Supplement to 'Deforestation in the Amazon: A Unified Framework for Estimation and Policy Analysis'

EDUARDO SOUZA-RODRIGUES[*]
Department of Economics, University of Toronto

FOR ONLINE PUBLICATION

November 2018

1 Introduction

In this supplemental document, I assess the robustness of the estimated regressions presented in the main paper (Section 2), and provide a detailed description of how the data set is constructed (Section 3).

Section 2 is divided into the following six subsections. In Subsection 2.1, I discuss the estimated coefficients on the covariates in the land use quantile regressions. In Subsection 2.2, I present the estimated results when excluding the proxies for monitoring effort: the number of fines issued and the distance to the nearest IBAMA office (the Brazilian Environmental Protection Agency). Consistent with the discussion in the main paper (that, historically, enforcement in the Amazon rainforest had been low), the inclusion and exclusion of these variables does not substantially affect the coefficients on costs to ports, as well as on other regressors. In Subsection 2.3, I explore how spatial dependence can affect the results. Coefficients on spatially lagged covariates are statistically significant, but the results are not sensitive to the choice of weighting matrix nor to the cutoff distance. Subsection 2.4 investigates whether the choice of instruments drives the main results. I drop distance to capital as an instrument for costs to ports, and find that the policy implications are robust to the choice of instrument. Subsection 2.5 drops the logit assumption in the paper

[*] University of Toronto, Department of Economics. E-mail: e.souzarodrigues@utoronto.ca. Financial support from Charles V. Hickox Fellowship, Arthur K. Watson Fund, Yale University, University of Toronto Mississauga, and Kernan Brothers Environmental Fellowship is gratefully acknowledged. All errors are mine.

and estimates a semiparametric model. The results establish the suitability of the logit model given that there are no significant differences between the logit and the semiparametric models. Finally, in Subsection 2.6, I discuss the use of different productivity indices to calculate the demand for deforestation. I describe the underlying economic model associated with the different indices briefly, present the corresponding estimated demands for deforestation, and provide evidence that local yields do not respond to changes in transportation costs.

The data set is described in Section 3. It includes a detailed explanation as to how the share of deforestation (Subsection 3.1), transportation costs (Subsection 3.2), productivity indices (Subsection 3.3), and covariates (Subsection 3.4) were constructed. In Subsection 3.1, I also compare deforestation calculated using the Brazilian Agricultural Census (my chosen measure) with satellite-based deforestation measures. I discuss their differences and similarities, and show that when they are most comparable (i.e., when the share of private land in the municipality is large and the share of clouds/unobserved areas in satellite images is small), their correlation in the data is high. Furthermore, I find no evidence that the census deforestation measure varies systematically with the exogenous variables after conditioning on the satellite deforestation measure.

2 Robustness

2.1 Land Use Quantile Regressions – Main Specification

In this subsection, I discuss the estimated coefficients on the covariates for the main specification briefly. Table 1 presents the results for the IVQR estimator at the median. For brevity, I omit the coefficients on the spatially lagged regressors, but present and discuss them in Subsection 2.3.

First, the coefficients on altitude are small in magnitude and not significant. Second, higher levels of temperature and precipitation in the Amazon are worse for agriculture, as expected. The already high levels of temperature and rainfall in the region make agriculture unattractive because, among other reasons, cattle are more susceptible to parasites and insect pests, crops are more subject to rotting, yields are depressed by light-limiting cloud cover, mechanization is difficult, and forest burning is incomplete (Chomitz and Thomas (2003)). Third, locations with steeper slopes are worse for agriculture as well (although not significantly so). Fourth, soils with worse quality induce less deforestation when compared to good soils, as expected, although the effects are not monotonic in the rank of soil quality (the proportion of good soils is omitted in the regressions).

As discussed in the main text, protected areas (PAs) may have spillover effects that affect the value of nearby forestlands. Except for large farms, the coefficients on distance to protected areas are small in magnitude and are not statistically significant. Most of them have a positive sign: the smaller the distance, the lower the deforestation. This is consistent with Herrera (2015), who finds causal evidence that protected areas in the Brazilian Amazon reduce deforestation in surrounding areas. Herrera (2015) proposes (and finds evidence for) two potential mechanisms for that: PAs may increase outmigration and decrease infrastructure development, both of which can reduce profits from agricultural activities. My estimated results are also consistent with the regularity that protected areas tend to be located on land facing lower deforestation pressure (as documented by Pfaff, Robalino, Herrera, and Sandoval (2015)), and the possibility that the aggregated nature of the data may mask multiple heterogeneous spillover effects. (Robalino, Pfaff, and Villalobos-Fiatt (2017) present evidence of heterogeneous spillovers in Costa Rica, using detailed disaggregated data.)

Controls for mining, the presence of power plants either in the municipality or within 75 km from the municipal seat, and the local population are meant to capture factors that influence local demand and supply for non-tradables. The signs of the impacts are not clear *ex-ante*, as they depend on how these factors affect the demand for local products, the demand and supply for labor, and which sector (agriculture or forestry production) is more labor-intensive. Most of the estimated coefficients are negatively correlated with the share of agricultural land, but they are not significant in most cases.

While the location of power plants may depend on siting decisions related to deforested areas, these decisions depend fundamentally on both local demand for electricity and cost factors (see, e.g., Lipscomb *et al.* (2013)). Demand for electricity is driven mostly by the size of the local population, and the costs of hydropower dam construction depend mainly on topographic factors. Given that most of these factors are controlled for in the land-use regressions, the location of power plants is, arguably, plausibly exogenous to farmers' land use decisions.

One might also be concerned with the potential endogeneity of population. I discuss robustness results related to local population below and why it may be difficult to estimate effects of factors that affect local labor market conditions in the Amazon rainforest.

I also discuss the robustness of the results with respect to the proxy for property rights, and with respect to the proxies for monitoring efforts (the number of fines and the distance to the nearest IBAMA agency). Here, I emphasize only that their coefficients are not statistically significant.

Robustness: Local Population. I now investigate whether the exclusion of local population in the land use regressions affects the main results. Table 2 compares the estimated coefficients on the regressors for each farm size. I present results only for the median, but the pattern is similar for other quantiles. As mentioned previously, the local population does not have significant coefficients. Furthermore, omitting it in these regressions does not affect the coefficients on the other regressors significantly. The population in the Amazon is sparsely distributed and so labor is likely a scarce factor in the region. Medium-sized and large farms are therefore likely to have production functions that are intensive in land, but not in labor. Indeed, the average number of agricultural workers per hectare of agricultural land is: 4/ha for small farms; 0.24/ha for small–medium farms; 0.054/ha for medium–large farms; and 0.012/ha for large farms. Therefore, the share of wages in costs is probably small, which may help explain why factors that shift the demand for and supply of labor are not significant in the regressions presented in Table 1. So, even though one may argue that the population is endogenous, instrumenting for it likely does not change the primary results.[1]

Robustness: Proxy for Property Rights. As discussed in the main text, a possible problem that may invalidate the instruments is the potential lack of property rights. For this reason, I include a proxy for property rights in the regressions. The best proxy for property rights in the data is the proportion of private land with a land title – presumably, the higher the tenure security, the larger is the proportion of land with land titles. However, the ‘tenure security’ proxy may be endogenous because farmers had incentives to deforest as a way to secure their land tenure – i.e., locations with more deforestation may have led to a higher proportion of land titles, which means the tenure proxy may suffer from simultaneity bias. Finding sources of exogenous variation for property rights within a country is an extremely difficult task – there are no clear instruments to hand. Yet, the inclusion or exclusion of the proxy for property rights in the land-use regressions does not affect the estimates significantly. Table 3 presents the results for the median regressions with and without the proportion of private land with land title. Although both estimates could be equally biased, the results are reassuring: most farmers do have their land titles (85 percent) and it is possible that they may not need to deforest as much to guarantee their property rights, or at least their land tenure status may not affect how their deforestation decisions relate to transportation costs.

[1] One may expect local labor markets to be imperfect, perhaps as a result of the sparsely distributed population. A symptom of this imperfection is that farmers rely more on family labor than on hired labor. Indeed, the proportion of family workers in the total number of workers averaged 83% in the Amazon, ranging from a low of 10% in more developed regions in the South Amazon to a high of 99%–100% in the isolated areas of the Western Amazon.

2.2 Penalties and Monitoring Efforts

I consider two proxies for monitoring efforts: the total cumulative number of fines (over the period 2002–2005), and the distance to the nearest IBAMA agency. The number of fines may capture the recent increase in penalties, while the distance to IBAMA may capture more permanent monitoring efforts. Table 4 presents the results for the IVQR median regressions. It omits the coefficients on all the regressors except for the costs to ports and the proxies for penalties. The first column of the table presents the specification that excludes both proxies; the second column includes only the number of fines; the third column includes only the distance to IBAMA; and the last column includes both (the main specification).

The estimated coefficients on both proxies are not statistically significant, and the magnitudes are small. Specifically, when covariates are fixed at the sample mean and when I increase the number of fines by 10 (corresponding to approximately one-fifth of its sample standard deviation) the estimated marginal effects range from negligible to a 0.5 percentage point decrease in the share of deforestation, depending on the farm size and the quantile. Recall that the average number of fines in the data is 23; increasing that amount by 10 is therefore a substantial change in the number of fines.

Although the signs of the coefficients are in the expected direction, it is clear that the number of fines may suffer from simultaneity problems, which may cause upward biases on their estimated coefficients (i.e., the estimated magnitudes are closer to zero). It is a nontrivial task to find suitable instruments for environmental transgressions in a cross-section of Amazonian municipalities; and the biases on the coefficients on costs to ports caused by this simultaneity problem are not clear *ex-ante*. Against that, ignoring the number of fines in the regressions may lead to omitted variables biases. The direction of the omitted variables bias depends on the correlation between fines and the instruments for TC_m. In the data, this correlation is small, which suggests the omission may result in small biases.[2] The fact that the inclusion and the exclusion of the number of fines do not affect substantially the coefficients on costs to ports, as well as on other regressors, is reassuring. This is consistent with the interpretation that the increased monitoring efforts were too recent to substantially affect the total accumulated deforested land.

The coefficients on distance to IBAMA are in the expected direction for medium-sized and large farms: the greater the distance, the lower the monitoring, and so the larger the deforested

[2] The raw correlation between number of fines and straight-line distance to ports is 0.1; and the raw correlation with distance to the nearest capital is 0.04. When regressing the total number of fines on the instruments for costs to ports and on all other covariates, I obtained small and insignificant coefficients for the instrumental variables.

area. However, similar to the number of fines, distance to IBAMA may suffer from a simultaneity problem, given that highly deforested locations may lead to the presence of IBAMA offices on nearby areas. This may bias the estimated coefficients downwards (i.e., estimated values are closer to zero, or negative). Still, ignoring this variable may result in omitted variables bias instead.

