already been matched. It guarantees that the resulted allocation will full fill the condition of Proposition 1 and the condition for Proposition 2 in any possible subset of lands in which the derivatives do not cross.

Measuring the benefit of relocation. After computing the candidate allocation, we compare the average product in the two scenarios: with the estimated allocation of farmers and the counterfactual one. To empirically assess the gain from the relocation, we compute an equivalent increase in the overall productivity (Δ) that would result in the same gain in average return given by our new candidate matching. Denoting by x_{i^c} the counterfactual productivity matched with land ℓ, we wish to find the constant $\Delta > 0$ that solves

$$\frac{1}{T}\sum_t\sum_\ell \pi_{\ell,t}(\Delta x_i) = \frac{1}{T}\sum_t\sum_\ell \pi_{\ell,t}(x_{i^c}) \tag{8}$$

3 Estimation

We derive a likelihood expression that allows us to estimate the parameters of the discrete choice model using standard maximum likelihood techniques. Denoting by $\mathbb{I}_{\ell,\kappa,t}$

the indicator function of land ℓ being observed with choice κ in time t, we can write the log-likelihood criterion function as

$$\mathcal{L}(\theta) = \sum_{t} \sum_{\ell} \mathbb{I}_{\ell,\kappa,t} \log P_{i,\ell,\kappa,t} \tag{9}$$

Here, θ denotes the set of parameters to be estimated. This set can be divided in two groups. The first one is composed of the parameters $\{\{\beta_\kappa\}, \{c_{\kappa,t}\}\}$, that is, the parameters of controls and adoption costs. The second group controls the spatial distribution of x_i, which we further explain now.

We do not observe farmer's productivity x_i. Nonetheless, with the discrete choice model we recover the spatial distribution of x_i that rationalizes the choices observed in the data. Ideally, we would like to estimate a non-parametric spatial distribution of x_i. Nonetheless, this would be unfeasible since it would amount to estimate hundreds of thousand of objects. For that reason, we fit a flexible function that parameterizes the spatial distribution of x_i. Formally, let the variables lat_ℓ and lon_ℓ denote location variables – in a two dimensional plane – of ℓ. Then, for each x_i matched with a ℓ we parameterize the farmer's productivity as

$$\alpha x_i = \left(\sum_{n=1}^{\eta} (h_n lat_\ell^n + v_n lon_\ell^n) \right)^2 \tag{10}$$

Where η determines the order of the polynomial that we fit. This parametrization does not mean that farmer's productivity is an inherit characteristic of the land. After the matching of farmers and land is completed, this matching generates a spatial distribution of farmers and, as a result, a spatial distribution of farmers' productivity. Our objective with the expression above is to use a function that can flexibly fit possible spatial distributions. By using a big enough η, we will be using higher order polynomials which can approximate any function. We take the polynomial to a square power to constraint αx_i to be positive, since a negative value would not have an interpretation in our model.

The functional form that we fit is a continuous one. This is necessary because we estimate the model using the maximum likelihood estimator, which assumes that the likelihood function varies continuously with the parameters. Nonetheless, this functional form does not imply that the distribution of farmer's productivity we estimate is continuous, since in our empirical specification our location variables are defined by the observed units, which are non-continuously distributed in space. For the estimation, it means that the function in Expression 10 can freely adapt in regions of the domain that we do not observe a unity, in order to better fit the function in regions of the domain where we do observe a unit.

Also, it is important to highlight that we parametrize αx_i instead of just x_i. The reason is that we cannot separately estimate α and the parameters of the polynomial in Expression 10. Nonetheless, this restriction does not affect the counterfactual, since Expression 8 do not rely on the estimation of α. Notice also that the parameters of the polynomial is fixed across time, that is, we assume farmers do not move. We decided to take this as our main specification because when estimating the model separately year by year we found a correlation above 0.97-0.99 for the ranking of farmers' productivity across all years.

An important caveat on the identification of farmer's productivity: the parametrize distribution of x_i estimates a residual heterogeneity that we interpret as farmer's productivity. The variable x_i could well be composed of other important factors, such as, measurement error, unobservable heterogeneity of land productivity, and local institutions[4]. As in our data we cannot identify farmers nor can see farmers moving, it is not possible to identify what portion of x_i can actually be traced back to farmer's productivity. Nonetheless, we see our data serving as an application to motivate the theory of multidimensional sorting, in a setting where multidimensionality has a natural application, given that the same plot of land can be used to produce different crops.

Finally, we can define define θ the vector of parameters to be estimated as $\theta = \{\{\beta_\kappa\}, \{c_{\kappa,t}\}\} \cup$

[4]Gollin and Udry (2021) is an example of a paper that deals with disentangling these different sources of heterogeneity.

$\{\{h_n\}, \{v_n\}\}$. That is, our estimation recovers θ^*

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_t \sum_\ell \mathbb{I}_{\ell,j,t} \log P_{i,\ell,j,t} \tag{11}$$

4 Data

Our study area is the state of Mato Grosso, Brazil. This state, illustrated in Figure 1a, is a global agricultural hub responsible for 10% of the global production of soybeans. At the same time, it is home for three different biomes: the Cerrado - a Savannah like biome that covers 38% of the territory - the Amazon - which covers 50% - and the Pantanal - wetlands that covers 12% of the territory. The state is in the transition from the Cerrado to the Amazon, and has become a hotspot of deforestation. It is located in a region denominated the Arch of Deforestation, a frontier of rapid expansion of agricultural land in Brazil, advancing towards the Amazon. Almost 21% of its territory is demarcated as a protected area, such as, conservation units and indigenous land.

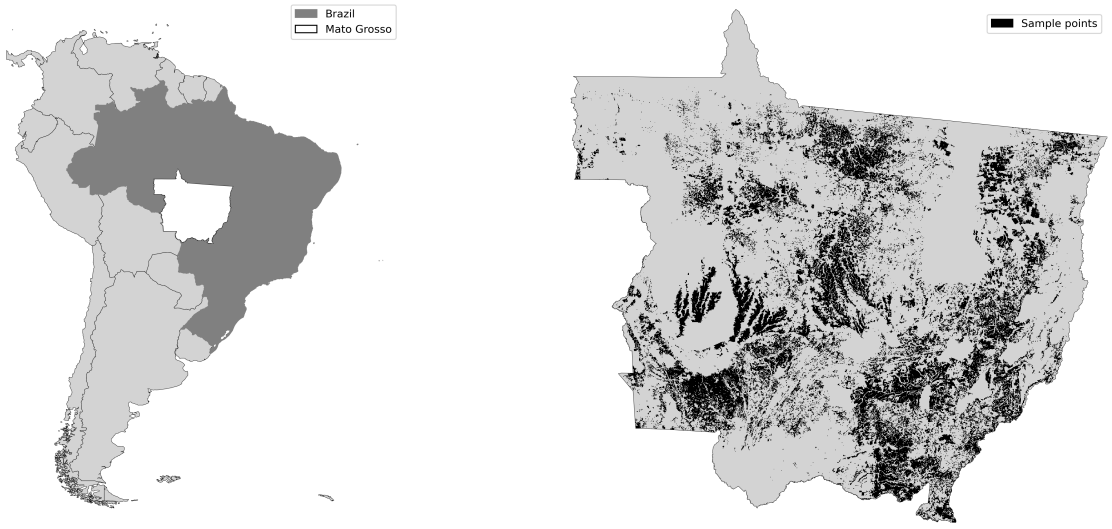
In order to estimate the model we need data on potential land productivity, crop prices, transportation cost, land use and additional control variables such as precipitation, temperature, and slope of the terrain.

