Comparing market instruments for forest conservation in Brazil using farm-level census data
Online Appendix

Guilherme DePaula[*] Leandro Veloso[†]

[*] Corresponding author. Assistant Professor, Department of Economics and Center for Agricultural and Rural Development (CARD), Iowa State University. gdepaula@iastate.edu

[†] Graduate Student, Electrical Engineering Department, Federal University of Rio de Janeiro. leandroveloso@gmail.com

Appendix A. Data

A1. Data availability and replication of results

All results presented in this study can be replicated at the Centro de Documentacão e Disseminacão de Informacões (CDDI) of the Brazilian Institute of Geography and Statistics (IBGE) located in Rio de Janeiro, Brazil. Although the Agricultural Census information is confidential to protect the identity of farmers, it is not proprietary. IBGE grants access to the farm-level census data as well as to other confidential surveys in Brazil to Brazilian and International researchers.

In order to replicate the results of this study a researcher must have four files provided by the authors: (1) IBGE project submission form (TR) with complimentary datasets, (2) SAS code for construction of the census dataset, (3) STATA code for descriptive statistics and estimation of land-use models, and (4) STATA code for market simulation. The IBGE charges a fee of approximately $300 for access to the CDDI computer lab. The authors will provide the TR with the census tables and variables as well as the complimentary datasets. The census is the only confidential dataset used in this study. All other datasets are available from the authors. The agricultural census is organized in a series of text files for corresponding tables. The SAS program reads the text files and integrate them into a single dataset. The first STATA program estimates the discrete choice models, compute the OC for each farm, and estimate the supply function for reforestation in each market.

The market simulation can be replicated outside of the CDDI, without restrictions. We created a dataset of simulated reforestation for each market. This dataset is not confidential and is available from the Authors. The second STATA program replicates all the market analysis presented in the article, including calculation of the optimal tax and the market price as well as parameters of the reforestation function for each market.

In the following sections of this online appendix we describe the dataset, the theoretical framework with the definition of the parameters estimated, and the tree-step empirical analysis. We focus on describing all the assumptions used in the analysis and providing guidance for the replication of all the results of this article.

A2. Data description

Brazilian agricultural census

The primary dataset used in the analysis is the farm-level version of the 2006 and 2017 Agricultural Census surveys completed by IBGE (IBGE, 2006, 2017). IBGE surveys over 5 million farmers every 10 years to collect information on farm and farmer characteristics, including land-use choices at the crop level and production output and technology. The main advantage of using farm-level data is to capture the large heterogeneity in the opportunity cost of forest land based on observed farmer land-use choices.

We restrict our analysis to the population of large commercial farms in Brazil as large commercial farmers are more likely to participate in forestland markets because of the requirements for formal land ownership and the high transaction cost of trading. Also, after the latest FC revision, small farmers are not required to reforest their land up to the legal reserve requirement. Finally, focusing on large farms simplify the land-use model estimation as large commercial farms in Brazil tend to specialize in one type of land-use. We define large commercial farms based on the total production value and the farm size, following the analysis of Alves *et al.* (2013) who found that about 86% of total agricultural production value in Brazil was generated by about 10% of the farmers, which had a monthly production value equal to or above ten minimum wages (MW) in 2006.[1] In our preferred sample, we select all farms in Brazil with total production value above 2 minimum wages, or R$7,200 in 2006, and farm size above 5 ha. The final dataset contains 1,195,450 commercial farms in 2006 and 1,129,400 commercial farms in 2017.

Description of census variables:

Land-use choice. The dependent variable for the crop discrete choice model is the farmer's land-use choice. We use the IBGE definition of the primary economic activity of the farm, variable W462900 (*Class. Ativ. Econ. Classe – Table Dados Gerais – Variaveis Derivadas*). We use five classes of economic activities defined by IBGE: 111 – Cereals, 115 – Soy; 113 – Sugarcane; 131 – Citrus; and 134 – Coffee. The baseline choice is others, the remaining classifications of economics activity.

Land-use area. In the second stage of our empirical analysis we model the share of land allocated to agriculture within a farm. The area allocated to agriculture is the sum of crop and pasture area. Crop area is defined as the sum of permanent crop area, variable W041100 (*Área de Lavoura Permanente – Table Utilizacão das Terras Lavouras em Ha – Variáveis Derivadas*), and temporary crop area, variable W041400 (*Área de Lavoura Temporária – Table Utilizacão das Terras Lavouras em Ha – Variáveis Derivadas*). The pasture area is defined as the sum of natural pasture area, variable W041700 (*Área de Pastagem Natural – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*), degraded pasture area, variable W041800 (*Área de Pastagem Degradada – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*), and non-degraded pasture area, variable W041900 (*Área de Pastagem Não Degradada – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*). The forest area is defined as the sum of the area of natural forest for preservation, variable W04200 (*Área de Florestas Nat. Preservacão – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*), the area of natural forest for commercial

[1] The minimum wage in Brazil is defined in terms of monthly income. In 2006, the minimum wage in Brazil was R$300 per month. For example, an annual gross revenue of 10 minimum wages per farm would correspond to annual revenue of R$36,000 per year. This annual gross revenue threshold is used to define commercial farms.

use, variable W042100 (*Área de Florestas Nat. Exploracão – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*), and the area of planted forest, variable W042200 (*Área de Florestas Plantadas – Table Utilizacão das Terras Pastagens em Ha – Variáveis Derivadas*).

Water access. We use indicator variables to identify farms with water sources within the property. The census has three variables that identify the existence of a spring water source, a river, or a lake in the farm. We use these indicators of water source to control for preservation requirements in the Forestry Code as farmers are required to preserve natural vegetation along rivers and water sources. The water source indicator variables are Spring Water, variable v043300 (*Se Tem Nascentes no Estabelecimento – Table Área do Estabelecimento – Características*), River, variable v043400 (*Se Tem Rios ou Riachos no Estabelecimento – Table Área do Estabelecimento – Características*), and Lake, variable v034500 (*Se Lagos Naturais ou Acudes no Estabelecimento – Table Área do Estabelecimento – Características*).

Farm size. The total area of the farm measured in ha is the census variable W040100 (*Área Total do Estabelecimento – Table Área do Estabelecimento em Ha – Variáveis Derivadas*).

Production value. The total gross revenue of the farm in year 2006 measured in Reals is the census variable w462704 (Valor Total da Producão – Table Dados Gerais – Variáveis Derivadas).

Municipality compliance in 1996. We use the 1996 Agricultural Census to construct an indicator variable at the municipality level to identify compliance with the forest reserve requirements defined in the forestry code. The objective is to control for possible endogeneity of enforcement efforts. We contrast the total agricultural land in a municipality with the required forestry reserve to identify the municipalities in 1996 that were in compliance. The census variables used are the same as defined in land-use area above.

Share of commercial siviculture. We identify municipalities with significant revenue from the commercialization of forestry products to control for the market benefits of forestland. We first identify farms that have value of siviculture production above R$1,000 using the value of siviculture production variable from the census, variable w462500 (*Valor da Producão Sivicultura – Table Valor da Producão Vegetal – Variáveis Derivadas*). We then compute the share of farms with commercial siviculture at the municipality level based on the number of farms within the municipality with sivivulture production value above R$1,000.

Complementary datasets

We combine the agricultural census data with information on the potential yield of crops, the transportation cost of agricultural production, and municipality socio-economic characteristics to measure the relative profitability of each land-use and to control for market access in the discrete choice model.

Potential yields. IIASA and FAO estimate the potential yield of 154 crops under three levels of land management and input use and differentiate between rain-fed and irrigated farming (IIASA, 2018). IIASA and FAO estimate crop potential yield for millions of grid cells (0.5 by 0.5 degrees latitude and longitude) using measurements of climate and soil characteristics. In this analysis, we use potential yield measurements for high input use, given the focus on commercial farmers, and for rain-fed farming as less than 3% of agriculture production in the sample uses irrigation. We use a geographical information system to combine the IIASA/FAO measures of potential yield for soy and alternative crops such as sugarcane, rice, cotton, coffee, and corn with the Census dataset at the census block level. There were 70,000 rural census blocks in Brazil in 2006.

The main advantage of using the potential yield variable constructed by IIASA and FAO is that it is determined independently of the choices of Brazilian farmers, and therefore it is a source of exogenous cross-sectional information about the productivity of agricultural land. We use the potential yield variable for different crops, at the high input use level, to estimate the discrete choice land use model. In contrast to climate and soil characteristics, the potential yield variable is defined at the crop level and therefore contains information regarding the relative productivity of each crop at each farm.