Although I do not have information about how the placement decisions of IBAMA offices were made exactly, Figure 1 provides relevant geographic information that can shed light on the issue. The left panel of Figure 1 presents a map with the geographic location of the IBAMA agencies, together with the Amazonian state capitals. The right panel adds to the map the deforested areas in 2006. It is apparent that the agencies are located throughout the Brazilian Amazon. Perhaps not surprisingly given this geographic distribution, the raw correlation between distance to IBAMA offices and distance to ports is small and insignificant in the data (specifically, the correlation is 0.015). Given that distance to ports is the main instrumental variable driving the estimated results (see Section 2.4), the small correlation suggests that omitting the distance to IBAMA in the land use regressions is likely to result in only very small biases. Again, similar to the number of fines, the coefficients on distance to IBAMA are not statistically significant, their magnitudes are small, and the inclusion and exclusion of this variable does not change results significantly. Because distance to IBAMA captures more permanent monitoring effects, this is consistent with the interpretation that the legislation has not been fully enforced historically – at least up to 2006.

Policy Implications. Table 5 presents the policy implications for the different specifications. The first column presents the value of the uniform tax that would lead to 20 percent of agricultural land in private properties in the Amazon; the second column presents the associated farmers' lost surplus from the uniform tax; the third column presents farmers' lost surplus under the (perfectly enforced) '80 percent rule;' and the fourth column, the amount of avoided emissions of carbon under a carbon tax of one dollar per ton of CO_2. The top panel presents the case in which the counterfactual policies are superimposed on the status quo scenario (denoted here by 'actual monitoring' counterfactual); the bottom panel considers the case in which the total number of fines during 2004–2005 is fixed at the same level as in the previous period 2002–2003. That is, it "shuts down" the recent increase in the monitoring efforts (see Section 7.2 in the main text).

In the counterfactual in which taxes are superimposed on the status quo scenario, the values of the uniform tax and their associated lost surpluses vary little across specifications: uniform taxes range from US$ 37.50/ha to US$ 42.50/ha; and lost surpluses, from US$ 425 million to US$ 482 million. The lost surpluses from the '80 percent rule' range from US$ 3.08 billion to US$ 4.81,

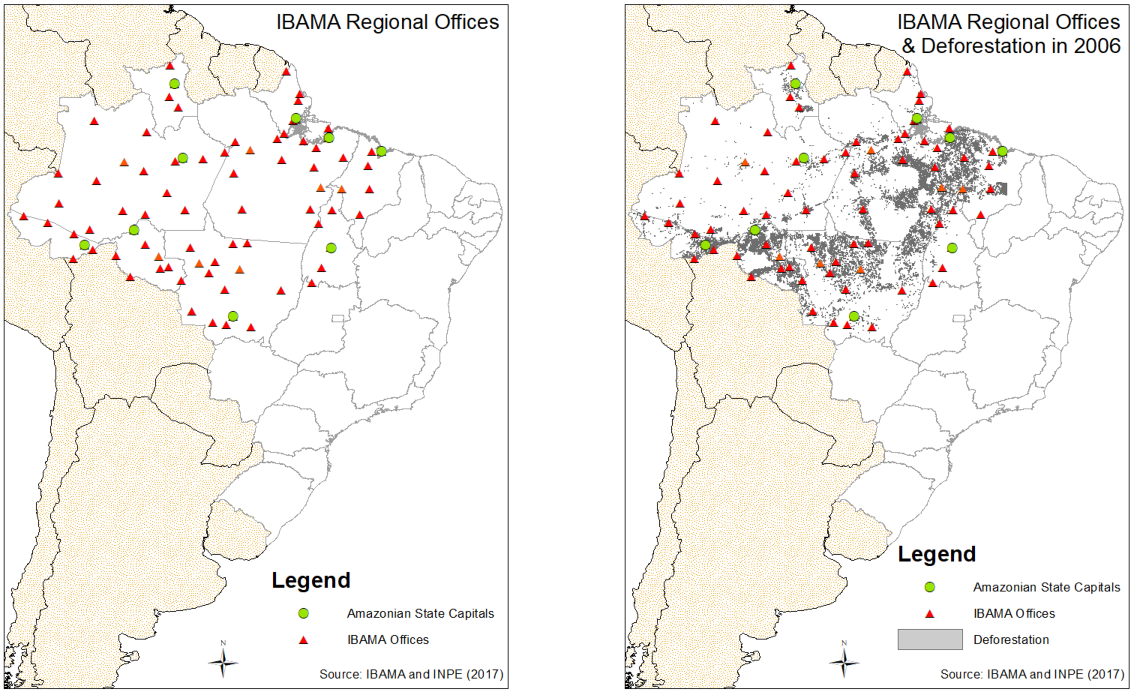


Figure 1: Location of IBAMA Regional Offices and Deforestation

which is approximately 7 to 9 times more expensive than the uniform taxes for farmers. The avoided emissions are all close to each other: between 4.17 billion and 4.21 billion tons of carbon. The policy implications are therefore robust to the inclusion and exclusion of the proxies for monitoring in the model specification.

For the counterfactual in which the number of fines post-2004 is fixed at pre-2004 levels, I consider only the main specification and the specification with no distance to IBAMA in the regressions. Both specifications deliver similar results. The value of the uniform tax increases to US$ 47.50/ha, compared to US$ 42.50 in the 'actual monitoring' counterfactual. The higher tax is not surprising given that the burden of reducing deforestation is much more relegated to the counterfactual policy. Similarly, the lost surpluses from the tax, from the '80 percent rule,' and the amount of avoided emissions also increase compared to the 'actual monitoring' counterfactual, but they are all of comparable magnitudes. Note that the estimated lost surpluses consider only the losses from the counterfactual policy and ignore the benefits farmers enjoy from reducing the existing monitoring efforts.

2.3 Spatial Dependence

There are different potential sources of spatial dependence in the data. I consider several possibilities, which I discuss briefly now.

First, I have included spatially lagged regressors of the local demand and supply shifters in the land use equations because inter-related local markets may create spatial dependence in farmers' decisions across municipalities. Spatial lags for proximity to protected areas (PAs) are also included among regressors to allow for further spatial spillover effects of PAs on deforestation.

I have chosen to not include a spatially lagged dependent variable among the regressors in my final specification. When I estimated such a model using spatially lagged regressors as instrumental variables, I obtained unstable and unreasonable estimates. Such a model would capture direct interactions between farmers across municipalities. My interpretation is that this type of interaction across municipalities seems less likely or less important than interactions through equilibrium prices or local neighbor interactions.

Although local interactions among neighbor farmers are likely important (evidence of that can be found in Robalino and Pfaff (2012) and the references cited therein), the aggregated nature of the present data set limits my ability to incorporate them in the analysis. Another possible specification involves interactions among farms of different sizes in the same municipality. Doing so would require an extended model with a simultaneous system of equations in which deforestation in a given type of farm would directly affect the land use on other types of farm. To identify and estimate such a system, I would have to find instrumental variables satisfying exclusion restrictions as in standard simultaneous equations problems. Finding such instruments is challenging; they are not available in the present data set (none of the regressors are farm-size specific). I therefore interpret my estimated results as the (reduced-form) equilibrium outcome of these local interactions.

Table 6 presents the results for the IVQR median regressions when spatial dependence is taken into account. As before, the table omits the coefficients on all the regressors except for the costs to ports and the spatially lagged regressors. The first column presents the specification that ignores spatial dependence; the second column includes the spatially lagged regressors and considers a cutoff distance of 50 km; the third column considers a cutoff distance of 75 km (the main specification); and the fourth column, a cutoff distance of 100 km. The spatial weight matrices are of the power functional type with a distance-decay parameter that equals one. Most of the estimated coefficients on the spatially lagged regressors are not statistically significant; an exception is the presence of mining for medium-sized and large farms. As previously discussed in Subsection 2.1 of this supplemental document, it is difficult to predict the direction of the impacts of the local demand and supply shifters. The results are not highly sensitive to the choice of the cutoff distance.[3]

Finally, it is conceivable that there exists spatial correlations in the unobservables, which may affect the standard errors of the parameters estimates. Although there is no currently available technique that incorporates such correlations into instrumental variables quantile regressions formally (as far as I am aware), I have implemented a geographic cluster-bootstrap variance matrix estimate to investigate this issue. To implement this bootstrap, I obtained G geographic clusters of municipalities (see more on that below), and executed the following steps B times: (a) I formed G clusters by resampling with replacement from the original sample of clusters, (b) I estimated the IVQR model for each b^{th} bootstrap sample, and then (c) given the B estimates of the vector of coefficients, I computed the variance-covariance matrix of the estimated coefficients across the bootstrap samples. This is known as the "pairs cluster bootstrap" (Cameron and Miller (2015)).

Given that there is no unambiguous way to define geographic clusters in the present context, I opted for two definitions. First, I make use of the official division of the Brazilian territory. The country is divided in 'immediate regions,' which are areas that are smaller in size than states but larger than municipalities (IBGE (2017)). Specifically, immediate regions are groups of adjacent municipalities that are clustered together taking into account their geographic features, the physical connection of nearby cities (logistics), and the economic and social relations between them (e.g., trade of durable and non-durable consumer goods; typical geographic reach of workers' job search; local demand for health and education services; and the provision of public services, such as service stations of the National Social Security Institute and the Ministry of Labor, as well as judicial services, among others). The regional division has the objective of helping planning and managing public policies at federal and state levels. The number of immediate regions (clusters) in the data is $G = 82$, and the average number of municipalities per cluster is 6.4 (with a standard deviation of 3.5).[4] I have not clustered the data at the state level because it would lead to the "few clusters" problem (Cameron and Miller (2015)), given that there are only nine states in the Brazilian Amazon.

For the second definition, I created geographic clusters using an unsupervised machine learning algorithm in ArcGIS. For a given number of clusters G, the algorithm searches for a solution in

[3] Consistent with the discussion in Le Sage (2014), results are robust to the choice of the weighting matrix. In light of the results presented in Assunção, Lipscomb, Mobarak, and Szerman (2016), I kept the presence of power plants in neighbor municipalities in all specifications.

[4] The immediate regions were established in 2017. They correspond to a revision of the old microregions, which were set in 1989. The microregions overlap but do not perfectly coincide with the immediate regions. I have also computed the cluster-bootstrap standard errors using the microregions as clusters. The results are essentially the same.

which observed features within each cluster are as similar as possible, while the clusters themselves are as different as possible. To guarantee that clusters are formed by adjacent municipalities, I impose the constraint that municipalities can only be part of the same group if they share an edge. The number of clusters I consider varies from 50 to 150 (with an average number of municipalities per cluster varying from 10.4 to 3.5).[5]

Table 7 shows the results. For each farm size and quantile, I present both the IVQR standard errors as developed by Chernozhukov and Hansen (2008) (the main specification presented in the paper), and the cluster-bootstrap standard errors as described above. For brevity, I present results for the second definition of clusters (based on the unsupervised algorithm) only for $G = 50$ and $G = 150$, as the results for different number of clusters are qualitatively similar.[6] I implemented $B = 500$ bootstrap samples. As expected, the IVQR standard errors are smaller than the cluster-bootstrap standard errors for almost all farm sizes and quantiles. However, most coefficients that are statistically significant under the IVQR standard error remain significant under the cluster-bootstrap standard errors. Overall, the qualitative results are similar for both types of standard errors.