Land use choice is from Simoes et al. (2020). This data set classifies each pixel of 250m resolution in the state of Mato Grosso from 2003 up to 2017 in several categories: pasture, idle, soy and corn, soy, soy and cotton, fallow and cotton, soy and sunflower, and sugarcane.[5] We sample the data at a resolution of 1 km for computational reasons. We exclude from the sample pixels that lie inside protected areas - e.g., conservation units and indigenous land - since land use in those areas are subject to a complete different legislation[6].

[5]The data set described in Simoes et al. (2020) starts at 2001. We exclude 2001 and 2002 from our estimation because prior to 2003 sugarcane had not been introduced in the state of Mato Grosso. We also collapse the land use soy and millet to only soy, since millet is not marketed.

[6]The polygons of protected area is from the Ministry of Environment and can be accessed at https://antigo.mma.gov.br/areas-protegidas/cadastro-nacional-de-ucs/dados-georreferenciados.html

Figure 1: South America - Brazil - Mato Grosso - Sample



(a) South America - Brazil - Mato Grosso

(b) Sample points

The map on the left shows the regions of South America (light gray), Brazil (gray), and the state of Mato Grosso (white). Our study area is the state of Mato Grosso. The map on the right shows our sample points inside the state of Mato Grosso, after excluding protected areas and pixel that were not deforested before 2001.

We also exclude pixels that were classified as native vegetation in 2001, prior to our first year in the sample. We do this because modeling the drivers of deforestation is out of the scope of this paper. Figure 1b shows the location of our sample points.

In Table 1, columns 1-8, we present a transition matrix for the land use data. Each row-column shows the proportion of pixels that moved from row to column, across all the years. Pasture grazing, sugarcane and fallow are the most persistent activities, as evidenced by a total of 96%, 94% and 88% pixels in these activity not moving to any other land use. Indeed, sugarcane is a semi-perennial crop, which should partially explain this behavior.[7] On the other hand, the persistence in fallow suggests that a great portion of land is abandoned, instead of being used to rotate crops. Soy and sunflower see a small proportion of pixels coming into the activity. Among cotton, soy and corn, soy, and soy and cotton there is a much higher degree of transition. In our model, these transitions

[7]For a more in depth study of the economics of sugarcane in Brazil, see Sant'Anna (2017)

Table 1: Transition matrix and proportion of land use (%)

	Cotton	Beef	Soy Corn	Soy Cotton	Soy	Soy Snflwr	Sugar	Fallow	Prop.
Cotton	40	11	21	16	13	0	0	0	1
Beef	0	94	1	0	3	0	0	1	69.7
Soy-Corn	2	6	72	4	17	1	0	0	10.7
Soy-Cotton	7	5	25	56	5	0	0	0	1.2
Soy	1	16	24	1	56	0	0	0	11.3
Soy-Snflwr	2	3	59	3	28	4	0	0	0.1
Sugar	0	2	1	0	1	0	96	0	0.9
Fallow	0	12	0	0	0	0	0	88	5.2

This table presents in columns 1-8 a transition matrix for the land use data across the years. Each row-column shows the percentage of pixels that transited from land use row to land use column. In column 9 the table shows the proportion of each land use with respect to all the sample. The land use data is from Simoes et al. (2020). Protected areas data is from the Ministry of Environment.

are explained by market prices for the product, as well as varying costs of adoption. In column 9 of Table 1 we show the proportion of each land use in our sample. Most of land is devoted to pasture grazing, followed by a single-crop of soybeans and a double-crop of soybeans and corn.

The potential soil suitability for each crop data - variable z_c in the model - is from the Global Agro-Ecological Zones of the Food and Agriculture Organization of the United Nations project (FAO GAEZ)[8]. We retrieve this information for pixels inside the state o Mato Grosso for corn, soybeans, cotton, sunflower, and sugarcane. Importantly, this dataset do not depend on farmers decisions. It is built only with information related to climate and soil characteristics. We convert the production of sugarcane to sugar using a technology parameter of 0.16 and convert sunflower to sunflower oil using a technology parameter of 0.42[9]. Table 2 shows descriptive statistics for the soil suitability variable. For the activity of pasture grazing, we assign a constant productivity that equals the average productivity of pasture in the state o Mato Grosso of 54 kg of beef per hectare. This data comes from

[8]Available at http://www.fao.org/nr/gaez/en/

[9]Both parameters is from Brazilian Agricultural Research Corporation (Embrapa). The conversion parameter for sunflower is from www.embrapa.br/girassol and the parameter for sugarcane is from www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_109_22122006154841.html

Table 2: Descriptive statistics.

	temp.	prec.	slope	corn	soybeans	cotton	sugar	snflwr
				Soil suitability (ton/ha)				
mean	0.987	0.478	0.909	5.980	3.833	0.514	1.503	0.667
std	0.006	0.088	0.795	1.252	0.237	0.121	0.156	0.335
25%	0.987	0.419	0.427	5.275	3.664	0.415	1.383	0.478
50%	0.987	0.473	0.744	5.449	3.821	0.534	1.549	0.639
75%	0.990	0.527	1.149	5.841	3.981	0.587	1.609	0.924
				Transportation cost (Brazilian reais)				
mean				132.209	131.587	169.376	138.680	153.052
std				15.376	15.676	15.870	14.480	16.365
25%				121.832	121.006	158.665	128.907	142.007
50%				132.437	131.818	169.610	138.894	153.294
75%				143.753	143.356	181.290	149.551	165.338

This table shows descriptive statistics for soil suitability of each crop (in tons by hectare), transportation cost for each crop (in Brazilian reais of 2008), and remaining variables: temperature (temp.), precipitation (prec.), and slope of the terrain (slope). The statics are mean, standard deviation (std) and the percentiles 25%, 50%, and 75%. Data of soil suitability is from FAO GAEZ. Transportation cost data is from Araujo et al. (2020), Ministry of Transportation, and Esalq. Temperature and Precipitation is from Hersbach et al. (2020), slope data is from Farr et al. (2007).

the 2010 Census of Agriculture.[10]

Average crop prices for each year are from the Federal Reserve Economic Data (FRED). We collect international prices for crops and beef across the years, and convert it to Brazilian reais (R$) using the average exchange rate of that year. Table 3 shows descriptive statistics for the crop prices. Data for the average precipitation and temperature in the agricultural year is from Hersbach et al. (2020) and data on the slope of the terrain is from Farr et al. (2007). Table 2 shows descriptive statistics of the geographical variables.