Transportation cost. We use variation in transportation cost to capture variability in the profitability of crops produced in different regions of the country. Transportation costs are important determinants of farm profitability in a large country such as Brazil because crops are produced over 1,000 km from markets. We estimate transportation costs using freight data from the System of Freight Information (SIFRECA) maintained by the Luiz Queiroz College of Agriculture at the University of Sao Paulo (ESALQ/USP) (ESALQ, 2009). Most of the agricultural production in Brazil is transported through roads and SIFRECA contains average transportation cost by road routes for the main agricultural products in Brazil.

We estimate transportation cost as a quadratic function of distance traveled using the SIFRECA route/product dataset. The transport cost dataset used is a subset of the SIFRECA annual report containing 1,039 routes, 625 for soy transportation and 423 for corn transportation. The estimated transportation cost equation is:

$$\widehat{tc} = 9.086 + 0.0087d - 5.24 \times 10^{-6}d^2, \tag{A1}$$

where d is the distance traveled measured in km and $(\widehat{tc})$ is the estimated transportation cost measured in 2008 Reals per ton. The quadratic function of distance is consistent with a standard road transport model with a fixed charge per km and an adjustment component for large distances. The quadratic model in equation (A1) explains 89% of the variation in transportation cost in the SIFRECA dataset and was estimated for the combination of soy and corn routes. We find no statistically significant difference in models estimated separately for corn and soy routes.

We use the estimated transportation cost model (A1) to calculate the transportation cost of agricultural products, excluding cattle, for each rural census block in Brazil. We use the straight-line distance from each census block to the closest large city or port as an estimate for the distance to the market. A market is defined as either a large city, a city with more than 1 million habitants, or a port of large, medium, or small size. Our calculation of the distance to market considers sea ports located along the coast and inland ports located on rivers. Figure A1 shows the cross-sectional variation in the estimated transportation costs in Brazil by microregion based on the minimum distance metric. The transportation costs in the north and northwest agricultural frontiers of Brazil can be four times larger than the transportation cost for farms along the coast.

Forest debt. The forest debt is the amount of reforestation required in each state/biome market to meet the legal requirements of the FC. The FC regulates land use within private properties in Brazil. The main objective of the code is to preserve the endowment of natural vegetation inside the farms, recognizing the value of biodiversity and ecosystem services such as freshwater protection. The FC specifies two land diversion requirements, the legal reserve and the areas of permanent preservation. The legal reserve requirement specifies, at the biome level, the fraction of the farm that must be preserved in the original natural vegetation. The reserve requirement is 80% in the Amazon biome, 35% in the Savanna biome (Campos Gerais), and 20% in the remaining biomes, including the Atlantic Forest. In addition to the reserve requirement, the DC defines the areas or permanent preservation to protect natural vegetation along rivers and other water sources, and on hilltops. These areas are defined in terms of the width of rivers or watercourses and the slope and height of hilltops.

In May of 2012, a new version of the FC was approved by congress and signed into law (Law 12.651/2012). The new FC differentiated the land diversion requirement from reforestation requirements. The reserve requirements at each biome remained the same but the formula for calculating the amount each farm must reforest in order to comply with the FC changed. For example, the new FC exempted small farmers from their reforestation obligations, revised the definition of permanent preservation areas, and incorporated land in permanent preservation areas into the definition of reserve requirements, reducing farmer's reforestation debt.[2] The reforestation debt is defined as the amount of land a farmer must convert to natural vegetation to comply with the FC. In this study, we use the calculation of the forest debt from Soares-Filho *et al.* (2014), as it reflects the latest revision of the FC.

[2] The legal definition of a small farmer in Brazil varies geographically and ranges from 20 ha in the southern regions of Brazil to 440 ha in the Northwest regions, which include the Amazon biome.

The primary market boundary defined in the FC is the state-biome combination. There are 44 different state/biome combinations in Brazil. Figure A2 shows the six biomes and the 44 state/biome markets in Brazil.

Municipality characteristics. We use the variables income per capita and population density to capture variation in market access and in demand for agricultural products across Brazil (Embrapa, 2014). Income per capita is measured in Reals per person and population density is measured in the number of persons per square kilometer. Also, we use mean elevation and the standard deviation of elevation to measure the suitability of the land for mechanized agriculture and the additional demand for forestland due to the FC requirements for natural vegetation in hilltops (EMBRAPA 2012). The unit of measurement of elevation is meters.

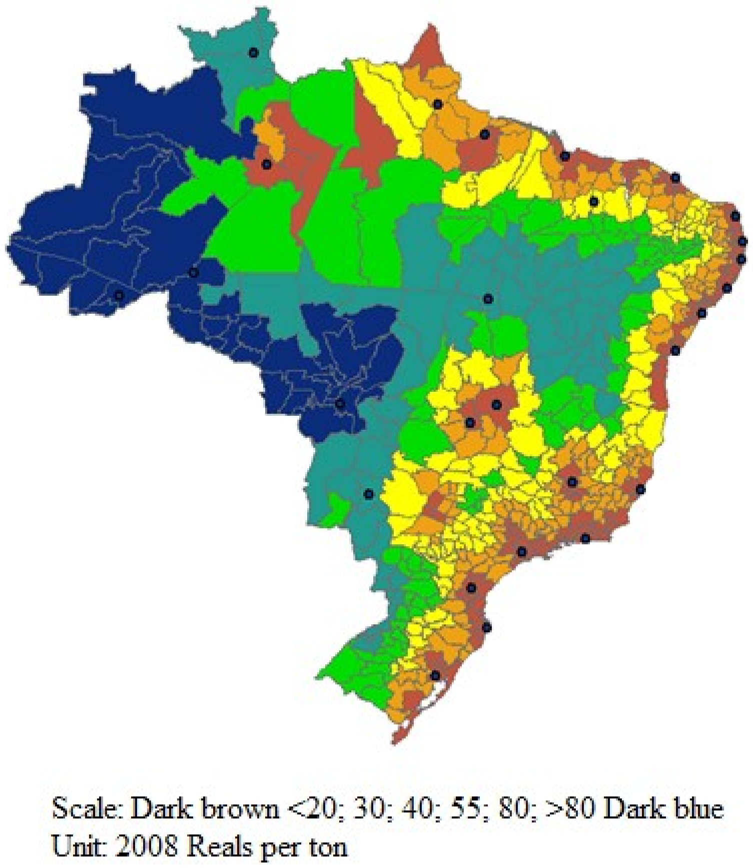


Figure A1: Cross-sectional variation in transportation costs in Brazil.

Notes: Figure A1 maps the average transportation cost at the microregion level. There are about 500 microregions in Brazil. We estimate the transportation cost using the average transportation cost by road routes for main agricultural products in Brazil based on the SIFRECA dataset and the straight-line distance from the rural census block to the closest large city or port (ESALQ, 2009). There are approximately 70,000 rural census blocks in Brazil. The estimated transportation cost equation is a quadratic function of distance to market: $\widehat{tc} = 9.086 + 0.0087d - 5.24 \times 10^{-6}d^2$.