It is important to stress that, while intuitively appealing, the cluster-bootstrap standard errors are not necessarily correct. As mentioned previously, there exists no formal theoretical result that incorporates spatial correlations of unobservables into instrumental variables quantile regressions. Further, Abadie, Athey, Imbens and Wooldridge (2017) have recently shown in the context of treatment effects models that cluster-robust standard errors tend to be too large unless (a) the number of clusters sampled is very small relative to the number of clusters in the population, or (b) there is no treatment effect heterogeneity. They argue that conventional cluster adjustments assume (often implicitly) that the clusters in the sample are only a small fraction of the clusters in the population of interest, and that this is not common in economic applications – on the contrary, it is common to have most, if not all, clusters in the population sampled.[7] Conditions (a) and (b) are not satisfied in the present case, so it is conceivable that the estimated cluster-robust standard errors presented in Table 7 are unnecessarily conservative. On a positive note, the fact that the inference results are qualitatively similar for the different types of standard errors is reassuring.

[5] Specifically, I use the ArcGIS toolbox "Grouping Analysis." The algorithm employs a connectivity graph (minimum spanning tree) to find natural groupings. The observed features that I consider to create the groups are the latitude and longitude of the municipalities' centroids.

[6] The number of clusters that I consider are $G = 50, 75, 100, 125$, and 150. The cluster-bootstrap standard errors tend to increase when the number of clusters decreases, but not substantially so (nor monotonically).

[7] Abadie, Athey, Imbens and Wooldridge (2017) also show that, in some cases, heteroskedasticity-robust standard errors result in more accurate inferences than cluster-robust standard errors, even in the presence of substantial intra-cluster correlation.

2.4 Instrumental Variables

In this section, I investigate briefly whether the estimated results are sensitive to the use of distance to capital as an instrument for costs to ports. To that end, I drop 'distance to capital' as an instrumental variable, re-estimate the model parameters, and simulate the counterfactual policies, using distance to ports as the only instrument for costs to ports. (Note that by dropping one instrument, the model becomes just-identified.)

Table 8 presents the results for the 2SLS and IVQR estimators. For each farm size, I present the estimated coefficients on costs to ports using distance to ports as the only instrument ("IV – Dist to Ports"), and using both straight-line distances as instruments ("IV – Both Instruments"), which corresponds to the main specification presented in the paper. Both specifications produce very similar results for all farm sizes. Table 9 shows the policy implications in terms of (a) the value of the uniform tax that would lead to 20 percent of agricultural land in private properties, (b) the corresponding farmers' lost surplus from the uniform tax, (c) the farmers' lost surplus under the (perfectly enforced) '80 percent rule,' and (d) the amount of avoided emissions of carbon under a carbon tax of one dollar per ton of CO_2. All policy implications are robust to the choice of instruments. I conclude that the results are driven primarily by the distance to ports as the relevant instrument for costs to ports.

2.5 Semiparametric Model

As a robustness exercise, I relax functional form restrictions by dropping the logit assumption to check whether it may drive the results. Specifically, I estimate the semiparametric model:

$$G_s\left(Y_m\left(s\right), u\right) = X_m \beta_{su} - TC_m, \tag{1}$$

where the function $G_s\left(., u\right)$ is unknown. Because scale and location normalizations are necessary, the coefficient on transportation cost is normalized to be minus one and the constant, to be zero. The normalization is without loss as long as TC_m affects deforestation negatively.

2.5.1 Implementation of the Estimator

To estimate the model (1), I use the Penalized Sieves Minimum Distance (PSMD) estimator proposed by Chen and Pouzo (2009, 2012). I assume (Z_m, X_m) is independent of $U_m\left(s\right)$ for any s, where $Z_m = (Z_m^p, Z_m^c)$ denotes the straight-line distances to the nearest port and to the nearest capital. For each farm size s and quantile $u \in (0,1)$, the moment restriction implied by the model is:

$$E\left[\rho_{su}\left(Y_m\left(s\right), X_m, TC_m; G_{su}, \beta_{su}\right) \mid Z_m, X_m\right] = 0, \tag{2}$$

where the residual function is

$$\rho_{su}\left(Y_m\left(s\right), X_m, TC_m; G_{su}, \beta_{su}\right) \equiv 1\left\{G_s\left(Y_m\left(s\right), u\right) \leq X_m\beta_{su} - TC_m\right\} - u, \tag{3}$$

and the conditional moment function is

$$m\left(Z_m, X_m; G_{su}, \beta_{su}\right) \equiv E\left[\rho_{su}\left(Y_m\left(s\right), X_m, TC_m; G_{su}, \beta_{su}\right) \mid Z_m, X_m\right]. \tag{4}$$

I approximate G_{su} using an artificial neural network (ANN) sieve approximation, because this non-linear sieve is often, in practice, better able than alternatives to allow for nonlinearities in the unknown function (Chen (2007)). Specifically, I use a sigmoid ANN defined by:

$$sANN\left(k_m\right) = \left\{{}_{j=1}^{k_m}\alpha_j S\left(\gamma_j Y_m + \gamma_{0,j}\right) : \alpha_j, \gamma_j, \gamma_{0,j} \in \mathbb{R}\right\} \tag{5}$$

where $S : \mathbb{R} \rightarrow \mathbb{R}$ is a sigmoid activation function. Note that the dimension of $sANN\left(k_m\right)$ is $3k_m$. I opted for a Gaussian activation function, i.e., I take $S\left(.\right)$ to be a normal distribution function.

For each s and u, the function G_{su} in $sANN\left(k_m\right)$ and the finite dimensional parameter β_{su} are chosen to minimize the criterion function:

$$Q\left(G_{su}, \beta_{su}\right) = \left\{\frac{1}{M}\sum_{m=1}^{M}\widehat{m}\left(Z_m, X_m; G_{su}, \beta_{su}\right)'\,\widehat{m}\left(Z_m, X_m; G_{su}, \beta_{su}\right) + \lambda_m\widehat{M}\left(G_{su}, \beta_{su}\right)\right\}$$

where M is the number of municipalities; $\widehat{m}\left(.\right)$ is an estimator for $m\left(.\right)$; the penalization parameter, $\lambda_m \geq 0$, converges to zero as $M \rightarrow \infty$; and $\widehat{M}\left(G_{su}, \beta_{su}\right)$ is the penalization function.

I take $\widehat{m}\left(.\right)$ to be a series least square estimator of $m\left(.\right)$. Let $\left\{p_1\left(Z_m, X_m\right), p_2(Z_m, X_m), ...\right\}$ be a sequence of known basis functions that approximate any square integrable real-valued function. Denote $p^{J_M}\left(Z_m, X_m\right) = \left(p_1\left(Z_m, X_m\right), ..., p_{J_M}\left(Z_m, X_m\right)\right)'$ a $(1 \times J_M)$-vector, and define the $(M \times J_M)$ matrix $P = \left(p^{J_M}\left(z_1, x_1\right), ..., p^{J_M}\left(z_M, x_M\right)\right)'$. The series least square estimator is given by:

$$\widehat{m}\left(z, x; G_{su}, \beta_{su}\right) = p^{J_M}(z, x)'(P'P)^{-}\sum_{m=1}^{M}p^{J_M}(Z_m, X_m)\rho_{su}\left(Y_m\left(s\right), X_m, TC_m; G_{su}, \beta_{su}\right)$$

where $(P'P)^{-}$ is the pseudo-inverse matrix of $P'P$. The penalization function used is:

$$\widehat{M}\left(G_{su}, \beta_{su}\right) = \left\|\nabla_y^2 G_{su}\right\|_{L^2\left(\widehat{f}_{Y(s)}\right)}$$

where ∇_y^2 denotes the second derivative with respect to y and $\|.\|_{L^2\left(\widehat{f}_{Y(s)}\right)}$ denotes the L^2-norm with the empirical measure $\widehat{f}_{Y(s)}$ of $Y_m\left(s\right)$.

Next, I discuss the specific approximations I adopted. Because the asymptotic theory provides guidance on the rate at which k_m must increase with the data, but not the specific value for k_m, I choose k_m so that the number of coefficients estimated does not exceed the number of moment restrictions.[8] Specifically, the total number of parameters to be estimated is $\dim\left(\beta\right)+3k_m$. The number of covariates in the regressions is 20 (including the spatially lagged regressors), so $\dim\left(\beta\right) = 20$. By choosing $k_m = 3$, the total number of parameters becomes 29. To approximate the conditional moment $m\left(Z_m, X_m\right)$, I use a basis function $p^{J_M}\left(Z_m, X_m\right)$ with dimension $J_M = 31$, so that the number of moment restrictions ($J_M = 31$) is greater than the number of model parameters ($\dim\left(\beta\right) + 3k_m = 29$). With respect to the penalization parameter, I use $\lambda_m = 10^{-6}$.

Although the nonparametric function G_{su} depends only on the scalar $Y_m\left(s\right)$, approximating the function $m\left(Z_m, X_m\right)$ is difficult because of the curse of dimensionality. For this reason, I had to be very parsimonious in choosing the number of terms for $p^{J_M}\left(Z_m, X_m\right)$. I opted to use P-Splines(2,4) for both Z_m^p and Z_m^c, i.e., quadratic splines with four knots (chosen at the respective 0.01, 0.25, 0.75 and 0.99 quantiles of Z_m^p and Z_m^c), and an interaction term $Z_m^p \times Z_m^c$. Each spline has dimension 5; the two splines plus the interaction therefore result in 11 terms. The remaining exogenous terms, X_m, enter linearly in $p^{J_M}\left(Z_m, X_m\right)$, yielding the dimension $J_M = 31$. When I experimented by approximating G_{su} with higher dimensions, say, $k_m = 4$, the dimensions of the splines for $Z_m = \left(Z_m^p, Z_m^c\right)$ were also increased to cubic splines, i.e., P-Splines(3,4). Including many more higher order terms for any of the instruments in $p^{J_M}\left(Z_m, X_m\right)$ made the estimation routine more difficult, because the matrix $\left(P'P\right)$ becomes singular very quickly. Note that to restrict the function $m\left(Z_m, X_m\right)$, it would be necessary to specify the conditional distribution of the endogenous variables $\left(Y_m\left(s\right), TC_m\right)$ given the exogenous $\left(Z_m, X_m\right)$. Such a specification is not imposed by the model.