Transportation cost is from Araujo et al. (2020), but we expand their data of soybeans and corn to the rest of the crops in our land use data. The transportation cost data is built using georefenced data on federal and states roads, Brazilian ports and waterways from

[10] Available here https://www.ibge.gov.br

Table 3: Descriptive statistics for crop prices

	Soybeans	Corn	Sugar	Cotton	Sunflower Oil	Beef
mean	733.37	356.26	590.05	3,352.32	2,125.35	6,596.95
std	119.45	58.86	93.92	706.44	519.93	1,436.25
25%	649.98	310.84	535.92	2,810.26	1,805.58	5,433.35
50%	712.19	339.40	577.11	3,114.91	2,052.83	6,862.50
75%	824.59	410.86	669.19	3,739.11	2,411.81	7,535.67

This table shows descriptive statistics for crop prices measured as Brazilian reais per ton across years. The statics are mean, standard deviation (std) and the percentiles 25%, 50%, and 75%. The data is from Federal Reserve Economic Data (FRED).

the Ministry of Transportation[11].

This transportation network data is combined with data on transportation cost by road of different products from the Group of Research and Extension in Agroindustrial Logistics of the College of Agriculture Luiz de Queiroz (Esalq)[12], which provides estimated transportation cost per ton of each product (corn, soybeans, cotton, beef, sugar, and soy oil[13]) between multiples Brazilian municipalities for the years of 2008-2013. We apply Dijkstra's shortest path algorithm on the raster of roads to fit the model described in Expression 12

$$cost_{i,j,c,t} = \alpha_c + \beta_c cost\_raster_{i,j,c} + \epsilon_{i,j,c,t} \tag{12}$$

Where $cost_{i,j,c,t}$ denotes the monetized cost of transportation of one ton of product c between municipalities i and j in year t and $cost\_raster_{i,j,c}$ denotes the computed cost of transportation from the raster – using the cost for each transportation mode from Araujo et al. (2020) – between the centroids of municipalities i and j for product c. Finally, we apply Dijkstra's algorithm to compute the raster cost from every sample point of our data to the nearest final port. This raster cost is transformed to a monetary value through the

[11] Available at https://www.gov.br/infraestrutura

[12] This data is available at https://sifreca.esalq.usp.br/

[13] We do not have information on transportation cost of sunflower oil. Nonetheless, given that this product is very similar to soy oil, we use the transportation cost of soy oil as the transportation cost of sunflower oil

Table 4: Regression results of transportation cost

Product	Soybeans	Corn	Cotton	Sunflower	Beef	Sugar
Cost	0.070115	0.068772	0.070980	0.073196	0.092168	0.064765
from raster	0.000391	0.000428	0.003279	0.002624	0.003533	0.000884
Constant	10.47	13.41	46.76	26.61	68.12	26.802
	0.39	0.52	5.61	2.22	7.80	0.68
# obs	4147	2557	191	166	258	1000
R^2	0.88	0.90	0.71	0.82	0.72	0.84

This table shows regressions for each product. The unit of measure of the independent variable is R$(2008)/ton, that is, Brazilian local currency as of 2008 for each ton of product. Note that, I do not have data on transportation cost of sunflower product, so instead I use data of soybeans oil, a product that shares similar characteristics and specificities of transportation. Data is from Araujo et al. (2020), Ministry of Transportation, and Esalq.

fitted model 12. Table 4 shows the results for the regression of each product and Table 2 shows descriptive statistics of the transportation cost variable.

The average spatial distributions of the net revenue $z_k(p_{k,t} - \tau_k)$ for different k's are illustrated in Figure 2. We see that there is no clear spatial pattern of net revenue among different crops. This is the main reason for which a multidimensional model is necessary. If it were the case that all the different revenues were strongly spatially correlated than it would be possible to simply order land from the one with highest revenue of all to the one with the lowest revenue of all.