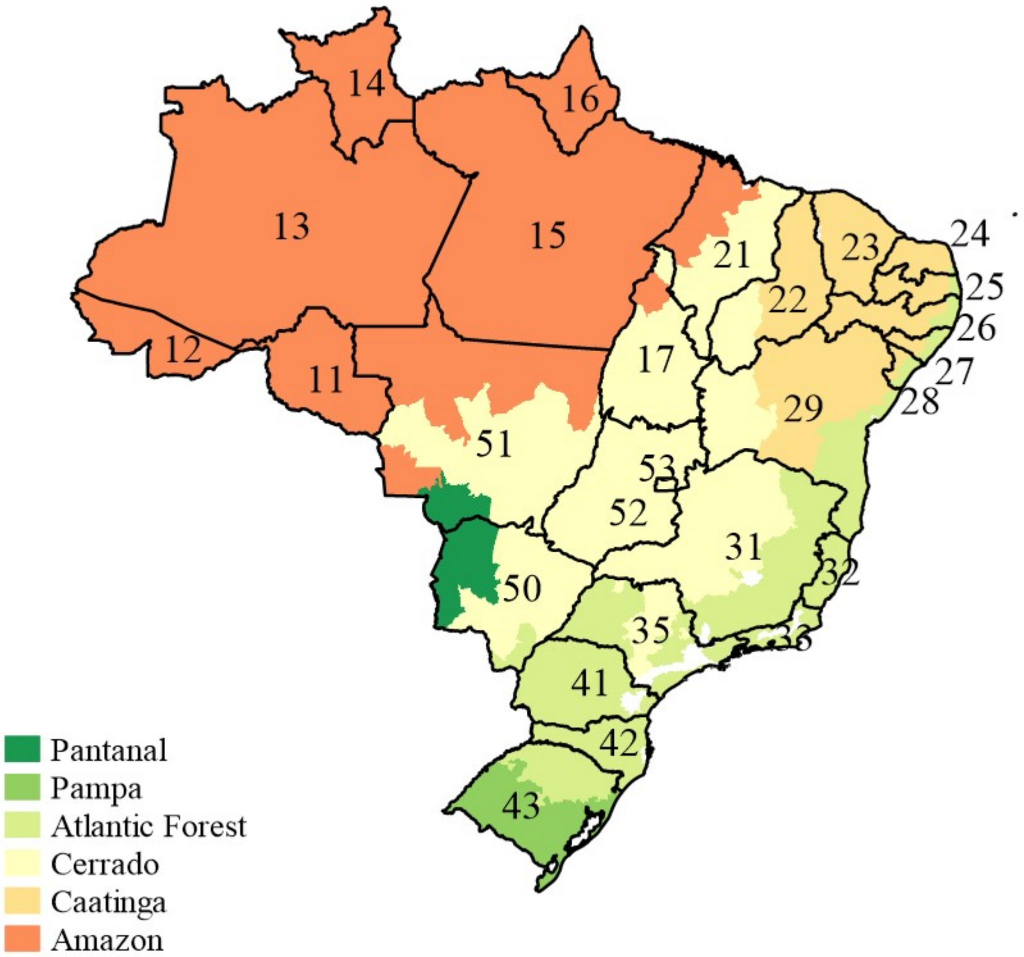
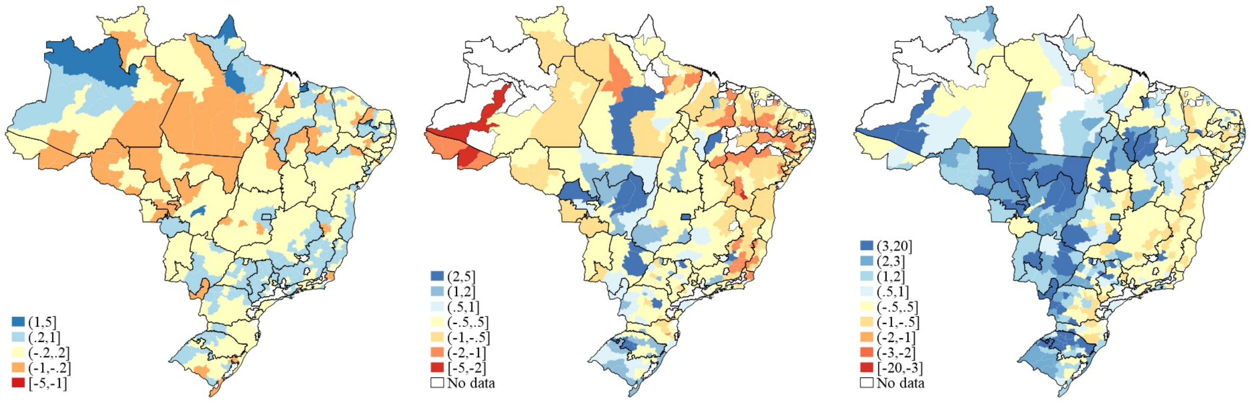


Figure A2: Biomes and biome-state markets in Brazil.

Notes: Figure A2 maps the six biomes, the 26 states in Brazil, and the 44 biome-state markets. The map shows the state boundaries and the state codes. The biomes are identified based on the six colors. The markets are the combinations of states and biomes. For example, the state of Mato Grosso, code 51, has three markets. The Mato Grosso-Amazon market is located in the north of the state, the Mato Grosso-Pantanal market is located in the south, and the Mato Grosso-Cerrado market is in the center and east part of the state. São Paulo state has two markets, the São Paulo-Cerrado in the middle of the state, and the São Paulo-Atlantic Forest market in the east and west parts of the state.



(a) Change in share of forestland (SD)

(b) Change in OC (SD)

(c) Change in standard deviation of OC (SD)

Figure A3: Changes in market characteristics: 2006-2017.

Notes: The geographical unit in each map is a microregion and the black boundaries identify each state-biome market. There are 44 state/biome markets in Brazil and 551 microregions. Panel a maps the change in the average share of forestland across all farms of a microregion. The unit of measure is the standard deviation of the share of forestland in 2006. Panel b maps the change in the average OC across all farms in a microregion. The unit of measure is the standard deviation of the OC in the microregion in 2006. Panel c maps the change in the standard deviation of the average OC in a microregion. The unit of measure is one standard deviation in year 2006. An increase in the standard deviation in OC in a microregion means that the OC within that microregion has a larger spread around the mean OC.

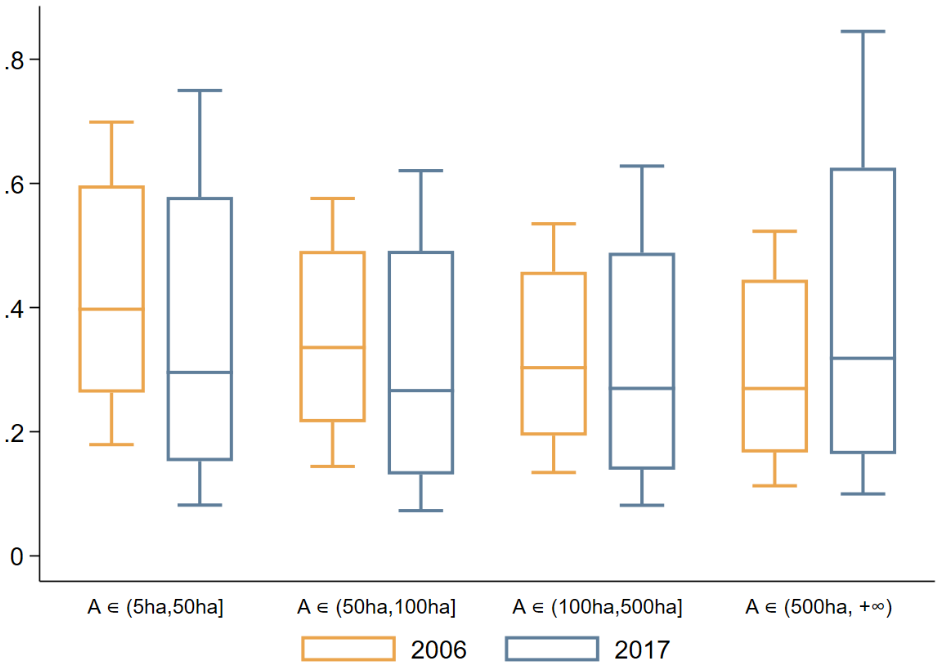


Figure A4: Changes in the distribution of OC by census year and farm size.

Notes: Figure A3 shows the distribution of OC estimates for the census years 2007 and 2017 and for four classes of farms by farm size. The boxplot graph highlights five quantiles of the distribution of OC: Q5, Q25, Q50, Q75, and Q95.

Table A1: Summary statistics: land use by census year

Land use	2006				
	Amazon (1)	Cerrado (2)	Atl. Forest (3)	Other (4)	Total (5)
Area (million hectares):					
Forest	17.3	23.5	9.4	8.3	58.4
Pasture	21.9	44.0	19.6	19.5	105.0
Crops	2.7	17.0	15.7	4.7	40.2
Total	42.3	87.2	46.8	34.9	211.3
Number of farms (thousands):					
Compliant	11.5	28.7	190.1	58.0	288.3
Total	133.1	224.9	635.7	201.8	1,195.5
Land value (Billions USD)	33.9	146.6	156.9	34.6	372.0
	2017				
	Amazon (6)	Cerrado (7)	Atl. Forest (8)	Other (9)	Total (10)
Area (million hectares):					
Forest	22.9	29.7	12.9	7.2	72.7
Pasture	29.2	43.3	19.1	17.2	108.8
Crops	4.5	25.6	17.8	4.9	52.8
Total	57.7	99.7	50.2	32.1	239.7
Number of farms (thousands):					
Compliant	9.5	36.2	208.6	46.6	300.9
Total	164.1	241.2	571.5	152.5	1,129.4

Notes: Table A1 reports summary statistics for land use choices of commercial farms based on the agricultural census dataset. We define commercial farms based on total annual production value and farm size. We select all farms in Brazil with a total production value above 2 minimum wages and farm size above 5 ha. The 2017 agricultural census does not have farmland use values. The number of compliant farms is the number of farms with native vegetation area equal to or above the reserve requirement in the FC.