In the estimation routine, I used the estimates from the IVQR model as an initial guess for β_{su} and the best $sANN\left(k_m\right)$ approximation for the rescaled logistic function as an initial guess for

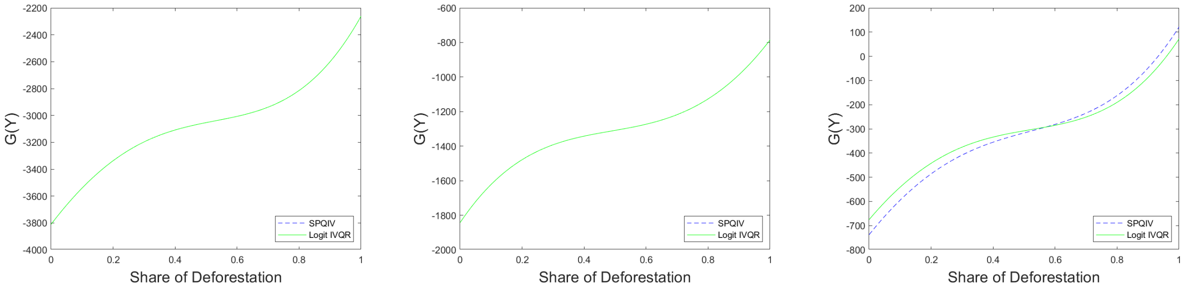
[8] Chen and Pouzo (2012) provide very general results for the rates of convergence for NPQIV models. However, to date, there exists no explicit rate of convergence in terms of the number of observations in the data for non-linear sieves such as the *sANN* used here.

G_{su} (where the rescaling used IVQR estimates). Therefore, as the minimization routine proceeds, it indicates whether or not the estimates of the IVQR are good estimates for the SPQIV. As shown below, the results indicate that they are indeed good estimates. Even when I started the minimization routine using different initial guesses, I could not obtain better results for the criterion function.

2.5.2 Results for the Semiparametric Model

In this section I present results only for medium-sized and large farms. I do not present results for small landholders because, as shown in the main text, they do not seem to respond significantly to transportation costs.

I focus on the specification with $k_m = 3$ because I obtained very similar results by setting $k_m = 4$. Figure 2 compares the estimates of the approximating function for G_{su} using $sANN\,(3)$ and the logistic function for the median. The left panel presents the results for small-to-medium farms (5–50 ha); the middle panel reports estimates for medium-to-large farms (50–500 ha); and the right panel, for large farms (more than 500 ha). To make the comparisons fair, I rescaled the logistic function from the IVQR estimates to make the coefficient on TC_m equal -1 and the constant term in the single-index equal zero. The $sANN\,(k_m)$ approximations for G_{su} are surprisingly similar to the rescaled logistic functions (and similar results hold for almost all the quantiles).[9]



(a) Small-Medium Farms

(b) Medium-Large Farms

(c) Large Farms

Figure 2: G(Y,u) and Log(Y/1-Y) at the Median

[9] Note that, in Figure 2, neither the logistic nor the G_{su} functions seem to tend to minus and plus infinity at the extremes: that is because of the scale of the figure and because the share of deforestation in the data is truncated away from 0 and 1. Although not presented in the text, I have also approximated G_{su} using P-Splines and Hermite polynomials, instead of $sANN\,(k_m)$. The estimated functions based on P-Splines and Hermite polynomials are close to the rescaled logistic function, but only when Y is around 0.5. The global approximations in these cases are too flat and fail to capture any curvature of G_{su} when Y is close to either 0 or 1. As a result, P-Splines or Hermite approximations might predict the share of agricultural land outside the unit interval. Despite the fact that the instruments are strong in a linear sense and able to identify β, they may be weak in a "non-linear" sense, i.e., may fail to identify curvatures of G_{su} in a finite data set. Using the $sANN\,(k_m)$ approximation helps impose more restrictions on G_{su} in the data.

Not only is the link function, G_{su}, similar for the SPQIV and the IVQR estimates, but the estimated finite dimensional coefficients, β_{su}, are also similar. Table 10 compares the estimated coefficients using the PSMD and the normalized IVQR estimates.[10] I illustrate only the results for the median regressions, but this pattern is similar for other quantiles. It is clear how (surprisingly) close the estimates are to each other. Although I do not run any formal tests between the IVQR and SPQIV estimates, the results illustrated in Figure 2 and Table 10 suggest that the logit assumption is well suited for this data set.

2.6 Local Yields

In this section, I discuss the use of local yields to calculate the demand for deforestation. As mentioned in the main text, because the data are aggregated up to the municipality level, and because there are hundreds of products being produced in the Amazon, some care is needed in defining $q_m(s)$. I selected the most representative products in the Amazon to construct two productivity indices. The details of how I calculated the indices are explained in Section 3.3 of this supplemental document. Here, I briefly (a) discuss the underlying economic model associated with them, (b) present the corresponding estimated demands for deforestation, and (c) provide evidence that local yields do not respond to changes in transportation costs.

I consider two productivity indices. The first one, denoted by $q_m^{cp}(s)$, is based on the production of beef and the yields of the most representative crops (those discussed in Section 3 of the main paper); it corresponds to the main specification presented in the paper. The second index, denoted by $q_m^{c}(s)$, includes only the main crops and ignores pasture land.

Multinomial Choice Models. Different productivity indices are associated with different underlying multinomial choice models. The first index, $q_m^{cp}(s)$, is associated with a simple "fixed-proportions" model. The underlying assumption in this case is that once the land is cleared for agriculture, it is used in fixed proportions for pasture and for the main crops: the proportions are allowed to differ for different municipalities, but they are fixed within the municipality. If the importance of substitution patterns within municipalities is not central to the present problem, the "fixed-proportions" model should provide a suitable estimate of the demand for deforestation.

The second index, $q_m^{c}(s)$, is associated with a "vertical" multinomial choice model: once the land is cleared for agriculture, it is used only for the representative crops. In this case, farmers do not substitute forest for pasture. The "vertical" model severely restricts the substitution patterns

[10] The t-statistics for the normalized IVQR estimates were computed using the delta method.

among land uses and provides a conservative upper bound on the demand for deforestation. To see why, note first that the counterfactual share of agricultural land $Y_m(s,t)$ is increasing in $q_m(s)$ (see Section 4.1 in the main paper). Because the production of beef is land-intensive, $q_m^{cp}(s)$ must be smaller than $q_m^{c}(s)$ (as it is in the data). Therefore, the demand curve based on the "vertical" model should be above the demand curve based on the "fixed-proportions" model, and also should be above any demand curve that allows for rich substitution patterns between forest and pasture.

In general, rich substitution patterns can be recovered by exploiting choice-specific variables that shift the value of each land use independently of the value of the other options (Berry and Haile (2014)). However, there is no variable satisfying such a requirement in the present data set. Note that although it is possible to estimate a parametric multinomial choice model in the absence of the choice-specific variables, the estimated parametric model may be inconsistent with the underlying model associated with the productivity indices; and the indices are necessary to recover the demand for deforestation. For this reason, I choose to be agnostic in terms of the way the agricultural area is divided when estimating the impacts of TC_m on deforestation, and experimented with the two different indices associated with the different multinomial models.

Demand for Deforestation. Figure 3 compares the demand for deforestation based on the "fixed-proportions" model and the demand based on the "vertical" model. The demand based on the "vertical" model indeed lies above the curve based on the "fixed-proportions" model, implying smaller impacts from taxes. As mentioned previously, the "vertical" demand curve provides an upper-bound for any demand curve with flexible substitution patterns among the different agricultural land uses. It results in a very conservative upper bound because the assumption that all forests can only be converted to cropland (which in turn can then be converted to pasture) is too restrictive for the Brazilian Amazon.[11]

Yields Regression. I now investigate whether the local yields respond to transportation costs. Specifically, I estimate the model

$$q_m(s) = X_m \gamma_s + \delta_s TC_m + \varepsilon_{ms}^{q},$$

[11] Based on remote-sensing data, Almeida *et al.* (2016) provide evidence that a significant fraction of deforested areas was converted directly to pasture. I do not present the supply curve of avoided emissions for the "vertical" model because it provides an extremely conservative bound that is not very informative for carbon emissions.

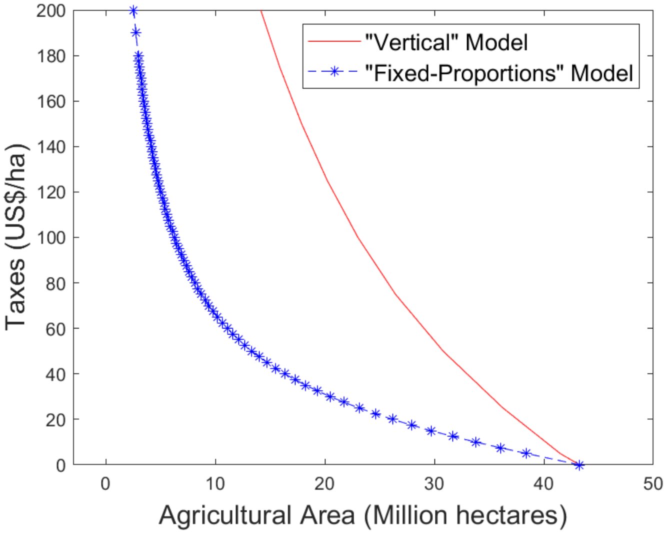


Figure 3: Demand for Deforestation – Different Specifications

for both $q_m^{cp}(s)$ and $q_m^{c}(s)$, using the same set of covariates and instrumental variables as in the land-use regressions. Recall that the crops-pasture index $q_m^{cp}(s)$ corresponds to the main specification in the paper (the "fixed-proportions" model).

Tables 11 and 12 report the estimated coefficients on the transportation costs and omit the coefficients on the other regressors. Table 11 presents results for the crops-pasture index, $q_m^{cp}(s)$. Except for the small farms, all estimated coefficients on costs to ports are insignificant and small in magnitude.

Table 12 presents the results for the crops-index, $q_m^{c}(s)$, as well as the results for each individual crops. The estimated effects of transportation costs on $q_m^{c}(s)$ are statistically significant (except for small to medium-sized farms). However, when separately regressing the yields of each individual product on covariates, I obtained impacts of transportation costs that are always insignificant. This conforms with Roberts and Schlenker's (2013) results: farmers seem to adjust the extensive margin (land use), but not the intensive margin (yields), in response to changes in transportation costs. The significant coefficients observed in Table 12 are therefore explained by the impacts on the weights of the indices, i.e., on shares of land use. This suggests some substitution patterns among land uses. However, as previously discussed, identifying rich substitution patterns in the present context is difficult. I conclude that the response of local yields to transportation costs is not a first-order issue.[12]

[12] The number of observations in the yields regressions are slightly smaller than in the land use regressions because of the presence of some outliers. The yields outliers are due to manioc: few observations report an unreasonably large amount of manioc produced. I consider manioc yields above 10 ton/ha measurement error. Because manioc is

3 Data

This section provides a detailed explanation about how the share of deforestation, the transportation costs, and the productivity indices were constructed. It also provides a brief explanation about the covariates and other useful variables that were not utilized in the regressions, but that were helpful in providing a sense of magnitude for the results.

The richest data set available for the agricultural sector is the Brazilian Agricultural Census, produced by the IBGE (Brazilian Institute of Geography and Statistics).[13] The unit of analysis in the census is the "agricultural establishment," be it a household or firm producing any animal or plant output. The data are aggregated up to the municipal level. There are 523 municipalities in the data set. All municipalities with a positive fraction of their area in the Amazon Biome and with complete information about all covariates are included in the sample.