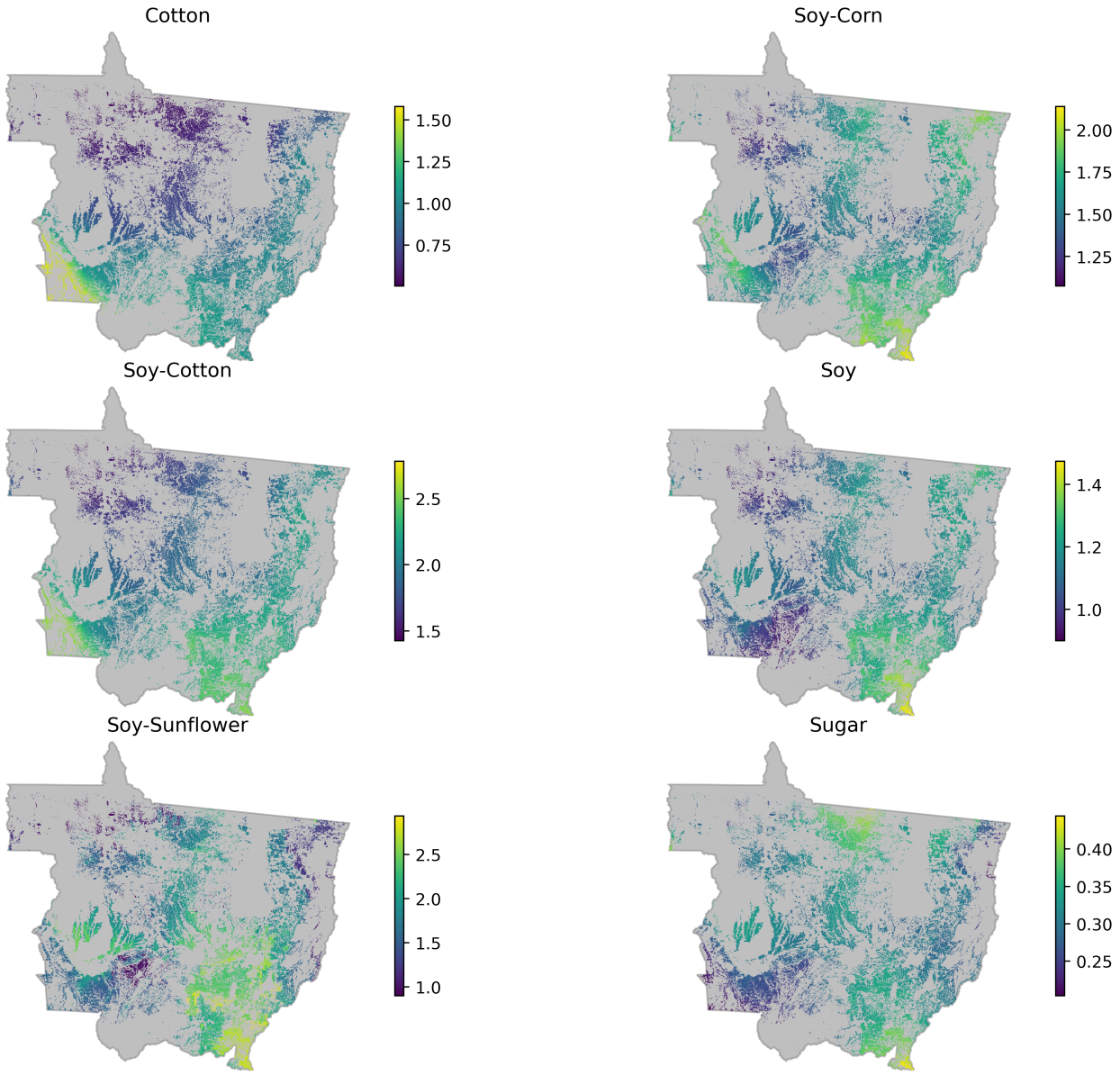
5 Estimation results

We estimate Expression 11 by maximum likelihood. To easy visualization we split the estimation results in three parts: the cost variables $(c_{k,t})$; the polynomial that characterizes the spatial distribution of farmer's productivity (x_i); and the control variables (β_κ).

Cost variables. Figure 3 plots the coefficients $c_{\kappa,t}$ for every crop and year. Although we estimate some variation across time, the order of the coefficients remains stable.

The two double crop systems of Soy-Sunflower and Soy-Cotton have the highest costs

Figure 2: Average values of net revenue



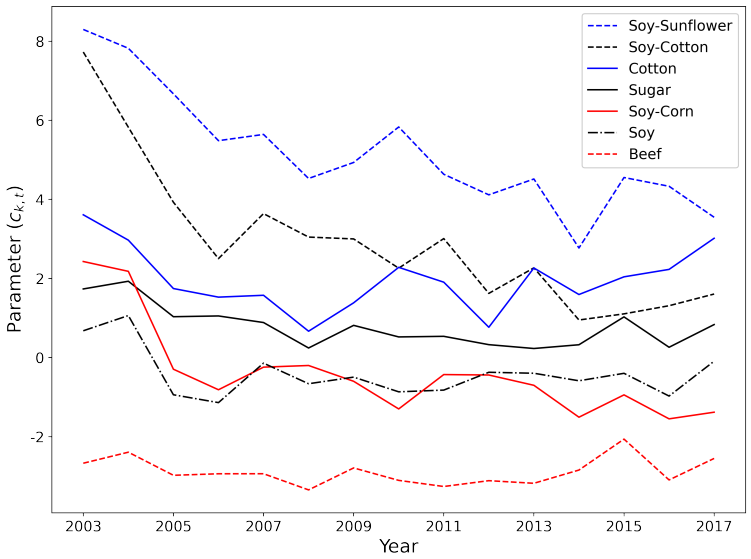
These maps shows the spatial distribution of $z_{\ell,k}(p_{k,t} - \tau_{r,k})$, which is the potential revenue of crop system k in time t net of transportation cost. The unit of measure is one thousand Brazilian reais per hectare, as of 2008. Data is from the FAO GAEZ project, FRED, Araujo et al. (2020), Esalq and Ministry of Transportation .

of adoption for almost the entire series. Nevertheless, for the final period, we see that the cost of adopting a single crop of cotton surpasses the cost of adopting the double crop soy-cotton. This can be rationalized by possible benefits that a crop can have on the amount of inputs used for a subsequent crop. The same inversion of costs is found in the adoption of

the double crop soy-corn and the single crop soy. In the final period, the cost of the double crop system is smaller than the single crop.

All those estimates are relative to the cost of leaving the land fallow, since we normalized $c_{fallow,t} = 0$. Therefore, a negative coefficient indicates that keeping some activity in the plot of land may be a cost effective way of mitigating risks involved in leaving the land idle, such as different taxes and the possibility of encroachment. Table 6 in the Appendix shows the estimated values and standard deviations of $c_{\kappa,t}$.

Figure 3: Adoption cost for each crop-system

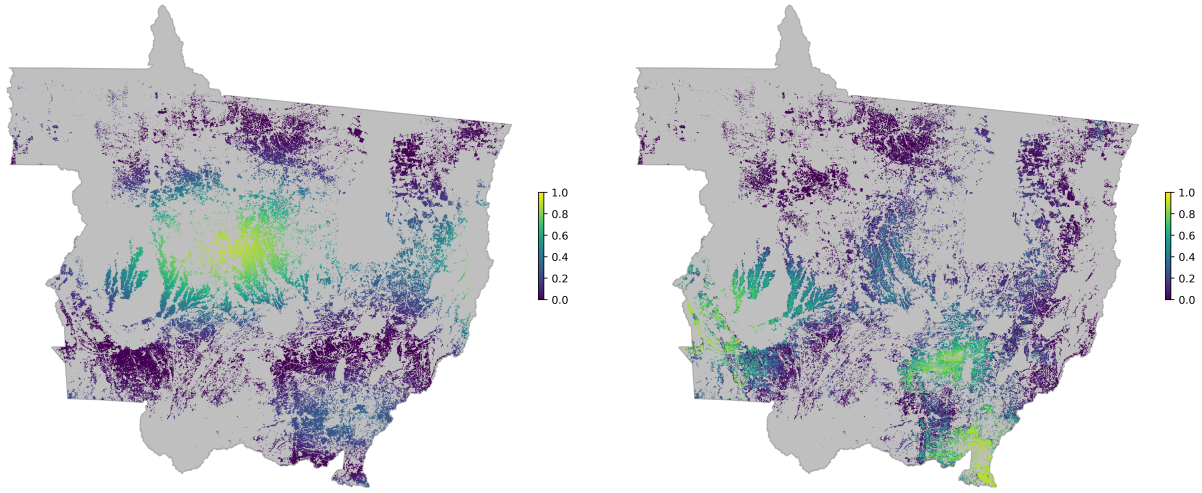


Estimated values of $c_{k,t}$ for each choice and year. Due to the size of our sample, the standard deviations are too small to be seen for most crops, therefore we do not show it here. See Table 6 in the Appendix for the values.

Productivity parameters. Table 7 in the Appendix presents the estimates of the polynomial that parametrize the spatial distribution of x_i defined in Expression 10. It is difficult to interpret these estimates, since it is just a parametrization of a spatial distribution. Therefore, in Figure 4a we plot the estimated polynomial surface evaluated at our sample points. This distribution shows a high concentration of high values of x_i in the center of the state, even tough there are high values of x_i also on the south-east and east regions.