‘

Table A2: Land use choice frequency by census year

Number of farms	All farms		Large farms		Top farms	
by land use choice	2006	2017	2006	2017	2006	2017
	(1)	(2)	(3)	(4)	(5)	(6)
00-Grazing	863,029	894,746	143,391	146,803	41,020	43984
01-Cereaks	101,307	44,755	20,701	11,330	3,074	2,535
02-Soy	110,285	145,578	39,396	64,652	8,681	15,753
03-Sugarcane	30,758	22,842	10180	9,137	2,036	2,897
04-Citrus	15,332	5,967	3,410	1,905	308	295
05-Coffee	88,301	56,396	13,583	8,788	768	618
Total	1,209,012	1,170,284	230,661	242,615	55,887	66082

Notes: Table A2 reports the number of farms by land use choice, defined using the farm economic classification from IBGE. We define our reference class as the grazing class because it is mostly formed by grazing farms which occupy most of the agricultural land in Brazil. The grazing reference class in our land use choice model also has other classifications with lower frequency. The subsample of top farms is defined as farms with more then 10 minimum wages in annual gross revenue and farm size larger than 500 hectares.

Appendix B. Discrete choice model

Inferring the OC function from land-use choices

We use a discrete choice model of land use to infer the OC of forestland from the observed land-use choices of farmers. Discrete choice models map the characteristics of the decision maker and the choice alternatives into a vector of utilities or returns for each choice available (Ben-Akiva and Lerman, 1985; Berry, 1994; Cardell, 1997; Lubowski *et al.*, 2006). The function takes the form $(Z_{ij}) \to \mathbb{R}^j$, where Z_{ij} is a vector of the characteristics of farmer i and land-use choice j, and J alternative land uses are available. In commercial agriculture, the profit function does the mapping. Farmer i can be represented by a vector of agricultural land-use returns Π^*:

$$\Pi_i^* = (\Pi_1^*, \Pi_2^*.., \Pi_J^*) = (\pi_1^*(Z_{i1}, p_1, w), \ldots, \pi_J^*(Z_{iJ}, p_J, w)), \tag{A2}$$

where p_j and w are the vectors of output and input prices, respectively. In this framework, the OC can then be represented in terms of the vector of land-use returns Π^*:

$$OC_i(Z_{i1}, p_1, w) = Max(\pi_1^*, \pi_2^*.., \pi_J^*) \sim F_{OC/Z,p,w}, \tag{A3}$$

where $F_{OC/Z,p,w}$ is the conditional distribution of the OC.

As prices and characteristics change, the OC also changes. We use a simple nested land-use model to calculate the OC defined in equation (A3). The farmer's land-use decision process is broken down into two nested choices: the choice of farming (f) and the choice of farm type (j) (figure A5). The farmer decides for each plot of land whether to farm or leave it as forest, and conditional on farming, he chooses the farm type corresponding to specific land use. Examples of farm types are grazing, cereals, and soy.

The observed land-use choices of farmers reveal the relative profitability of each land use (each pi_j^*). The land rent from land use f_j, where f specifies the choice of farming and j specifies the choice of farm type, can be expressed as $\pi_{fj} = \tilde{\pi}_f + \tilde{\pi}_{fj} + \tilde{\epsilon}_f + \tilde{\epsilon}_{fj}$, where the rents are decomposed into an observed component such as $\tilde{\pi}_f$ and an unobserved component such as $\tilde{\epsilon}_f$ (Berry, 1994; Cardell, 1997). The conditional probability equation for the choice of agriculture shows how the OC fits into the discrete choice model for farming:

$$Pr(Agr) = Pr(\tilde{\pi}_{agr} + \tilde{\epsilon}_{agr} + max_{jagr}(\tilde{\pi}_j + \tilde{\epsilon}_{agr,j}) \geq \tilde{\pi}_{forest} + \tilde{\epsilon}_{forest}). \tag{A4}$$

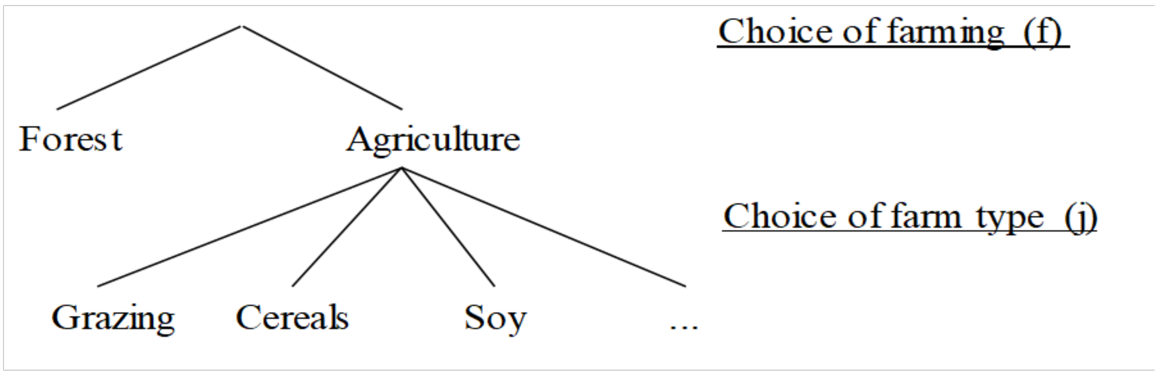


Figure A5: Discrete choice model of land use for a commercial farmer.

Notes: Figure A5 shows the nested choice set of a farmer in our discrete choice land use model. The farmer land use decision process is divided into two nested choices: farming (f) and farm type (j). The farmer decides for each plot of land whether to farm or leave it as forest, and conditional on farming, he chooses the farm type corresponding to specific land use. Examples of farm types are grazing, cereals, and soy. The OC can be represented by the standard inclusive value formula of the nested logit model: $OC = \sum_j e^{\pi_j}$. The benefit of using a discrete choice framework to estimate OC is that we can easily model changes in the OC.

Assuming that $\tilde{\epsilon}_{fj}$ is Gumbel distributed with the parameter μ_j, the OC can be represented by the standard inclusive value formula of the nested logit model:

$$OC = max_{j \in agr}(\tilde{\pi}_j + \tilde{\epsilon}_{agr,j}) = \frac{1}{\mu_j} Ln \sum_j e^{\pi_{agr,j} \times \mu_j}. \tag{A5}$$

The benefit of using a discrete choice model to estimate the OC is explicitly modeling the changes in the OC function (A5).

Identification of the discrete choice model

Our identification strategy follows the special regressor approach of Lewbel (2014). A special regressor is an exogenous observed covariate with large support. The intuition for the identification strategy is that this special regressor creates variation in the returns of alternative land uses for similar farmers (Lewbel, 2014; Dong and Lewbel, 2015). We use the interaction between a potential yield measure for each crop from IIASA (2018) and the straight-line distance to markets as a special regressor. This interaction captures the variation in the maximum potential net revenue specific to each alternative land-use choice. The potential yield measure is exogenous to farmers' choices and captures the variation in agricultural productivity, whereas the straight-line distance captures the variation in transportation costs.[3]

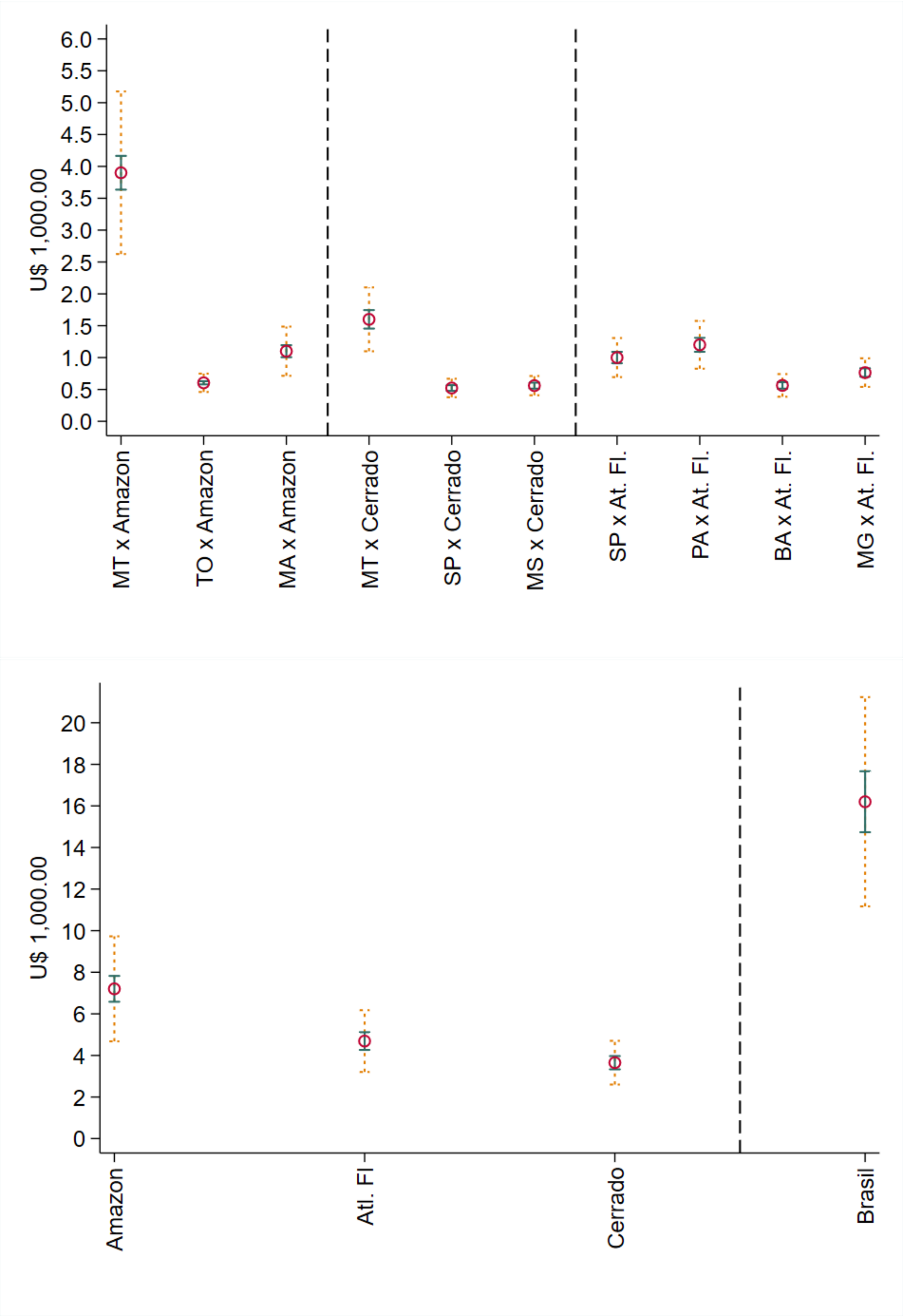
A concern with our identification strategy is the potential correlation between the straight-line distance and infrastructure. We thus use several measures of market access and the interaction between the biome and state fixed effects to control for the unobserved variation in policies and infrastructure investments across states. Therefore, our land-use