3.1 Dependent Variable: Deforestation

The land use in the Agricultural Census is divided into the following categories: annual cropland, perennial cropland, pasture (planted and natural), forest (planted and natural), ponds and lakes, constructions, degraded land, and unusable land (for economic activities). I define agricultural land as the sum of annual and perennial cropland, pasture, constructions and degraded land. Forested area is the sum of the remaining land uses, which includes managed forests and forests that are not currently being exploited.

The online information available for the Agricultural Census provides land use data separately for 18 classes of farm size. For confidentiality reasons, the land use data are missing whenever the municipality has only one or two farmers in a given class. Yet, by aggregating different classes of farm sizes so that the final aggregated class has more than two farmers in the municipality, it is possible to recover the total land-use information. I therefore aggregated the 18 classes into four in trying to achieve a good balance among: (a) not losing too much information from the missing data, (b) obtaining more or less homogeneous classes, and (c) easily exposing the results. The final group classification is: (i) small farms (those with fewer than 5 hectares); (ii) small-to-medium farms (those with an area between 5 and 50 hectares); (iii) medium-to-large farms (those with an area between 50 and 500 hectares); and (iv) large farms (those with more than 500 hectares). For each group of farm size, the share of deforestation in a municipality divides the total agricultural

typically produced for subsistence, it is conceivable that is more likely subject to measurement error problems.

[13] Available at http://www.ibge.gov.br/home/estatistica/economia/agropecuaria/censoagro/default.shtm and http://www.sidra.ibge.gov.br/bda/agric

area of the farms by the total area they occupy.

3.1.1 A Comparison of Census Data and Satellite Data

Because the existing policies restrict the amount of deforestation on private lands, one may wonder whether farmers underreport their share of agricultural land in the census. To investigate whether this is an important issue, I contacted the IBGE staff for further clarifications about the census methodology; I also contacted staff at the Brazilian National Institute of Space Research (INPE) who are responsible for the official measurement of deforestation based on satellite imagery; and I compared the census and the satellite deforestation data myself.

Agricultural Census Data. In personal communication with the IBGE staff, they told me that, although they do not select a subset of farmers to serve as validation data for the Agricultural Census, they do verify the information by comparing the census with several other sources of data that are collected regularly (the oldest survey started in 1935, and the most recent, in 1972). The other data sets are the Systematic Survey of Agricultural Production, the Municipal Agricultural Production Survey, the Extraction of Forest Products and Silviculture Survey, the Municipal Livestock Production Survey, and the Quarterly Survey of Livestock Slaughter.[14] In all surveys, IBGE verifies the information by checking first for filling errors (such as fields left blank or invalid values) and calculation errors (such as planted area smaller than harvested area). Based on central tendency measures (mean, median, mode), IBGE constructs acceptable ranges to correct for outliers. It also checks variables' percentage variation over time to detect extremely discrepant differences. Information on weather conditions and any records of abnormalities that may have affected the normal crop evolution are taken into account (IBGE (2002)). The staff argue that the local IBGE agencies have interacted with farmers systematically and continuously and so accumulated substantial local knowledge (there exist more than 500 IBGE agencies in the Brazilian territory, and

[14] The *Systematic Survey of Agricultural Production* is a monthly survey that forecasts and monitors the performance of agricultural crops for the current year with the objective of helping public policies. It provides planted and harvested area estimates, production, and average yield at the municipal level, and this informattion is available to the public only aggregated at the state level. The *Municipal Agricultural Production Survey* collects detailed and disaggregated annual information on crop area planted and harvested, quantity, and production value for several crops at the municipality level. The *Extraction of Forest Products and Silviculture Survey* provides annual municipal-level information on the quantity and value of production obtained by the exploitation of natural forest resources (timber, medicines, vegetal oil, fruits, etc.), as well as the exploitation of planted forest regions. The *Municipal Livestock Production Survey* collects annual municipal-level information on the number of cattle, pigs, chickens and other animals, as well as animal products (such as milk and wool). Finally, the *Quarterly Survey of Livestock Slaughter* provides quarterly data on the number of animals slaughtered and the total carcass weight per animal species investigated. In all cases, the information is obtained from local farmers, and complemented by cooperatives, agribusinesses, companies that provide technical assistance, slaughterhouses, etc. (IBGE (2002)).

each agency is responsible for collecting data covering approximately 11 municipalities on average). They also argue that farmers have little incentive to misreport because (a) they are informed that the census is intended solely for statistical purposes and cannot legally be used against farmers; and (b) farmers are themselves interested in having access to accurate IBGE statistics to make well-informed planting decisions.[15]

Although IBGE's answers are encouraging, I investigate further whether misreporting is a first-order problem by comparing the deforestation measurements based on satellite and census data. Before presenting the comparison, a brief explanation about how INPE estimates deforestation using remote-sensing data is in order (INPE (2017)).

Satellite Sensor Data. INPE uses the Landsat Thematic Mapper images (TM/Landsat), with spatial resolution of 60×60 meters, to calculate the following land covers in the Brazilian Amazon: forest, deforested, non-forest (mostly cerrado, which is similar to savanna), hydrography, clouds, and unobserved. Land cover classification is performed in several steps. First, INPE estimates a linear spectral mixture model for each pixel in the data to obtain the pixel's fraction of different components that help predict land cover.[16] It then groups adjacent pixels in larger regions based on their spectral similarities. After the image segmentation, it implements a cluster unsupervised classification algorithm to generate the land cover classifications. Finally, it verifies the classifications and calculates the deforestation rate.[17]

There are several differences between the deforestation measured by the census and that mea-

[15] IBGE staff also told me that the Agricultural Census follows the United Nations' recommendations established by the Food and Agriculture Organization (FAO) contained in the "Programa del Censo Agropecuario Mundial 2010" (available at http://www.fao.org/economic/ess/ess-wca/es/).

[16] The pixel components they consider are soil, vegetation, and shade. The images-fraction shade and soil help in the process of identification of deforested areas; images-fraction shade is helpful for areas dominated by tropical forests due to the various strata in the forest structure and the irregularity of the canopy, which contrasts with a low amount of shade in deforested areas; and finally the images-fraction soil helps detect transition/contact areas between forest formations and those of cerrado. Formally, the linear spectral mixture model is the following: for each pixel, let $i = 1, ..., I$ be the number of spectral bands (e.g., blue, red, near-infrared, etc.); and $j = 1, ..., J$ be the number of components assumed for the problem (soil, shadow, and vegetation). Assume the linear model $R_i = a_1 X_{i1} + ... + a_J X_{iJ} + e_i$, where R_i is the mean spectral reflectance for the $i - th$ spectral band (e.g., how much the pixel reflects red in the sensor data); X_{ij} is the spectral reflectance of the $j - th$ component in the pixel for the $i - th$ spectral band (e.g., how much soil reflects red – which is assumed known, obtained from other studies); a_j is the proportion of the $j - th$ component in the pixel; this is the parameter of interest. Assuming all a_j are positive and add up to 1, a constrained least squares is used to estimate the coefficients a_j (Camara, Valeriano, and Soares (2006)).

[17] Several photointerpreters have the task of analyzing the thematic polygons generated by the classifications (directly on the computer screen having as background, for comparability, the original image in color composition). The classified polygons are accepted or reclassified into other land use categories based on the experience of the photointerpreter, whose evaluation depends on specific contexts and, if necessary, relies on historical data (Camara, Valeriano, and Soares (2006)).

sured by the interpretation of satellite imagery. First, while the census data capture land use within private properties, satellite data cannot distinguish between deforestation on private and public land. Farms are not geo-referenced in the census data, which makes it impossible to obtain deforestation on private land using satellite images. As explained in the main text, this is the reason why I opted for the census data in the econometric analysis.

Second, the census data include the short-term fallow (land not used for four years) in either the crop land or the pasture land, depending on the farmer's activity; while the long-term fallow (land not used for more than four years) is included in the forested land. INPE, in contrast, considers deforestation irreversible, so that land converted back to forest is counted as deforested. In this sense, the satellite measure underestimates the amount of forest. To have a better sense of whether the underestimation is substantial, I make use of the TerraClass data, which consider a finer classification of land uses (including forest regrowth) and which have been released recently by INPE.[18] According to this data, forest regrowth is responsible for about 22 percent of the total accumulated deforested area. Although highly informative, TerraClass cannot be used directly to compare deforested versus forested areas and the census data for two reasons: first, TerraClass data cover every two years from 2004 to 2014, except for 2006 (exactly the year of the agricultural census). Second, forest regrowth classification is not based on biologic methods, and so it includes various stages of development, from initial stages, when the canopy is homogeneous and has few species, to advanced stages, when canopy heterogeneity and species diversity are similar to original forest (Almeida *et al.* (2016)).[19]

In addition to forest regrowth, deforestation measured by INPE includes more categories than the agricultural land in the census. It includes urban area, mining, and others (such as rock outcrops, river beaches, and sandbars). However, these land covers are responsible for only about 2 percent of the total accumulated deforested land, according to the TerraClass data.

Finally, while the census data may suffer from misreporting, satellite data may suffer from prediction errors. There are three reasons for prediction errors. First, cloud cover prevents the classification of true land use; this is an important issue for the Amazon rainforest. Second, it is common to include a residual classification to allow for cases in which the classifier algorithm is not able to clearly identify the land cover. INPE includes such a class, denoted by "unobserved area,"

[18] TerraClass data are available at http://www.inpe.br/cra/projetos_pesquisas/dados_terraclass.php

[19] The spectral similarities in the reflectance of forests at different stages of succession and regeneration lead to high classification errors when one tries to distinguish secondary forests from deforested areas, cerrado vegetation, and mature forests. Assessments of forest regrowth based on remote-sensing data are commonly viewed as a nontrivial difficult task (Caviglia-Harris *et al.* (2015)).

that essentially results from poor radiometric quality of the sensor data. The average proportion of the sum of clouds and "unobserved area" per municipality in my data is 10 percent, which is a non-negligible share of municipality areas.[20] Finally, there are the unavoidable statistical misclassification errors. Unfortunately, INPE does not report the land cover misclassification probabilities, so I cannot assess the accuracy of their deforestation measurement.

Direct Comparison. Although the two deforestation measures are not entirely comparable, in order to make a comparison that is as clean as possible, I proceed as follows: first, I calculate the municipalities' share of deforestation according to both census and satellite data sets. Then I compare the two measures in municipalities in which both the share of public land and the share of clouds/unobserved areas are small. Specifically, I consider municipalities with more than 90 percent of private land and with less than 3 percent of clouds/unobserved areas.[21] The idea is that in municipalities that have negligible share of public land and that are not contaminated by clouds or by "unobserved areas," the two measures should be close to each other.[22]

As expected, the correlation between the two deforestation measurements increases as we move from the entire sample to the restricted sample: the correlation increases from 0.74 to 0.84. The average shares of deforestation also increase and get closer to each other when I use the restricted sample, again as expected. The average share of deforestation in the full sample according to the census is 27 percent, while according to the satellite data it is 42 percent (which is consistent with the previous discussion). In the restricted sample, the census average share is 68 percent, and the satellite average share is 75 percent. Although the gap is reduced by half, there is still a non-negligible difference. Yet, when I subtract the 2008 TerraClass data's forest regrowth from the INPE deforestation, I obtain an average deforestation of 65 percent, which is much closer to the census average. Although TerraClass data are not directly comparable to INPE or census data (as I mentioned earlier), these results are encouraging.