Comparing Figure 4a with Figure 2 we see that there is no clear pattern of correlation between x_i and $\tilde{z}_k$. This is reassuring, since a strong spatial correlation between the two

Figure 4: Spatial distribution of x_i



(a) Actual spatial distribution (b) Counterfactual spatial distribution

The on the left map shows the estimated spatial distribution of αx_i as described by Expression 10. The map on the rigth shows the spatial distribution of αx_i. We normalized the values so that the highest αx_i equals one.

figures could indicate that the estimated x_i is actually only a higher resolution land heterogeneity that we do not capture correctly. The lack of a clear spatial correlation indicates that observable land heterogeneity would need to be significantly large in order to be the major explanation of our estimated distribution of farmer's productivity. Furthermore, the concentration of farmers with estimated high productivity in the center of the state implies that, if it were the case that we are only estimating land observable heterogeneity, then our data should have an strong skewed bias toward the center of the state.

Controls variables. Table 5 presents the estimates of the parameters β_k, that is, the parameters that creates spatial heterogeneity in the adoption cost. Conditional on land suitability and transportation cost, an increase in temperature and precipitation from the mean and an increase in the slope of the terrain negatively affects the return of the crop system. The only exception is that an increase in temperature, from its mean, increase the return of the pasture activity. These results are consistent with Spangler et al. (2017) and Cohn et al. (2016).

Table 5: Estimated coefficients of temperature, precipitation, and slope

Land use (κ)	Temperature	Precipitation	Slope
Cotton	-548.647	-16.931	-1.305
	(2.984)	(0.118)	(0.01)
Beef	4.023	-7.297	-0.157
	(1.289)	(0.034)	(0.002)
Soy-Corn	-392.524	-11.673	-1.465
	(1.625)	(0.049)	(0.005)
Soy-Cotton	-695.571	-12.06	-1.963
	(2.67)	(0.091)	(0.01)
Soy	-314.494	-12.155	-0.644
	(1.529)	(0.045)	(0.004)
Soy-Sunflower	-249.436	-15.106	-1.345
	(6.996)	(0.289)	(0.035)
Sugar	-163.008	-10.579	-0.879
	(2.371)	(0.092)	(0.011)

This table presents the estimation results for the parameters that govern how the controls affect the cost of adoption (β_k). Values in parentheses are standard errors computed via Delta method.

6 Counterfactual results

In our counterfactual exercise we want to relocate farmers (x_i) among plots of land (ℓ) in order to increase the aggregate production, as explained in subsection 2.3. To build a lower bound of how far the actual allocation is from the optimal one we relocate farmers as described in the counterfactual section, that is, we use our candidate allocation that fulfills the conditions of propositions 1 and 2. Recall that this is a lower bound, since we do not see the twist condition being satisfied by our data. We find that the relocation of farmers from the actual allocation to the candidate one would generate an increase in production equivalent to increasing the overall productivity by 34%, that is, from Expression 8 we have that $\Delta = 1.34$. In other words, a relocation of farmers would have the same effect on aggregate production as of increasing everyone's productivity by 34%. This result does

not take into account the possibility of relocating farmers to plots of land that have never been used to agricultural production, since our sample only have plots of land that was already being used for agriculture at least since 2001. That way, the relocation of farmers increases production without generating deforestation and/or abandonment of previous deforested areas.

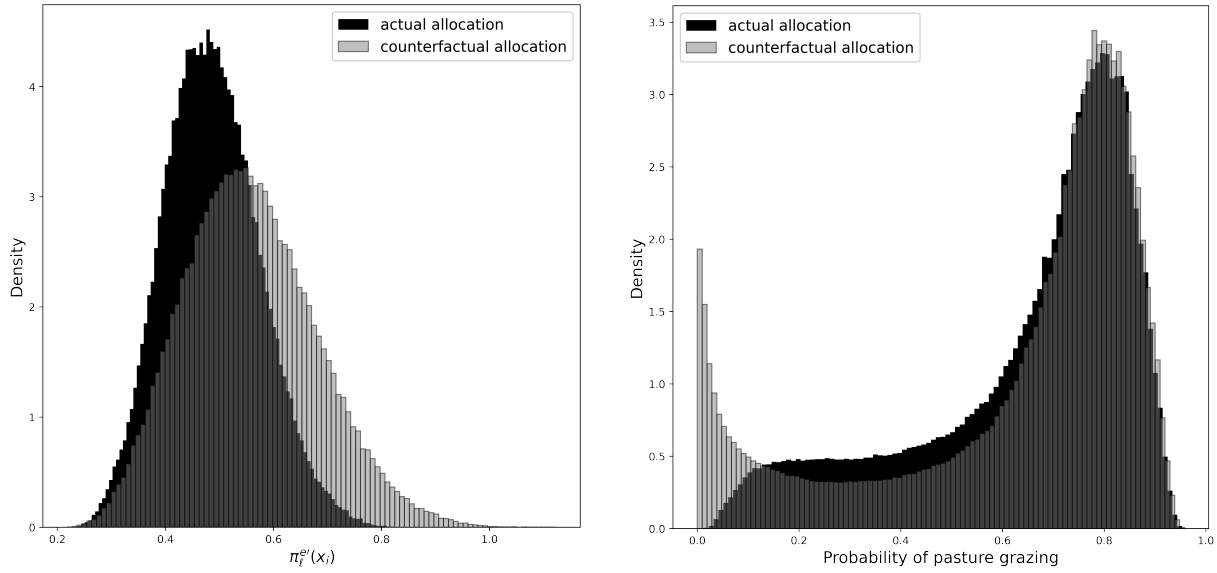
In Figure 4b we show the spatial distribution of x_i in our counterfactual allocation. Compared with Figure 4a, where we have the spatial distribution of x_i for the actual allocation of farmers, we see that an increase in aggregated production can be achieved by spreading farmers with higher productivity to the southeast and southwest regions.

The increase in production is the result of an first order stochastic dominant increase in marginal productivity of the economy. In Figure 5a we plot the distributions of $\pi_\ell^{e\prime}(x_i)$ - the marginal expected return - in the two alternative scenarios. The relocation of farmers considerably shifts the distribution of the marginal returns to the right. The effect of a relocation of farmers can be seem as the result of two components. One is the direct effect of a change in the productivity across all crops due to the change in x_i; the other component is the change in the probability of crop-system adoption due to the change in x_i. In the counterfactual scenario, a substantial portion of land sees a decrease in the probability of using land for pasture grazing, as illustrated in Figure 5b, where we compared the distribution of this probability in the actual and counterfactual scenarios. This decrease in the probability of pasture grazing activity is compensated by marginal increases in the probability of other crop systems, particularly Soy-Corn, Soy-Cotton, and Soy.