[3] See Bustos *et al.* (2016) and DePaula (2023) for empirical applications in Brazilian agriculture using IIASA/FAO potential yield measurements.

analysis uses the within-state variation in the maximum potential net revenue for each land use to estimate the OC. Two other sources of bias could be the endogenous variation in policy enforcement and demand for forestry products. To address these concerns, we create one variable for historical compliance with the Brazilian Forestry Code (FC) and one variable for the local value of forestry products. We find that our estimates of the supply function of forestland are robust to adding these controls. Finally, we test the robustness of our results to alternative measures of the potential returns for each land-use choice, different nesting structures and choice sets in the nested logit model, and six methods for computing changes in the OC. We find that our estimates of the OC and forestland supply function and our simulation results are robust to these changes in the model specification.

Additional Results. Figures and Tables

A - Sensitivity of reforestation



B - Sensitivity of market price of forest certificates

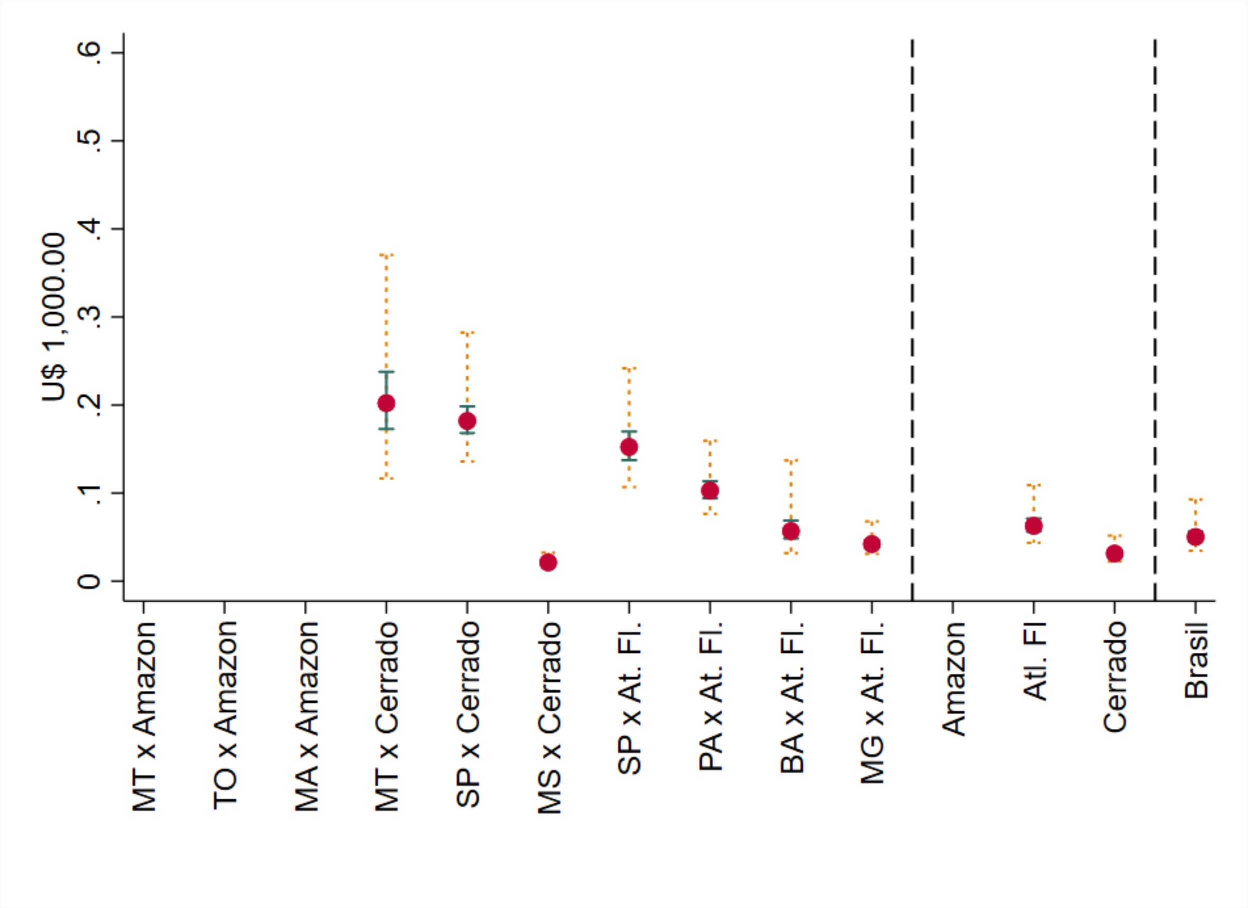


Figure A6: Sensitivity of market instruments to higher OC.

Notes: Figure A6 shows the variation in the optimal amount of land reforested (panel A) and in the optimal tax (panel B) and by state/biome market, biome, and country. The green bar represents the variation for a 0.5SD change in the OC, and the dashed yellow line represents the variation for a 2SD change in the OC. The states included in the graphs are MT – Mato Grosso; TO – Tocantins; MA – Maranhão; SP – São Paulo; PA – Para; BA – Bahia; MG – Minas Gerais. The biomes included in the graph are Amazon, Cerrado (Savanna), and AT. Fl.- Atlantic Forest. The optimal tax and optimal reforestation computed at different geographical levels are based on all commercial farms within the geographical boundary.

Table A3: Conditional logit model of land use: census year 2006

Land use		Land use choices				
variables		Cereals	Soy	Sugarcane	Orange	Coffee
	(1)	(2)	(3)	(4)	(5)	(6)
Land use alternative specific variables						
Maximum potential	0.0972***					
net revenue (MNR)	(0.0109)					
Interacted with:						
mean elevation	-5.06e-05***					
	-0.0000106					
Log pop. density	-0.0105***					
	(0.00226)					
Farm characteristics						
Transport cost		0.00142	0.0277***	-0.00958***	-0.0175***	-0.00688
		(0.00140)	(0.00195)	(0.00252)	(0.00349)	(0.00447)
Mean elevation		0.000286**	0.00161***	-0.000209	-0.000151	0.00410***
		(0.000121)	(0.000154)	(0.000201)	(0.000279)	(0.000392)
Income per capita		0.00751*	0.0339***	0.0282***	0.0117	-0.0738***
		(0.00456)	(0.00851)	(0.00822)	(0.00844)	(0.0193)
Log pop. density		0.131***	0.374***	0.663***	0.746***	0.816***
		(0.0265)	(0.0427)	(0.0484)	(0.0749)	(0.101)
Log farm size		-0.122***	0.0714***	0.170***	-0.119***	-0.404***
		(0.0133)	(0.0175)	(0.0188)	(0.0300)	(0.0173)
N (million farms)		1.195	1.195	1.195	1.195	1.195
Chi-squared statistics		1,947	1,947	1,947	1,947	1,947

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Table A3 reports the results of the discrete choice model for farm type, the first stage of the nested choice set in figure A5, for census year 2006. We estimate a conditional logit model with three land-use specific variables: maximum potential net revenue and its interactions with elevation and density. The land-use specific variables differ for each farm/land use combination. All standard errors are clustered at the municipality level. There are 5,336 municipalities in our sample with an average 220 commercial farms.