Perhaps more important than comparing raw correlations and simple averages is to investigate whether the census deforestation systematically varies with exogenous variables (particularly with the instrumental variables) *after* controlling for the satellite deforestation. The idea is that, if we consider the satellite data to provide a good measure of deforestation, then, after conditioning on it,

[20] Clouds result in a missing data problem. However, the data are not exactly missing-at-random because some areas are covered by clouds systematically more often than others. The data can be more convincingly considered missing-at-random only if *conditioned* on the latitude and longitude of a location.

[21] To minimize potential problems with INPE misclassification probabilities, the restricted sample also ignores municipalities in which the sum of land areas classified in INPE is greater than or equal to the municipality area. Correlations are not very sensitive to this restriction.

[22] I am grateful to an anonymous referee who suggested such comparison.

the exogenous variables should contain no additional information about deforestation in the census data.

To investigate this possibility, I regress the census deforestation measure on satellite deforestation and other explanatory variables. In this case, I augment the restricted sample to include municipalities with more than 80 percent of private land (because the sample with more than 90 percent of private land does not have enough degrees of freedom). Table 13 presents the results. The first column presents the estimated coefficients for the full sample, and the second column, for the restricted sample.[23]

As expected, the estimated coefficients on the satellite deforestation measure are positive and statistically significant in both samples. The estimated coefficients on most of the other explanatory variables are not statistically significant in the full sample, including the coefficients on the instrumental variables (distance to port and distance to capital), which are very small in magnitude. In the restricted sample, when the census and satellite measures are most comparable, none of the regressors' coefficients are significant except for the coefficient on satellite deforestation.[24]

The correlations I find are consistent with recent results obtained by Caviglia-Harris *et al.* (2015). To the best of my knowledge, theirs is the only published paper that combines and compares survey data (with geo-referenced household property boundaries) and remote-sensing data for the Brazilian Amazon. They perform a series of reliability tests, and find a high degree of consistency between households' responses and satellite-derived land cover.

3.2 Endogenous Regressor: Transportation Costs

The proxy for transportation costs is defined as the minimum unit cost (US$/ton) to transport 1 ton of goods to the nearest port. This cost was calculated by combining information from the Brazilian transportation network for 2006 produced by the Ministry of Transportation for the National Highway Plan (PNLT (2006)), and the freight rate data collected by SIFRECA (the Freight Information System). The implementation used ArcGIS cost distance tools – see Allen and Arkolakis (2014) for an excellent description of this type of algorithm. Calculating this proxy

[23] The regressions include the proportion of clouds/unobserved areas among the covariates to further mitigate measurement problems in the satellite deforestation. Note that the coefficient on clouds/unobserved areas increases substantially in the restricted sample; this is because the scale of that covariate changes from an average of 10 percent in the full sample to an average of 0.0004634 percent in the restricted sample. For both samples, I used the Eicker-Huber-White robust standard errors. I obtained qualitatively similar results when I used the spatially dependent HAC robust standard error proposed by Conley (1999). For the spatially dependent robust standard errors, I used a uniform spatial kernel, and a 75 km distance cutoff. The results do not change when cutoffs are 50 km or 100 km.

[24] Results are qualitatively similar when I subtract the TerraClass 2008 forest regrowth data from satellite deforestation.

requires definitions about (a) which ports are included in the calculations, and (b) how to combine the freight rate data with the network of roads, railroads, and navigable rivers.

As discussed in the main paper, when directed to international markets, the main products in the Amazon normally use either the ports in the Amazon (Port of Santana, Port of Belém and Port of Itaqui) or the ports in the South (Port of Santos and Port of Paranaguá); see Figure 2 in the main paper. Therefore, I selected these five ports to be the main destinations.

Given the selected destinations, the least cost path to the nearest port is computed in ArcGIS. The calculation divides the entire country into cells corresponding to 1 km^2, and the cost to travel over each cell in the grid depends on whether it contains a segment of road (paved or unpaved), railroad, navigable river, or no transportation mode. Given these travel costs, the optimization routine in ArcGIS determines the least cumulative cost path from each origin to the nearest destination.[25] To assign the travel costs for each transportation mode, I use the freight rate data collected by SIFRECA. I adjust for road quality in the calculations using the Vehicle Cost Module of the World Bank's Highway Design Model (HDM-VOC-4), which is designed to calculate unit road user costs for different types of road section (see below). Because almost all the information I obtained from SIFRECA for the Amazon corresponds to costs of transporting soybeans, the proxy TC_m measures the minimum cost of transporting 1 ton of soybeans to the nearest port. Note that, although the costs to transport different products may differ, they should be highly collinear: all products use the same transportation network and reach the same ports (under the no-arbitrage condition). Furthermore, bulk products and sacks must have the same transportation costs as soybeans (Castro (2003)).[26]

Table 14 summarizes the unit costs used to compute the least accumulative cost in ArcGIS. The first column discriminates between the possible modes of transportation considered in the calculations; the second column reports the unit travel costs used in Brazilian currency (R$); and the third column converts these unit costs into US$. Note that differences in unit costs are in the expected direction, as waterways with good infrastructure are the cheapest mode of transportation, followed by railroads and paved roads. Unpaved roads inside the Amazon are the worst mode of transportation followed by navigable rivers with poor infrastructure and unpaved roads outside the

[25] The network of modes of transport are in a polyline format in ArcGIS while the country is in a polygon format. I first have to transform these polylines and polygons into a raster format (i.e., a grid of cells with cost values). Then I use the command "cost distance" in ArcGIS to compute the least cumulative cost path to the nearest destination and "cost allocation" to identify which destination is the nearest for each cell. Finally, the command "extract values to points" is used to assign the total costs to the corresponding municipal seats.

[26] I am grateful to Prof. Newton de Castro, who suggested the use of the World Bank's HDM-VOC-4 model and the data from SIFRECA to compute transportation costs.

Amazon. Finally, because ArcGIS allows for traveling by land, I imposed high costs to transport goods by land with no mode of transportation so that ArcGIS would avoid computing traveling costs using these cells.

Next, I discuss the unit costs for each mode of transportation.

Roads. I purchased the freight values of routes from cities in the Legal Amazon to state capitals and ports for the main products in the Amazon.[27] I obtained the costs and distances to transport soybeans by roads for 105 routes and five destinations for 2006. The average cost to transport 1 ton of soybeans per km in these data is R$ 0.0767 (US$ 0.0353). That is the value I inputted as the unit cost to travel one cell with a paved road.[28]

To capture the higher costs of traveling on unpaved roads, I increased the unit costs for paved roads by using the World Bank's HDM-VOC-4. This model is designed to calculate unit road user costs for a road section with 1 km length and requires several inputs for the characteristics of the road. I maintained all inputs at the default values, except for changing the road characteristics from paved to unpaved and increasing the roughness of the road to the recommended value for a poor tertiary road. For heavy truck vehicles, the increase in the roughness raised the road user costs by 29 percent. I adopted the unit cost to travel 1 km on unpaved road to be 29 percent higher than the cost to travel on a paved road. Therefore, a cost of R$ 0.0992 (US$ 0.0457) to transport 1 ton of soybeans per km was assigned to unpaved roads.[29]

Transporting large quantities on unpaved roads within the rainforest is extremely difficult. The poor conditions of the roads, combined with the excess of rain, especially in the rainy season, can make these roads inaccessible. For this reason, I decided to differentiate unpaved roads within the rainforest from the unpaved roads elsewhere. Unfortunately, the freight rate data purchased from SIFRECA do not cover the dense rainforest. To overcome this limitation, I called local companies directly and asked them how much it costs to transport soybeans from Sorriso (in Mato Grosso state), one of the main producers of soybeans, to the Port of Santarém on the Amazon River. The

[27] The Legal Amazon is an administrative area in the northern part of Brazil that includes nine states and around 5 million km^2 of land (about 60% of the Brazilian national territory). The states are indicated in Figure 1 in the main paper. It consisted of 771 municipalities in 2006, and includes the Amazon Biome as well as other types of vegetation, in particular a savannah type of vegetation called the *cerrado*.

[28] In SIFRECA's sample, almost all cities were located in the state of Mato Grosso, the main producer of soybeans.

[29] The International Roughness Index (IRI) is an index that measures the deviations of a surface from a true planar surface with characteristic dimensions that affect vehicle dynamics, ride quality, dynamic loads, and drainage. It is measured in m/km units. The value recommended for a good primary paved road is 2 m/km. The value I used in order to increase the costs of unpaved roads is 8 m/km, which corresponds to a poor tertiary road. See Archondo-Callao (2008). The HDM-VOC-4 model is available at
http://www.lpcb.org/index.php/edocman-test/software/hdm4-road-user-cost-model-v-3-0

only route available in this case is the "Cuiabá-Santarém" road, which cuts the Amazon almost in the middle from the South to the North and provides a good measure of the difficulties in traveling in the dense jungle. The average cost per km that I obtained is R$ 0.15 (US$ 0.069). I used this unit cost for all unpaved roads within the dense rainforest.[30]

Railroads and Navigable Rivers. Freight values for navigable rivers and railroads are more difficult to obtain because of the reluctance of the firms to disclose their data. The information I was able to obtain from SIFRECA includes the freight values to transport 1 ton of soybeans in two routes for a railroad, with an average value of R$ 0.0608 (US$ 0.0279), and one route for one of the most important navigable rivers in the Amazon – the Madeira River waterway. Because of the difficulty in collecting these data, SIFRECA does not provide recent freight rate data for these modes of transportation anymore.[31]

Similar to roads, there are differences in the quality of navigable rivers, depending on the depth of the river, investments in signaling, investments in communications, and in the quality of the local ports. Based on conversations with governmental agencies responsible for the administration of the waterways, as well as local companies, I classified the navigable rivers in two types: those with good and those with poor infrastructure. The good rivers include the Madeira River waterway and the Amazon River waterway (linking Manaus to Belém), among a few others. The rivers with poor infrastructure are the remaining navigable rivers.[32]

From SIFRECA's information of the freight cost of the Madeira River waterway combined with the information I obtained directly from the companies for the cost of the Amazon River waterway (from Manaus to Belém), I arrived at a unit cost of R$ 0.0444 (US$ 0.0204) to transport 1 ton of soybeans per km. For the navigable rivers with poor infrastructure, I obtained an average value of R$ 0.1139 (US$ 0.0525) per ton of soybeans per km.[33]

[30] I define the dense rainforest as covering the states of Acre, Amazonas, Rondônia, Roraima, Pará, and Amapá. The remaining states in the Legal Amazon (Maranhão, Tocantins, and Mato Grosso) are not in the dense rainforest.