7 Conclusion

In this paper we studied sorting of farmers and land in a multidimensional setting. Farmers are characterized by a one dimensional space of productivity, while land is characterize by a multidimensional vector of characteristics, such as land productivity and transporta-

Figure 5: Distribution in actual and counterfactual scenarios



(a) Distribution of $\pi_\ell^{e\prime}(x_i)$

(b) Distribution of pasture

On the left, histogram for the $\pi_\ell^{e\prime}(x_i)$ variable in the estimated spatial distribution of x_i and in the counterfactual one. On the right, histogram of the probability of adoption of the pasture grazing activity.

tion cost. We explored a discrete choice model over crop systems to recover a residual heterogeneity which we interpret as being farmer's productivity. We then incorporated this discrete choice model inside a Central Planner's problem that wants to maximize the expected aggregated production by moving farmers around.

The Central Planner's problem gives us a sorting condition that must be full filled in an optimal allocation of farmers. This sorting condition connects farmer's productivity with the marginal expected return of the land for which this farmer is matched with. We showed that this sorting condition is sufficient to characterize the optimal allocation when the *twist* condition is satisfied. We also discussed that in the absence of the *twist* condition, our sorting condition can be used to compute a lower bound of the benefit of relocating farmers. After estimating the discrete choice model, we showed that a counterfactual allocation of farmers in the state of Mato Grosso could increase the overall agricultural productivity by at least 34%.

The theoretical results rely on the property that the surplus function is convex with

respect to the productivity of the market side defined by a one dimensional space. This property arises from the use of the discrete choice model. Therefore, applications that do not possess the convexity property is in need of further research to develop a full characterization of the optimal sorting.

Our empirical application introduces the possibility of employing discrete choice models as an alternative approach to discuss misallocation. An important caveat is our interpretation of the residual heterogeneity as being the farmer's productivity. We do not explore alternative data sets that would allow us to better disentangle farmer's productivity from potential (and likely) measurement errors, unobservable heterogeneity of land productivity, and local institutions. This is likely to overstate the benefit of relocating farmers, since relocation cannot be used to move measurement errors or unobservable heterogeneity of land productivity across plots of land. Therefore, further research is needed, with different applications and data, in order to incorporate discrete choice models and multidimensional sorting with a more complete estimation of productivity.

References

Abman, R. and Carney, C. (2020a). Agricultural productivity and deforestation: Evidence from input subsidies and ethnic favoritism in malawi. *Journal of Environmental Economics and Management*, 103:102342.

Abman, R. and Carney, C. (2020b). Land rights, agricultural productivity, and deforestation. *Food Policy*, 94:101841. Understanding Agricultural Development and Change: Learning from Vietnam.

Adamopoulos, T., Brandt, L., Leight, J., and Restuccia, D. (2017). Misallocation, selection and productivity: A quantitative analysis with panel data from china. Technical report, National Bureau of Economic Research.

Adamopoulos, T. and Restuccia, D. (2020). Land reform and productivity: A quantitative analysis with micro data. *American Economic Journal: Macroeconomics*, 12(3):1–39.

Alix-Garcia, J. M., Sims, K. R., and Yañez-Pagans, P. (2015). Only one tree from each seed? environmental effectiveness and poverty alleviation in mexico's payments for ecosystem services program. *American Economic Journal: Economic Policy*, 7(4):1–40.

Araujo, R., Costa, F., and Sant'Anna, M. (2020). Efficient forestation in the brazilian amazon: Evidence from a dynamic model. *Working paper*.

Assunção, J., Gandour, C., and Rocha, R. (2013). Deterring deforestation in the brazilian amazon: environmental monitoring and law enforcement. *Climate Policy Initiative*, 1:36.

Assunção, J., Gandour, C., and Rocha, R. (2015). Deforestation slowdown in the brazilian amazon: prices or policies? *Environment and Development Economics*, 20(6):697–722.

Assunçao, J., Lipscomb, M., Mobarak, A. M., and Szerman, D. (2017). Agricultural productivity and deforestation in brazil. Technical report, Mimeo.

Assunção, J., McMillan, R., Murphy, J., and Souza-Rodrigues, E. (2019). Optimal environmental targeting in the amazon rainforest. Technical report, National Bureau of Economic Research.

Ayerst, S., Brandt, L., and Restuccia, D. (2020). Market constraints, misallocation, and productivity in vietnam agriculture. *Food Policy*, page 101840.

Bolhuis, M., Rachapalli, S., and Restuccia, D. (2020). Misallocation in indian agriculture.

Burgess, R., Costa, F., and Olken, B. (2019). The Brazilian Amazon's Double Reversal of Fortune. Technical report, working paper.

Burgess, R., Hansen, M., Olken, B. A., Potapov, P., and Sieber, S. (2012). The political economy of deforestation in the tropics. *The Quarterly journal of economics*, 127(4):1707–1754.

Chade, H., Eeckhout, J., and Smith, L. (2017). Sorting through search and matching models in economics. *Journal of Economic Literature*, 55(2):493–544.

Chari, A., Liu, E. M., Wang, S.-Y., and Wang, Y. (2017). Property rights, land misallocation and agricultural efficiency in china. Technical report, National Bureau of Economic Research.

Chen, C., Restuccia, D., and Santaeulàlia-Llopis, R. (2017). The effects of land markets on resource allocation and agricultural productivity. Technical report, National Bureau of Economic Research.

Chiappori, P.-A., McCann, R., and Pass, B. (2016). Multidimensional matching. *arXiv preprint arXiv:1604.05771*.

Chiappori, P.-A., McCann, R., and Pass, B. (2020). Multidimensional matching: theory and empirics. Technical report, Working Paper.

Chiappori, P.-A., McCann, R. J., and Pass, B. (2017). Multi-to one-dimensional optimal transport. *Communications on Pure and Applied Mathematics*, 70(12):2405–2444.

Chiappori, P.-A., Oreffice, S., and Quintana-Domeque, C. (2012). Fatter attraction: Anthropometric and socioeconomic matching on the marriage market. *Journal of Political Economy*, 120(4):659–695.

Cochrane, M. A. and Schulze, M. D. (1998). Forest fires in the brazilian amazon. *Conservation Biology*, 12(5):948–950.

Cohn, A. S., VanWey, L. K., Spera, S. A., and Mustard, J. F. (2016). Cropping frequency and area response to climate variability can exceed yield response. *Nature Climate Change*, 6(6):601–604.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American economic review*, 100(1):35–69.

Duflo, E., Kremer, M., and Robinson, J. (2008). How high are rates of return to fertilizer? evidence from field experiments in kenya. *American economic review*, 98(2):482–88.

Duflo, E., Kremer, M., and Robinson, J. (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from kenya. *American economic review*, 101(6):2350–90.