Table A4: Conditional logit model of land use - Census year 2017

Land use		Land use choices				
variables		Cereals	Soy	Sugarcane	Orange	Coffee
	(1)	(2)	(3)	(4)	(5)	(6)
Land use alternative specific variables						
Maximum potential	0.0868***					
net revenue (MNR)	(0.0179)					
Interacted with:						
mean elevation	5.58e-06					
	(1.37e-05)					
Log pop. density	-0.0193***					
	(0.00360)					
Farm characteristics						
Transport cost		-0.00105	0.0192***	-0.00926***	-0.0201***	0.00328
		(0.00160)	(0.00181)	(0.00224)	(0.00429)	(0.00414)
Mean elevation		0.000152	0.00144***	-8.73e-05	-0.000486	0.00621***
		(0.000161)	(0.000164)	(0.000180)	(0.000301)	(0.000468)
Income per capita		0.0199***	0.0288***	0.0128***	0.0129***	-0.0796***
		(0.00287)	(0.00435)	(0.00317)	(0.00458)	(0.00843)
Log pop. density		0.320***	0.367***	0.728***	0.754***	1.405***
		(0.0290)	(0.0358)	(0.0423)	(0.0696)	(0.141)
Log farm size		-0.000849	0.169***	0.326***	0.110***	-0.476***
		(0.0190)	(0.0142)	(0.0201)	(0.0288)	(0.0258)
N (million farms)		1.129	1.129	1.129	1.129	1.129
Chi-squared statistics		2,143	2,143	2,143	2,143	2,143

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Table A4 reports the results of the discrete choice model for farm type, the first stage of the nested choice set in figure A5, for census year 2017. We estimate a conditional logit model with three land-use specific variables: maximum potential net revenue and its interactions with elevation and density. The land-use specific variables differ for each farm/land use combination. All standard errors are clustered at the municipality level. There are 5,336 municipalities in our sample with an average 220 commercial farms.

Table A5: Simulation of a tax on agricultural land and markets of forest certificates in Brazil - Census year 2017

Market (Biome-state)	Market characteristics				Tax on agricultural land			Market of forest certificates			
	Forest Debt	OC		Marginal OC	Tax ($/ha)	% Reduction reforestation		Price ($/ha)	Conservation cost	% Increase conservation cost	
	(million ha)	Mean	Std. Dev.	($/1,000 ha)		(0.5 SD)	(2 SD)		(Million $)	(0.5 SD)	(2 SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Amazon:											
11 - Rondonia	0.241	0.18	0.06	0.09	18	4	15	126	18	73	89
17 - Tocantins	0.604	0.13	0.07	1.02	222	2	12	-	-	-	-
21 - Maranhao	1.100	0.11	0.06	0.13	114	5	19	-	-	-	-
51 - Mato Grosso	3.900	0.20	0.12	0.06	198	4	19	-	-	-	-
Cerrado:											
17 - Tocantins	0.238	0.13	0.04	0.06	6	3	12	18	3	44	44
31 - Minas Gerais	0.234	0.34	0.22	0.04	6	3	10	18	3	33	33
35 - Sao Paulo	0.523	0.39	0.22	0.15	90	5	18	130	38	104	137
50 - Mato Grosso do Sul	0.560	0.41	0.43	0.03	18	6	20	30	10	66	66
51 - Mato Grosso	1.600	0.64	0.42	0.05	66	5	18	102	98	87	126
52 - Goias	0.432	0.29	0.22	0.02	6	4	16	6	3	0	0
Atlantic Forest:											
29 - Bahia	0.564	0.19	0.15	0.07	42	5	21	54	19	62	97
31 - Minas Gerais	0.764	0.55	0.48	0.04	30	4	14	42	20	59	104
32 - Espirito Santo	0.179	0.36	0.31	0.13	18	4	16	30	3	55	55
33 - Rio de Janeiro	0.121	0.32	0.22	0.18	18	5	16	18	2	29	114
35 - Sao Paulo	1.000	0.35	0.23	0.09	90	5	17	126	66	90	126
41 - Parana	1.200	0.36	0.19	0.07	78	5	18	102	69	78	119
43 - Rio Grande do Sul	0.218	0.27	0.17	0.11	18	4	14	30	4	50	50
50 - Mato Grosso do Sul	0.433	0.28	0.11	0.21	102	5	20	126	27	103	142

Notes: Table A5 reports 2017 market statistics for all markets with more than 100,000 hectares in forest debt. The forest debt is from Soares-Filho *et al.* (2014). The OC and the marginal OC are estimated using the 2017 census data. Market values are simulated using the empirical function of forestland for each market. We do not report results for the Amazon Para market because of insufficient data on the average yield for the class of cereals (corn, rice, beans, wheat, sorghum, barley, and rye) for the calculation of the rescaling factor across the state of Para.

Robustness analysis

We test the robustness of our empirical models using alternative subsets of the agricultural census data, alternative specifications, different measures of land-use returns in the first stage of the estimation process, and various clustering parameters. We find robust estimates for the agricultural land share model when we add a set of state, biome, and biome/state fixed effects.

Table A6 reports the estimates of the agricultural land share model using different subsets of the agricultural census dataset. Our preferred specification is the full set of commercial farms with an annual gross revenue higher than 2 MW and a farm size larger than 5 ha. We also estimate the model using a dataset of farmers with a gross revenue above 2 MW and farm size larger than 50 ha (table A6, column (2)), a dataset of farms with a gross revenue above 10 MW and farm size larger than 5 ha (table A6, column (3)), and a dataset of large farms with a gross revenue above 10 MW and farm size larger than 50 ha (table A6, column (4)). We find consistent results across these four datasets for the first- and second-stage models, choice of farm type and land-use choice (agriculture versus forest). Few of the coefficients are statistically different across the datasets. Moreover, all the estimated models have fixed effects for states, biomes, and biome/state.

Table A7 presents the estimates of our agricultural land share model when we change the measure of land-use returns in the first-stage model. Our preferred measure is the maximum potential revenue, calculated by first multiplying the maximum potential yield for each crop by the expected price of the commodity and subtracting the transportation cost. The alternative measure is the maximum incremental revenue, calculated by multiplying the expected price by the difference between the maximum and minimum potential yields for the crop (table A7, column (2)). This measure captures the return on crop production with intensive management and advanced technology. The third measure is the maximum potential revenue, the product of the expected price and potential yield (table A7, column (3)). The fourth measure is the maximum net revenue, calculated by subtracting the expected production cost for the crop from the maximum potential revenue. The coefficients of the agricultural land share models using these four measures of land-use return are not statistically different. The coefficients of each land-use return measure in the first-stage model are also not statistically different.

Table A8 reports the estimation results of the agricultural land share equation using different clustering parameters to calculate the standard errors. Owing to the concern that the estimated standard errors are subject to spatial correlation across farms, we use Conley standard errors, which allow for spatial correlation within a specified geographical area defined by a radius parameter. We estimate the model for four radius measures: 50 km, 100 km, 200 km, and 500 km. The estimation of Conley standard errors is computationally intensive, particularly with such a large census dataset and the inclusion of a large set of fixed effects. Therefore, we estimate the models for large farms in Mato Grosso, which has a land area of over 900,000 square km. We find that the standard errors do not change significantly as we increase the radius measure and that most of the estimated coefficients are statistically significant.

Finally, we test six alternative measurements of the estimation of the OC shock using the census agricultural data. The first four measures are estimated for each of the 44 markets in Brazil. Shock 1 is 1 SD of the farm OC. Shock 2 is defined as 2 SD of the farm OC. Shock 3 is the coefficient of variation of the OC, defined as the SD divided by the mean OC. Shock 4 is the difference between the 90th and 50th percentiles of the farm OC. Shock 5 is defined as the average of Shock 1 across all the markets. Our preferred definition of the OC shock is the Shock 5 measure. In our uncertainty analysis, we simulate each market using a lower and upper bound of the OC shock, defined as 0.5 and 2 times Shock 5. Figure A7 shows that the range defined by 0.5 to 2 times the OC shock measure includes most of the other OC shock measures. Figure A7 compares the five measurements of the OC shock. The vertical axis plots the size of the OC shock and the horizontal axis plots its rank. The OC shock across all 44 states ranges from 0 to 45. The OC shock tends to be high in large states with a significant variation in agricultural profitability across farms. The gray area in the graph represents the range of OC shocks defined as 0.5 to 2 times Shock 5. Most measures of the OC shock fall within this range. The exceptions are in the extreme OC shocks, particularly in states with the highest OC shocks. For these states, our preferred OC shock measure, Shock 5, would understate the uncertainty in the OC shock. Shock 5 is thus a conservative measure of uncertainty for these largest states.