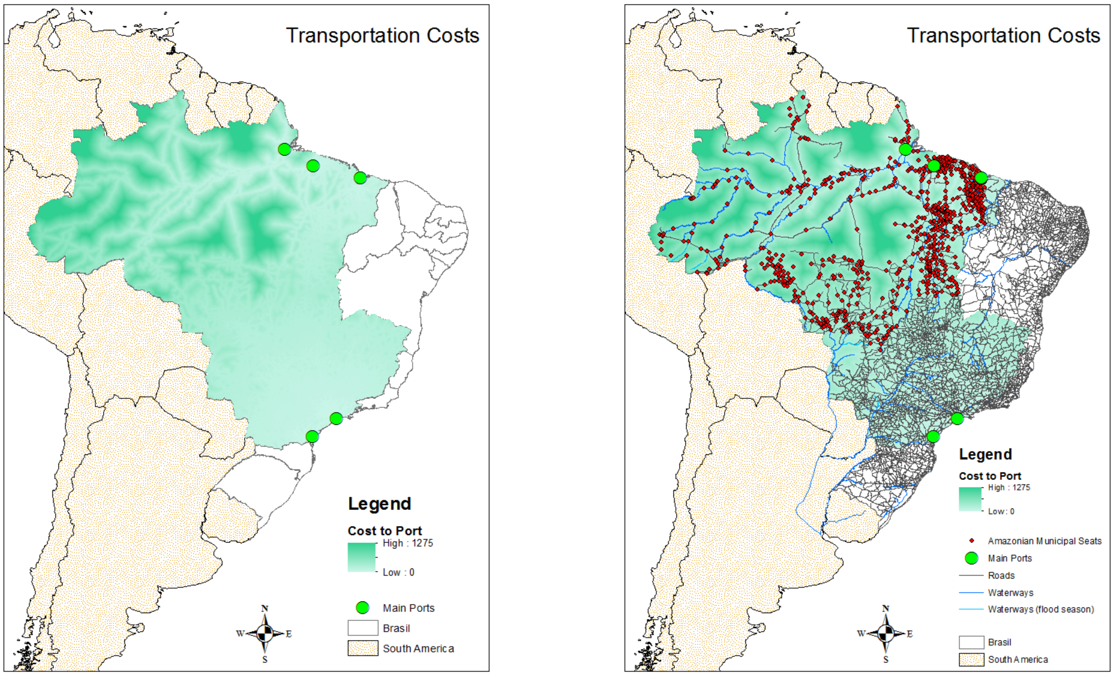
[31] The routes for railroads are from Cascavel to Ponta Grossa (both in the Paraná state in the South) and from Porto Franco to São Luís (both in the Maranhão state in the Legal Amazon). The Madeira River waterway connects Porto Velho (the capital of the Rondônia state) to Itacoatiara, which is close to Manaus (the capital of the Amazonas state).

[32] The Ministry of Transportation classifies two types of waterway: *more-navigable* and *less-navigable* rivers. I excluded the *less-navigable* rivers, because they do not seem to be used to transport large quantities of products. Hence, the classification of good- and poor-infrastructure rivers is restricted only to those *more-navigable* rivers.

[33] The unit cost for poor-infrastructure rivers is obtained from the routes Manaus–Tabatinga, Manaus–Barcelos and Manaus–Boca do Acre, all of them in pristine regions in the Western Amazon. The larger costs obtained reflect not only poor signaling and capacity constraints, but also difficulties with the excess of curves, the depth of rivers (some may have about 1 meter depth in the dry season, while the Amazon River has, on average, 16 meters' depth), as well as the presence of stones and sandbars that shift around over time. To gain a sense of magnitude, it can take about 10 days to go from Tabatinga (located at the border with Venezuela) to Manaus and about 18 days to navigate

Land. Finally, I imposed high costs to transport products by land with no mode of transportation so that ArcGIS avoids computing traveling costs using these cells. For the Amazon Biome, I imposed a cost of R$ 3 per km (US$ 1.38) and for land outside the Amazon Biome, a cost of R$ 1.5 (US$ 0.69). The rationale is that moving within the rainforest should be much more costly than for other types of vegetation. There is little guidance on which values should be adopted for transportation in land, but, because the municipal seats are connected to some segment of the network and to the extent that these unit cost values induce ArcGIS to use cells with some mode of transportation, the costs of moving on land should not impact the results significantly.[34]

Figure 4 presents the map of transportation costs calculated in ArcGIS. The figure also presents the network of modes of transportation and the municipal seats in the Legal Amazon in the right panel. The darker the color in the map, the higher the transportation cost.



(a) Costs to Port

(b) Transportation Network

Figure 4: Transportation Costs and Transportation Network

that stretch of the river in the opposite direction.

[34] For those cities not connected to any mode of network in the map, I created straight lines joining them to the nearest road and assigned the cost of unpaved road (depending on whether they are inside or outside the rainforest) to travel over these straight lines. This procedure is reasonable, because the official map from the National Highway Plan misses the unofficial roads, and they are most likely unpaved.

3.3 Productivity Index

From the Agricultural Census, I obtained the quantity sold and the area occupied of major agricultural outputs for each municipality for farms of different sizes. I calculated two productivity indices: one considers only the production of crops, while the other considers the production of both crops and beef.

I describe first the index based on crops only. I selected the main crops discussed in the paper: soy, corn, manioc, rice, and beans. For each product j, for each municipality m, and for each farm size s, I calculated the quantity of output sold per hectare for each product. The yields are denoted by q_{jm}^{s}, for $j = 1, ..., 5$. The productivity index is the weighted average of the q_{jm}^{s} across j, where the weights are the proportion of the area utilized for each crop, denoted by a_{jm}^{s}. The index for the crops is therefore:

$$q_{m}^{c}\left(s\right) = \sum_{j=1}^{5} \left(\frac{a_{jm}^{s}}{a_{m}^{s}}\right) \times \left(q_{jm}^{s}\right)$$

where $a_{m}^{s} =_{j=1}^{5} a_{jm}^{s}$.

To add pasture in the productivity index, I first assumed that each ox weighs half a ton and that the entire ox can be used for beef consumption. I therefore multiplied the number of cattle sold by 0.5 to obtain the quantity of beef sold in tons.[35] By dividing this result by the pasture area, I obtained the productivity for cattle, q_{m}^{p}. I do not have information on the number of oxen for different farm sizes, so q_{m}^{p} is the same for all farm sizes.

According to the freight data, it is 30 percent more expensive to transport frozen meat than soybeans. For this reason, I increased the weight for the beef in the productivity index by 30 percent so that the transportation costs of all products are measured in terms of the costs to transport 1 ton of soybeans. Let a_{mp}^{s} be the area occupied by pasture for farms of size s in location m. The second productivity index is a weighted average between q_{m}^{c} and q_{m}^{p}, where the weights are the proportion of crops and pasture areas:

$$q_{m}^{cp}\left(s\right) = \left(\frac{a_{m}^{s}}{a_{m}^{s} + a_{mp}^{s}}\right) q_{m}^{c}\left(s\right) + \left(\frac{a_{mp}^{s}}{a_{m}^{s} + a_{mp}^{s}}\right) \left(1.3\right) q_{m}^{p}$$

I averaged the indices over the micro-regions to reduce measurement error problems. Micro-regions are administrative areas larger than the municipalities; there are 85 micro-regions in the Amazon.

[35] Informal conversations with farmers suggest that a common rule of thumb is to assume a cow has about 150 kg of "available meat" and a bull has about 270 kg. Because a bull weighs around 500 kg, I am overestimating the productivity for pastures.

3.4 Covariates

Next, I briefly describe the set of covariates and the instrumental variables.[36]

Temperature and Precipitation. The Climate Research Unit (CRU) computed the annual average temperature and precipitation based on the average climate from 1961 to 1990. Given the high levels of temperature and rain in the region, one might expect these variables to be negatively correlated with deforestation.[37]

Altitude. This variable is prepared by the IBGE. Although high altitudes may be good for the production of rice, I do not have *a priori* information about whether this variable should be positively or negatively correlated with deforestation.[38]

Slope. To calculate the average slope in a municipality, I used data from the Shuttle Radar Topographic Mission (SRTM). Slope is the inclination of a surface in relation to the horizontal surface, and it is measured in degrees. One should expect places with steeper slopes to be worse for agriculture in the Amazon because of water runoff and because mechanization is more difficult.[39]

Soil quality. These data are produced by IBGE and EMBRAPA (the Brazilian Agricultural Research Corporation) and were kindly made available by Professor Eustáquio Reis. They consist of the proportion of the municipal area on each of five aptitude classes of soil. The classes of soil aptitude for agriculture are: high quality, medium-to-high, low-to-medium, low quality, and unsuitable. The five aptitude classes were aggregated from the 13 soil types that exist in the Brazilian territory. Using the 13 soil types in conjunction with the data on local topography, data from ground surveys, and general familiarity with the land, EMBRAPA soil scientists created the digital map of soil aptitude for agriculture that is used here. Factors that could lead to a low ranking include high metal content, poor drainage, high flood risk, uneven ground, low nutrients, and steep slope (Anderson and Reis (2007)).

[36] Some covariates were available at the municipal level for years other than 2006. Because municipal boundaries changed over time, I had to match the previous boundaries to convert the values of these covariates to the boundaries observed in 2006. I summed or averaged the values over the municipalities in 2006, taking a weighted average where appropriate.

[37] The data are available at http://www.ipeadata.gov.br/

[38] The data are available at http://www.ipeadata.gov.br/

[39] The data are available at http://www.dpi.inpe.br/Ambdata/English/index.php

Local Population. These data come from the Demographic Census of 2000 produced by IBGE. In principle, the local population may affect local demand for non-tradables and local wages. Land use on private land data, however, is affected by the local population depending upon whether agriculture is more or less labor-intensive than extractive activities.[40]

Presence of Local Mining and Power Plants. These data are available from the National Highway Plan. The presence of mining and power plant should shift the demand for non-tradables. To capture the effect that electrification may have on the neighborhood of power plants, I also created a dummy variable that equals one if a power plant exists within a 75 km radius of the municipal seat (but is not within the municipality itself), and zero otherwise.

Distance to Protected Areas. These data are available from the National Highway Plan. I computed straight-line distances from the municipal seats to the nearest protected area polygon using ArcGIS. Protected areas in the Amazon rainforest include indigenous reserves and conservation units (which in turn include national and state forests, wildlife reserves, extractive reserves, ecological stations, areas of relevant ecological interest, and national parks, among others).

Distance to IBAMA. I computed straight-line distances from the municipal seats to the nearest IBAMA agency using ArcGIS.

Number of Fines. The total number of fines includes all environmental infractions in a municipality for each year since 2002. The data are publicly available from IBAMA. For each municipality, I added all fines starting in the first year the data are available up to the year before the Agricultural Census.