Fajgelbaum, P. and Redding, S. (2018). Trade, structural transformation and development: Evidence from argentina 1869-1914. *NBER Working Paper*, 20217.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., et al. (2007). The shuttle radar topography mission. *Reviews of geophysics*, 45(2).

Gollin, D. and Udry, C. (2021). Heterogeneity, measurement error, and misallocation: Evidence from african agriculture. *Journal of Political Economy*, 129(1):000–000.

Gottlieb, C. and Grobovšek, J. (2019). Communal land and agricultural productivity. *Journal of Development Economics*, 138:135–152.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica, Journal of the Econometric Society*, pages 501–522.

Griliches, Z. (1980). Hybrid corn revisited: a reply. *Econometrica: Journal of the Econometric Society*, pages 1463–1465.

Hagedorn, M., Law, T. H., and Manovskii, I. (2017). Identifying equilibrium models of labor market sorting. *Econometrica*, 85(1):29–65.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, 124(4):1403–1448.

Jayachandran, S., De Laat, J., Lambin, E. F., Stanton, C. Y., Audy, R., and Thomas, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science*, 357(6348):267–273.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Koch, N., zu Ermgassen, E. K., Wehkamp, J., Oliveira Filho, F. J., and Schwerhoff, G. (2019). Agricultural productivity and forest conservation: Evidence from the brazilian amazon. *American Journal of Agricultural Economics*, 101(3):919–940.

Laurance, W. F., Cochrane, M. A., Bergen, S., Fearnside, P. M., Delamônica, P., Barber, C., D'angelo, S., and Fernandes, T. (2001). The future of the brazilian amazon. *Science*, 291(5503):438–439.

Lindenlaub, I. (2017). Sorting Multidimensional Types: Theory and Application. *The Review of Economic Studies*, 84(2):718–789.

Low, C. (2014). *Essays in gender economics*. PhD thesis, Columbia University.

McCann, R. J. (2012). A glimpse into the differential topology and geometry of optimal transport. *arXiv preprint arXiv:1207.1867*.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Nepstad, D., McGrath, D., Stickler, C., Alencar, A., Azevedo, A., Swette, B., Bezerra, T., DiGiano, M., Shimada, J., da Motta, R. S., et al. (2014). Slowing amazon deforestation through public policy and interventions in beef and soy supply chains. *science*, 344(6188):1118–1123.

Pellegrina, H. S., Sotelo, S., et al. (2019). Migration, specialization, and trade: Evidence from the brazilian march to the west. In *2019 Meeting Papers*, number 863. Society for Economic Dynamics.

Restuccia, D. and Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, 11(4):707–720.

Restuccia, D. and Santaeulalia-Llopis, R. (2017). Land misallocation and productivity. Technical report, National Bureau of Economic Research.

Sant'Anna, M. C. B. (2017). How green is sugarcane ethanol?

Shenoy, A. (2017). Market failures and misallocation. *Journal of Development Economics*, 128:65–80.

Simoes, R., Picoli, M. C., Camara, G., Maciel, A., Santos, L., Andrade, P. R., Sánchez, A., Ferreira, K., and Carvalho, A. (2020). Land use and cover maps for mato grosso state in brazil from 2001 to 2017. *Scientific Data*, 7(1):1–10.

Soares-Filho, B. S., Nepstad, D. C., Curran, L. M., Cerqueira, G. C., Garcia, R. A., Ramos, C. A., Voll, E., McDonald, A., Lefebvre, P., and Schlesinger, P. (2006). Modelling conservation in the amazon basin. *Nature*, 440(7083):520–523.

Souza-Rodrigues, E. (2019). Deforestation in the amazon: A unified framework for estimation and policy analysis. *The Review of Economic Studies*, 86(6):2713–2744.

Spangler, K. R., Lynch, A. H., and Spera, S. A. (2017). Precipitation drivers of cropping frequency in the brazilian cerrado: evidence and implications for decision-making. *Weather, Climate, and Society*, 9(2):201–213.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics*, pages 281–306. Elsevier.

Stabile, M. C., Guimarães, A. L., Silva, D. S., Ribeiro, V., Macedo, M. N., Coe, M. T., Pinto, E., Moutinho, P., and Alencar, A. (2020). Solving brazil's land use puzzle: Increasing production and slowing amazon deforestation. *Land Use Policy*, 91:104362.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica*, 79(1):159–209.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Appendix A Mathematical Appendix

Here we present details of mathematical derivations of the results presented in the text.

Our surplus function of a matching between farmer with land ℓ in time t is given by Expression 4:

$$\pi_t(x,\ell) = \ln\left(\sum_{\kappa}\exp\left(\alpha x\left[\sum_{k\in\kappa} z_{\ell,k}(p_{k,t}-\tau_{r,k})\right]+\beta_\kappa X_{\ell,t}-c_{\kappa,t}\right)\right)+\gamma$$

First, we verify that the expected payoff is strictly convex with respect to farmers productivity:

$$\frac{\partial \pi_t(x,\ell)}{\partial x} = \frac{\alpha\cdot\left[\sum_{\kappa\in\mathcal{K}}\left(\sum_{k\in\kappa}\tilde{z}_{\ell,k}\right)\cdot\exp\left(\sum_{k\in\kappa}[\alpha x\tilde{z}_{\ell,k}+\beta_k X_\ell]-c_\kappa\right)\right]}{\sum_{\kappa\in\mathcal{K}}\exp\left(\sum_{k\in\kappa}[\alpha x\tilde{z}_{\ell,k}+\beta_k X_\ell]-c_\kappa\right)} > 0 \tag{13}$$

$$\frac{\partial^2 \pi_t(x,\ell)}{\partial x^2} = \frac{\alpha^2}{\left(\sum_{\kappa\in\mathcal{K}}\exp\left(\sum_{k\in\kappa}[\alpha x\tilde{z}_{\ell,k}+\beta_k X_\ell]-c_\kappa\right)\right)^2}\cdot$$
$$\left[\left(\left(\sum_{k\in\kappa}\tilde{z}_{\ell,k}\right)-\left(\sum_{j\in\iota}\tilde{z}_{\ell,j}\right)\right)^2\cdot\exp\left(\alpha\left[\left(\sum_{k\in\kappa}x\tilde{z}_{\ell,k}+\beta_k X\right)+\left(\sum_{j\in\iota}x\tilde{z}_{\ell,j}+\beta_j X\right)\right]-(c_\kappa+c_\iota)\right)\right] > 0 \tag{14}$$

We assume that farmers do not move across lands, but prices and costs vary over the years. In order to describe the central planners problem, we define the expected return of the land over the year. This is the surplus function for each land ℓ that we consider in our problem, Expression 5:

$$\pi^e_\ell(x) = \mathbb{E}\left[\pi_t(x,\ell)\right] = \int_\Omega \pi_t(x,\ell)dP$$

where dP is the probability measure of prices and costs with support in Ω.