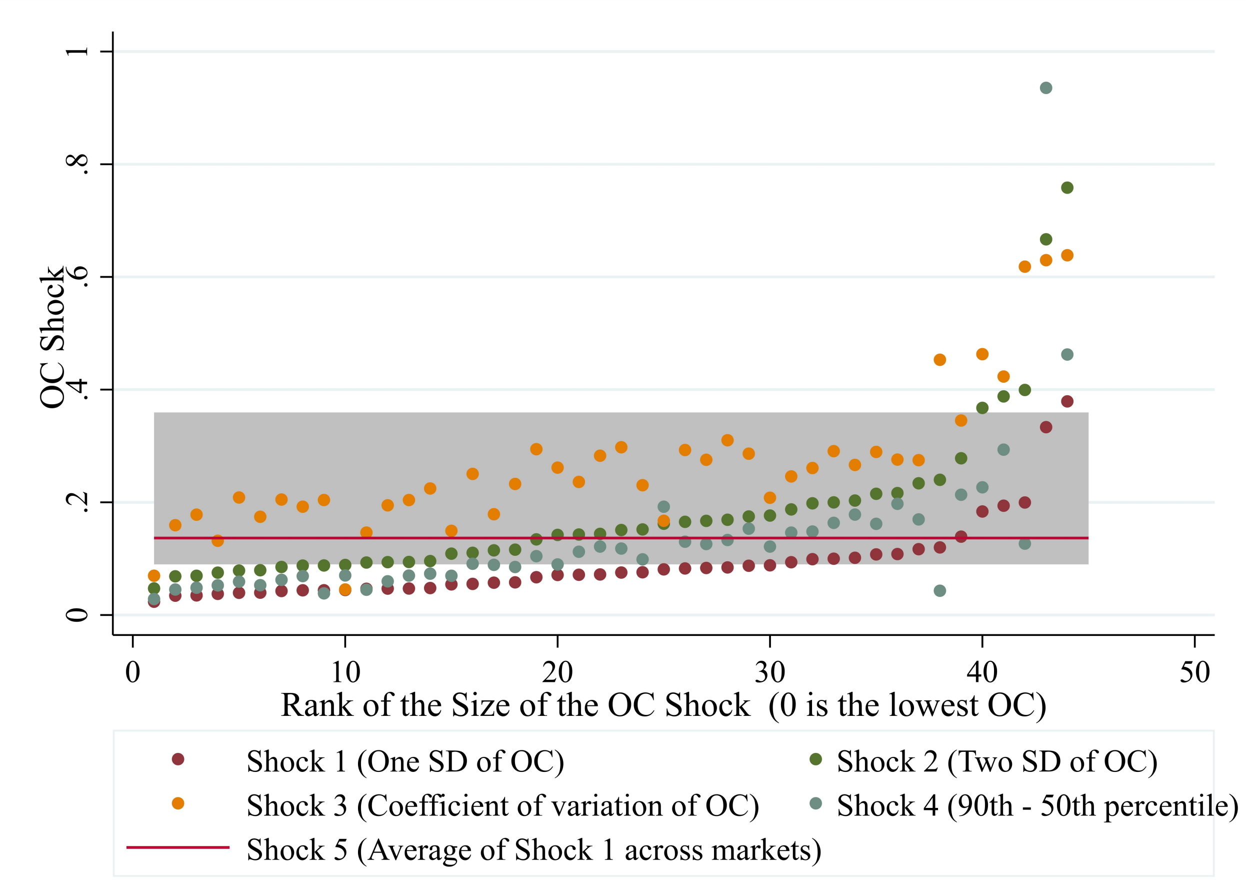


Figure A7: Comparison of alternative measures of the OC shock.

Notes: Figure A7 compares five measurements of the OC shock that we test to model the uncertainty in the OC. The vertical axis measures the size of the OC shock and the horizontal axis measures the rank of the OC shock. 0 in the horizontal axis is the lowest OC shock across all 44 states and 45 is the highest. The first four measures are estimated for each of the 44 markets in Brazil. Shock 1 is one SD of the farm OC. Shock 2 is defined as two SD of the farm OC. Shock 3 is the coefficient of variation of the OC, defined as the SD divided by the mean OC. Shock 4 is the difference between the 90th and 50th percentiles of the farm OC. Shock 5 is defined as the average of Shock 1 across all markets. The OC shock tends to be high in large states with significant variation in agriculture profitability across farms. The gray area in the graph represents the range of OC shocks defined as 0.5 to 2 times the OC shock 5.

Table A6: Comparison of agricultural land share model for alternative samples - Census year 2006

	Baseline	> 50ha	>10MW	>10MW & >50 ha
Variables	(1)	(2)	(3)	(4)
Panel A: Second stage - logistic function of agriculture landshare				
Opportunity cost of forestland	0.780***	0.460***	0.493***	0.550***
	(0.140)	(0.138)	(0.110)	(0.0733)
Transport cost (minimum distance)	0.0184***	0.00922***	0.0180***	0.0151***
	(0.00334)	(0.00328)	(0.00367)	(0.00326)
Municipality in compliance in 1996	-0.660***	-0.805***	-0.578***	-0.615***
	(0.0429)	(0.0493)	(0.0516)	(0.0565)
Municipality with siviculture	-2.321***	-3.252***	-2.592***	-3.159***
	(0.315)	(0.535)	(0.326)	(0.555)
Farm with spring water source	-0.370***	-0.438***	-0.427***	-0.462***
	(0.0160)	(0.0216)	(0.0181)	(0.0208)
Farm with river	-0.324***	-0.425***	-0.331***	-0.392***
	(0.0152)	(0.0189)	(0.0160)	(0.0194)
Farm with lake	-0.0496***	-0.0719***	-0.0731***	-0.0973***
	(0.0159)	(0.0189)	(0.0156)	(0.0164)
Mean elevation	-0.000224**	-0.000168	-0.000202*	-0.000199*
	(9.75e-05)	(0.000107)	(0.000107)	(0.000108)
Standard deviation of elevation	-0.00266***	-0.00324***	-0.00291***	-0.00304***
	(0.000324)	(0.000395)	(0.000418)	(0.000397)
Farm size	-5.79e-05***	-4.95e-05***	-4.72e-05***	-3.64e-05***
	(5.43e-06)	(4.80e-06)	(4.39e-06)	(3.88e-06)
Constant	2.467***	2.557***	2.464***	2.457***
	(0.190)	(0.184)	(0.210)	(0.197)
Biome-state fixed effects	Yes	Yes	Yes	Yes
Observations (number of farms)	1,184,071	435,070	417,940	226,492
R2	0.196	0.169	0.211	0.194
Panel B: First stage - conditional logit model of land use				
Maximum potential net revenue	0.0621***	0.0484***	0.0651***	0.0518***
	(0.00683)	(0.00680)	(0.00949)	(0.00905)
Observations	1,184,071	439,940	421,987	228,680
Chi-squared statistic	1,959	1,951	1,731	1,986

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are bootstrapped with 700 iterations and clustered at the municipality level. The baseline model is our preferred dataset and includes all farms with annual gross revenue greater than 2 minimum wages (MW) and farm sizes larger than 5 ha. Column (2) shows results for farms with gross revenue greater than 2 MW and farm size larger than 50 ha. Column (3) shows results for farms with annual gross revenue greater than 10 MW and farm size larger than 5 ha. Column (4) shows results for farms with annual gross revenue greater than 10 MW and farm size larger than 50 ha. This robustness analysis controls for population density at the microregion level. The main results reported in the main text are based on regressions with controls for the log of population density at the municipality level. The robustness of the OC effect is similar to the two different control variables for population density. The estimation with a more disaggregated control variable for population density leads to a more conservative estimate of the OC effect.