Expected return – $\pi^e_\ell(x)$ – is strictly increasing. As $\pi_t(x,\ell)$ is increasing, for any $x' > x$:

$$\pi_t(x',\ell) > \pi_t(x,\ell)$$

then,

$$\int_\Omega \pi_t(x',\ell)dP > \int_\Omega \pi_t(x,\ell)dP$$

$$\Rightarrow \pi^e_\ell(x') \geq \pi^e_\ell(x)$$

Expected return – $\pi^e_\ell(x)$ – is strictly convex. As $\pi_t(x,\ell)$ is stricly convex, for any $\lambda \in (0,1)$ and any productivity x' and x:

$$\pi_t(\lambda x' + (1-\lambda)x,\ell) < \lambda\pi_t(x',\ell) + (1-\lambda)\pi_t(x,\ell)$$

then,

$$\int_\Omega \pi_t(\lambda x' + (1-\lambda)x,\ell)dP < \int_\Omega \left(\lambda\pi_t(x',\ell) + (1-\lambda)\pi_t(x,\ell)\right) dP$$

$$\Rightarrow \int_\Omega \pi_t(\lambda x' + (1-\lambda)x,\ell)dP < \lambda\int_\Omega \pi_t(x',\ell)dP + (1-\lambda)\int_\Omega \pi_t(x,\ell)dP$$

Therefore,

$$\pi^e_\ell(\lambda x' + (1-\lambda)x) < \lambda\pi^e_\ell(x') + (1-\lambda)\pi^e_\ell(x)$$

Appendix B Data Appendix

Table 6: Estimated values of cost of adoption ($c_{\kappa,t}$)

year	Cotton		Beef		Soy-Corn		Soy-Cotton		Soy		Soy-Sunflower		Sugar	
2003	3.611	0.029	-2.671	0.010	2.429	0.017	7.726	0.058	0.678	0.013	8.297	0.136	1.736	0.030
2004	2.970	0.028	-2.392	0.010	2.181	0.016	5.826	0.032	1.063	0.014	7.821	0.090	1.931	0.031
2005	1.746	0.027	-2.976	0.011	-0.294	0.015	3.928	0.038	-0.940	0.013	6.670	0.112	1.034	0.028
2006	1.529	0.027	-2.938	0.011	-0.813	0.015	2.503	0.029	-1.139	0.013	5.485	0.141	1.053	0.027
2007	1.576	0.026	-2.937	0.011	-0.241	0.014	3.641	0.033	-0.141	0.014	5.643	0.130	0.887	0.026
2008	0.667	0.024	-3.346	0.012	-0.200	0.016	3.047	0.032	-0.661	0.015	4.529	0.047	0.243	0.026
2009	1.385	0.027	-2.790	0.011	-0.599	0.014	3.001	0.030	-0.494	0.013	4.934	0.074	0.815	0.025
2010	2.282	0.038	-3.105	0.011	-1.298	0.014	2.261	0.025	-0.866	0.014	5.831	0.140	0.523	0.025
2011	1.907	0.028	-3.260	0.012	-0.429	0.015	3.009	0.025	-0.822	0.014	4.636	0.071	0.538	0.026
2012	0.767	0.025	-3.112	0.011	-0.440	0.014	1.624	0.021	-0.371	0.015	4.116	0.058	0.327	0.025
2013	2.266	0.038	-3.178	0.011	-0.701	0.014	2.269	0.024	-0.396	0.014	4.516	0.077	0.230	0.024
2014	1.595	0.036	-2.844	0.011	-1.507	0.013	0.951	0.019	-0.586	0.014	2.772	0.041	0.325	0.024
2015	2.042	0.034	-2.061	0.011	-0.943	0.014	1.106	0.021	-0.397	0.014	4.553	0.059	1.029	0.025
2016	2.230	0.047	-3.094	0.012	-1.549	0.014	1.312	0.021	-0.973	0.014	4.332	0.073	0.261	0.025
2017	3.017	0.043	-2.550	0.011	-1.381	0.012	1.607	0.019	-0.091	0.014	3.548	0.049	0.835	0.024

This table shows the estimated values of cost of adoption ($c_{\kappa,t}$) and the estimated standard deviations. This table generates Figure 3. For each crop-system κ the table shows the coefficient (on the left) and standard deviation (on the right).

Table 7: Polynomial coefficients that fits x_i

Coefficient	Estimates
h_0	-3.401
	(0.115)
h_1	55.274
	(0.801)
h_2	-229.127
	(1.981)
h_3	339.535
	(2.123)
h_4	-163.066
	(0.829)
v_1	4.609
	(0.333)
v_2	-41.714
	(0.95)
v_3	78.572
	(1.158)
v_4	-42.95
	(0.509)

This table presents the estimation results for the parameters that define the spatial distribution of farmers' productivity (x_i) as defined by expression (10)

Definition 2. *Suppose a matching function M. The matching satisfies the nested criterion if for each given land $\ell, \ell' \in L$ and for each $x, x' \in X$ such that $x > x'$ and $M(x) = \ell, M(x') = \ell'$ the following conditions are satisfied:*

$$\left\{\tilde{\ell} \in L | \pi_{\tilde{\ell}}^{e\prime}(x') \leq \pi_{\ell'}^{e\prime}(x')\right\} \subseteq \left\{\tilde{\ell} \in L | \pi_{\tilde{\ell}}^{e\prime}(x) \leq \pi_{\ell}^{e\prime}(x)\right\} \tag{15}$$

$$\#\left\{\tilde{\ell} \in L | \pi_{\tilde{\ell}}^{e\prime}(x') \leq \pi_{\ell'}^{e\prime}(x')\right\} = \#x' \tag{16}$$

$$\#\left\{\tilde{\ell} \in L | \pi_{\tilde{\ell}}^{e\prime}(x') \leq \pi_{\ell'}^{e\prime}(x)\right\} = \#x \tag{17}$$

Definition 3. *The function $\pi_\ell^e(\cdot)$ satisfies the twist condition (discrete version) if for any $\ell, \ell' \in L$ such that $\ell \neq \ell'$, and for any $x, x' \in X$ such that $x > x'$, we have:*

$$\pi_\ell^{e\prime}(x') \geq \pi_{\ell'}^{e\prime}(x') \iff \pi_\ell^{e\prime}(x) \geq \pi_{\ell'}^{e\prime}(x)$$

Proposition 3. *If the twist condition (discrete version) is valid, then the candidate matching is nested and is the only nested matching.*

Proof. If the model is nested, then the largest farmer x_m must be matched with the land l_m that has the highest $\pi_\ell^{e\prime}$ for all $x \in X$. Otherwise, $\#\left\{\tilde{\ell} \in L | \pi_{\tilde{\ell}}^{e\prime}(x_m) \leq \pi_{\ell'}^{e\prime}(x_m)\right\} < \#x_m$. Doing the same argument iteratively yields the candidate matching function. ■