Table A7: Comparison of agricultural land share model for alternative measures of land use returns - Census year 2006

	Baseline	Max Incremental Revenue	Max Revenue	Max Net Revenue
Variables	(1)	(2)	(3)	(4)
Panel A: Second stage - logistic function of agriculture landshare				
Opportunity cost of forestland	0.780***	0.759***	0.787***	0.787***
	(0.140)	(0.157)	(0.139)	(0.139)
Transport cost (minimum distance)	0.0184***	0.0182***	0.0184***	-5.79e-05***
	(0.00334)	(0.00282)	(0.00319)	(5.30e-06)
Municipality in compliance in 1996	-0.660***	-0.661***	-0.656***	-0.660***
	(0.0429)	(0.0426)	(0.0427)	(0.0422)
Municipality with siviculture	-2.321***	-2.319***	-2.428***	-2.320***
	(0.315)	(0.356)	(0.319)	(0.320)
Farm with spring water source	-0.370***	-0.372***	-0.356***	-0.370***
	(0.0160)	(0.0171)	(0.0159)	(0.0158)
Farm with river	-0.324***	-0.324***	-0.311***	-0.324***
	(0.0152)	(0.0144)	(0.0144)	(0.0145)
Farm with lake	-0.0496***	-0.0505***	-0.0353**	-0.0495***
	(0.0159)	(0.0141)	(0.0150)	(0.0155)
Mean elevation	-0.000224**	-0.000218**	-9.84e-05	-0.000226**
	(9.75e-05)	(9.50e-05)	(9.91e-05)	(9.26e-05)
Standard deviation of elevation	-0.00266***	-0.00270***	-0.00276***	-0.00266***
	(0.000324)	(0.000316)	(0.000351)	(0.000344)
Farm size	-5.79e-05***	-5.84e-05***	-5.79e-05***	-5.79e-05***
	(5.43e-06)	(5.04e-06)	(5.30e-06)	(5.30e-06)
Constant	2.467***	2.484***	2.464***	2.464***
	(0.190)	(0.194)	(0.188)	(0.188)
Biome-state fixed effects	Yes	Yes	Yes	Yes
Observations (number of farms)	1,184,071	1,184,071	1,184,071	1,184,071
R2	0.196	0.196	0.196	0.196
Panel B: First stage - conditional logit model of land use				
Land use return measure	0.0621***	0.0694***	0.0622***	0.0622***
	(0.00683)	(0.00866)	(0.00682)	(0.00682)
Observations	1,184,071	1,195,000	1,195,000	1,195,000
Chi-squared statistic	1,959	1,850	1,966	1,966

Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the municipality level. Each column in table A7 reports estimates for the 2nd stage model using alternative measures of land use returns in the 1st stage. Column (1) shows our preferred model estimated using the maximum potential net revenue measure. Standard errors in column 1 are bootstrapped with 700 iterations. Column 2 uses an incremental revenue measure calculated using the difference between the maximum potential yield and the minimum potential yield for each land use. Standard errors in column (2) are bootstrapped with 250 iterations. We reduced the number of iterations in some robustness tests due to the computational intensity of bootstrapping with the large census datasets. Column 3 uses a maximum revenue measure for the return of land uses calculated by multiplying the expected price of each crop by its maximum potential yield in each location. Standard errors in column (3) are not bootstrapped. Column (4) uses a maximum net revenue measure for the return of different land uses calculated by subtracting the expected production of each crop from the maximum revenue. Standard errors in column (4) are not bootstrapped. All models are estimated for our preferred dataset of commercial farms with annual gross revenue greater than 2 MW and farm size larger than 5 ha. This robustness analysis controls for population density at the microregion level. The main results reported in the main text are based on regressions with controls for the log of population density at the municipality level. The robustness of the OC effect is similar to the two different control variables for population density. The estimation with a more disaggregated control variable for population density leads to a more conservative estimate of the OC effect.

Table A8: Spatial correlation test with agricultural land share model for Mato Grosso - Census year 2006

Clustering radius	50 km	100 km	200 km	500 km
Variables	(1)	(2)	(3)	(4)
Second stage - logistic function of agriculture landshare				
Opportunity cost of forestland	0.438*	0.438*	0.438*	0.438*
	(0.227)	(0.227)	(0.236)	(0.245)
Transport cost (minimum distance)	0.00448	0.00448	0.00448	0.00448
	(0.00371)	(0.00371)	(0.00347)	(0.00387)
Municipality in compliance in 1996	-0.375***	-0.375***	-0.375***	-0.375***
	(0.123)	(0.123)	(0.122)	(0.0867)
Municipality with siviculture	-6.744	-6.744	-6.744	-6.744
	(6.139)	(6.139)	(6.267)	(5.555)
Farm with spring water source	-0.196***	-0.196***	-0.196***	-0.196***
	(0.0646)	(0.0646)	(0.0601)	(0.0507)
Farm with river	-0.296***	-0.296***	-0.296***	-0.296***
	(0.0637)	(0.0637)	(0.0638)	(0.0577)
Farm with lake	-0.0977*	-0.0977*	-0.0977*	-0.0977*
	(0.0578)	(0.0578)	(0.0580)	(0.0576)
Mean elevation	0.000299	0.000299	0.000299	0.000299
	(0.000790)	(0.000790)	(0.000873)	(0.000878)
Standard deviation of elevation	0.00607***	0.00607***	0.00607***	0.00607***
	(0.00220)	(0.00220)	(0.00212)	(0.00204)
Farm size	-3.19e-05***	-3.19e-05***	-3.19e-05***	-3.19e-05***
	(4.66e-06)	(4.66e-06)	(4.86e-06)	(4.59e-06)
Constant	0.533	0.533	0.533	0.533
	(0.438)	(0.438)	(0.420)	(0.439)
Biome-state fixed effects	No	No	No	No
Observations (number of farms)	13,599	13,599	13,599	13,599
R2	0.447	0.447	0.447	0.447

Notes: *** p¡0.01, ** p¡0.05, * p¡0.1. Table A8 reports estimates for an agricultural land share model, the 2nd stage in our empirical approach, for the state of Mato Grosso in Brazil using different clustering parameters. We restrict our analysis to the state of Mato Grosso because the clustering tests are computationally intensive with the large census dataset. Also, all models are estimated without fixed effects due to computational restrictions in the calculation of Conley standard errors with fixed effects and the large census dataset (Conley, 1999). The state of Mato Grosso is the largest grain producer in Brazil and has an area of over 900 thousand km2. We test for spatial correlation using Conley spatial standard errors (20). Conley standard errors account for the spatial correlation of units within a radius measured in km. We estimate the Mato Grosso agricultural land share model with four different clustering radius: 50 km, 100 km, 200 km, and 500 km. This robustness analysis controls for population density at the microregion level. The main results reported in the main text are based on regressions with controls for the log of population density at the municipality level. The robustness of the OC effect is similar with the two different control variables for population density. The estimation with a more disaggregated control variable for population density leads to a more conservative estimate of the OC effect.

References

Alves, **E.**, **Souza**, **G. d. S.**, **Rocha**, **D. d. P.**, et al. (2013). Desigualdade nos campos na ótica do censo agropecuário 2006. *Revista de Política Agrícola*, *22*(2), 67–75.

Ben-Akiva, **M. E.**, & **Lerman**, **S. R.** (1985). *Discrete choice analysis: Theory and application to travel demand* (Vol. 9). MIT press.

Berry, **S. T.** (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242–262.

Bustos, **P.**, **Caprettini**, **B.**, & **Ponticelli**, **J.** (2016). Agricultural productivity and structural transformation: Evidence from brazil. *American Economic Review*, *106*(6), 1320–65.

Cardell, **N. S.** (1997). Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory*, *13*(2), 185–213.

DePaula, **G.** (2023). Bundled contracts and technological diffusion: E vidence f rom the brazilian soybean boom. *Journal of Development Economics*, *165*, 103163.

Dong, **Y.**, & **Lewbel**, **A.** (2015). A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews*, *34*(1-2), 82–105.

Embrapa. (2014). Climate and soil characteristics at municipality in brazil. socio-economics characteristics at microregion in brazil.

ESALQ. (2009). Sistema de informacoes de frete – anuario 2009. [Sistema de Informacoes de Frete – Anuario 2009. Coordinator Jose Vincente Caixeta Filho - Luiz Queiroz College of Agriculture – University of Sao Paulo.].

IBGE. (2006). Censo agropecuário - ano 2006. [Available at Instituto Brasileiro de Geografia e Estatística.].

IBGE (2017). Censo agropecuário - ano 2017. [Available at Instituto Brasileiro de Geografia e Estatística.].

IIASA (2018). Global agro-ecological zones (gaez v3. 0). [Available at http://webarchive.iiasa.ac.at/Research/LUC/GAEZv3.0/.].

Lewbel, **A.** (2014). An overview of the special regressor method.

Lubowski, **R. N.**, **Plantinga**, **A. J.**, & **Stavins**, **R. N.** (2006). Land-use change and carbon sinks: Econometric estimation of the carbon sequestration supply function. *Journal of environmental economics and management*, *51*(2), 135–152.

Soares-Filho, **B.**, **Rajão**, **R.**, **Macedo**, **M.**, **Carneiro**, **A.**, **Costa**, **W.**, **Coe**, **M.**, **Rodrigues**, **H.**, & **Alencar**, **A.** (2014). Cracking brazil's forest code. *Science*, *344*(6182), 363–364.