Proportion of Private Land with Land Title. These data come from the Agricultural Census and measure the share of the private land that has a land title. The land with no title is either occupied by squatters or covered under one of the official settlement programs, which have not yet given the definitive titles to the farmers.

Instrumental Variables. I computed straight-line distances from the municipal seats to the nearest port and to the nearest capital using ArcGIS. I excluded Palmas, the capital of the Tocantins

[40] Available at http://www.ibge.gov.br/home/estatistica/populacao/default_censo_2000.shtm

state, from the destinations for the straight-line distances because it is a planned city that was built in 1989 with the objective of helping to develop the region.

Carbon Stock. The amount of aboveground carbon stock is calculated by Baccini *et al.* (2012). I combined their raster data of carbon stock with the 2008 TerraClass shapefile of land uses in the Amazon to calculate the average carbon stock in forested and deforested areas in each municipality. There are 18 municipalities with missing data on the carbon stock, almost all of them in the state of Maranhão, in the Eastern Amazon.

Other Useful Variables. Useful variables that are not utilized in the estimation procedure include:

Satellite Data. Satellite-based measures of deforestation in the Brazilian Legal Amazon are calculated by INPE (the Brazilian National Institute of Space Research). They include both the PRODES data (Monitoring Program of the Brazilian Amazon by Satellite), and the TerraClass data.[41]

Labor. The Agricultural Census provides the number of workers in the farms (separating family member workers from hired workers), as well as rural wages.

Revenue and Expenditures. The Agricultural Census also contains information on both the revenues and the expenditures of the agricultural establishments. The revenue is the value from the sale of production and the expenditures include maintenance costs, salaries, rentals of machinery, and other expenses. Revenues and costs are not discriminated for each land use.

References

[1] Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2017). "When Should You Adjust Standard Errors for Clustering?" NBER Working Papers 24003.

[2] Allen, T. and C. Arkolakis (2014). "Trade and the Topography of the Spatial Economy," *Quarterly Journal of Economics*, 129(3), 1085–1140.

[3] Almeida, C. A., A. C. Coutinho, J. C. D. M. Esquerdo, M. Adami, A. Venturieri, C. G. Diniz, N. Dessay, L. Durieux, and A. R. Gomes (2016). "High Spatial Resolution Land Use and Land

[41] The methodology is explained by Câmara, Valeriano, and Soares (2006). The PRODES data are available at http://www.obt.inpe.br/prodesdigital/cadastro.php. And the TerraClass data are available at http://www.inpe.br/cra/projetos_pesquisas/dados_terraclass.php

Cover Mapping of the Brazilian Legal Amazon in 2008 Using Landsat-5/TM and MODIS Data." *Acta Amazonica*, 46(3), 291–302.

[4] Anderson, K. and E. Reis (2007). "The Effects of Climate Change on Brazilian Agricultural Profitability and Land Use: Cross-Sectional Model with Census Data," Final report to WHRC/IPAM for LBA project Global Warming, Land Use, and Land Cover Changes in Brazil.

[5] Archondo-Callao, R. (2008). "Applying the HDM-4 Model to Strategic Planning of Road Works," Transport Paper Series, no. TP-20. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/993961468338479540/Applying-the-HDM-4-model-to-strategic-planning-of-road-works

[6] Assunção, J., M. Lipscomb, A. M. Mobarak, and D. Szerman. (2016). "Agricultural Productivity and Deforestation in Brazil," Mimeo, Yale University.

[7] Baccini, A., S. J. Goetz, W. S. Walker, N. T. Laporte, M. Sun, D. Sulla-Menashe, J. Hackler, P. S. A. Beck, R. Dubayah, M. A. Friedl, S. Samanta, and R. A. Houghton (2012). "Estimated Carbon Dioxide Emissions from Tropical Deforestation Improved by Carbon-Density Maps," *Nature Climate Change*, 2, 182–185.

[8] Berry, S. T. and P. A. Haile (2014). "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82(5), 1749–1797.

[9] Camara, G., D. M. Valeriano, and J. V. Soares (2006). "Metodologia para o Cálculo da Taxa Anual de Desmatamento na Amazônia Legal," São José dos Campos, INPE.

[10] Cameron, A. C. and D. L. Miller (2015) "A Practitioner's Guide to Cluster-Robust Inference," *The Journal of Human Resources*, Spring, vol. 50, no. 2, 317-372.

[11] Castro, N. (2003). "Formação de Preços no Transporte de Carga," *Pesquisa e Planejamento Econômico*, 33(1), 167–189.

[12] Caviglia-Harris, J. L., M. Toomey, D. W. Harris, K. Mullan, A. R. Bell, E. O. Sills, and D. A. Roberts (2015). "Detecting and Interpreting Secondary Forest on an Old Amazonian Frontier," *Journal of Land Use Science*, 10(4), 442–465.

[13] Chen, X. (2007). “Large Sample Sieve Estimation of Semi-nonparametric Models,” Chapter 76 in *Handbook of Econometrics*, Vol. 6B, eds. James J. Heckman and Edward E. Leamer, North-Holland.

[14] Chen, X. and D. Pouzo (2009). “Efficient Estimation of Semiparametric Conditional Moment Models With Possibly Nonsmooth Residuals,” *Journal of Econometrics*, 152, 46–60.

[15] Chen, X. and D. Pouzo (2012). “Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals,” *Econometrica*, 80(1), 277–321.

[16] Chernozhukov,V. and C. Hansen (2008). “Instrumental Variable Quantile Regression: A Robust Inference Approach,” *Journal of Econometrics*, 142(1), 379–398.

[17] Chomitz, K. M. and T. S. Thomas (2003). “Determinants of Land Use in Amazônia: A Fine-scale Spatial Analysis,” *American Journal of Agricultural Economics*, 85(4), 1016–1028.

[18] Conley, T. (1999). “GMM Estimation with Cross-sectional Dependence,” *Journal of Econometrics*. 92, 1–45.

[19] Herrera, L. D. (2015). *Protected Areas' Deforestation Spillovers and Two Critical Underlying Mechanisms: An Empirical Exploration for the Brazilian Amazon*. PhD thesis, University Program in Environmental Policy, Duke University.

[20] IBGE (2002). “Pesquisas Agropecuárias,” *Série Relatórios Metodológicos*, Volume 6.

[21] IBGE (2017). “Divisão Regional do Brasil em Regiões Geográficas Imediatas e Regiões Geográficas Intermediárias 2017,” Available at https://www.ibge.gov.br/apps/regioes_geograficas/

[22] INPE, Instituto Nacional de Pesquisas Espaciais (2017), *Monitoramento da Floresta Amazônica Brasileira por Satélite – Projeto Prodes*. Available at http://www.obt.inpe.br/prodes/

[23] Le Sage, J. P. (2014). “What Regional Scientists Need to Know About Spatial Econometrics,” *The Review of Regional Studies*, 44(1), 13–32.

[24] Lipscomb, M., A. M. Mobarak, and T. Barham. (2013). “Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil,” *American Economic Journal: Applied Economics*, 5 (2): 200-231.

[25] Pfaff, A., J. A. Robalino, D. Herrera, and C. Sandoval (2015). “Protected Areas’ Impacts on Brazilian Amazon Deforestation: Examining Conservation – Development Interactions to Inform Planning,” *PLOS One*, 10(7): e0129460. https://doi.org/10.1371/journal.pone.0129460

[26] PNLT, National Highway Plan (2006). Available at
http://www.transportes.gov.br/index/conteudo/id/36604 and accessed on 11/24/2010.

[27] Robalino, J. A., A. Pfaff (2012). “Contagious Development: Neighbor Interactions in Deforestation,” *Journal of Development Economics*, 97(2), 427–436.

[28] Robalino, J. A., A. Pfaff, and L. Villalobos-Fiatt (2017). “Heterogeneous Local Spillovers from Protected Areas in Costa Rica,” *Journal of the Association of Environmental and Resource Economists*, 4(3), 795–820.

[29] Roberts, M. J. and W. Schlenker (2013). “Identifying Supply and Demand Elasticities of Agricultural Commodities: Implications for the US Ethanol Mandate,” *American Economic Review*, 103(6), 2265–2295.

Table 1: Land-use Median Regression by Farm Size

	Small	Small–Medium	Medium–Large	Large
Cost to Port	0.0045	-0.0053	-0.0079**	-0.0110***
	(0.0060)	(0.0034)	(0.0038)	(0.0038)
Altitude	0.0010	0.0000	-0.0002	-0.0011
	(0.0015)	(0.0006)	(0.0007)	(0.0008)
Temp	-0.0383	-0.4680***	-0.2649	0.0229
	(0.3073)	(0.168)	(0.1719)	(0.2074)
Rain	-0.0038	-0.0094***	-0.0068**	-0.0051*
	(0.0040)	(0.0027)	(0.0027)	(0.0031)
Slope	-0.0846	-0.1015	-0.2076	-0.1965
	(0.3645)	(0.2038)	(0.1989)	(0.2297)
Soil 2	-2.1386***	-1.8197***	-1.5805***	-1.6984***
	(0.7548)	(0.4289)	(0.5163)	(0.6469)
Soil 3	-0.9227	-0.8729*	-0.9222*	-0.9942**
	(1.1002)	(0.4532)	(0.4068)	(0.4360)
Soil 4	-1.4279**	-1.4894***	-1.5151***	-1.6160***
	(0.5929)	(0.3118)	(0.3083)	(0.3597)
Soil 5	-0.0730	-0.8029**	-0.5831*	-0.6423*
	(0.6693)	(0.3362)	(0.3179)	(0.3773)
Dist. to PA	-0.0059	0.0005	0.0028	0.0055*
	(0.0065)	(0.0028)	(0.0025)	(0.0031)
Mining	-0.0136	0.0061	-0.0432**	0.0075
	(0.0432)	(0.0413)	(0.0217)	(0.0272)
Power Plant	-0.5107	-0.2998	-0.1086	-0.0445
	(0.3972)	(0.2802)	(0.3311)	(0.3215)
Power Plant Neighbor	-0.4703	-0.3035	-0.3089	-0.1575
	(0.3526)	(0.2423)	(0.2260)	(0.2490)
Population	-0.0008	-0.0005	0.0002	0.0003
	(0.0008)	(0.0009)	(0.0015)	(0.0007)
Prop. Land Title	-0.3951	0.4597	0.5894	0.6563
	(0.9071)	(0.4506)	(0.7070)	(0.8808)
Fines	-0.0009	-0.0012	-0.0015	-0.0015
	(0.0019)	(0.0012)	(0.0019)	(0.0016)
Dist. to IBAMA	-0.0005	-0.0009	0.0013	0.0005
	(0.0022)	(0.0012)	(0.0011)	(0.0013)
Constant	5.6863	16.317***	10.267**	3.4441
	(8.6866)	(4.7843)	(4.9321)	(5.7670)

Notes: Land-use IVQR regression results for the median. Spatially lagged regressors are omitted from the table. The number of municipalities for small farms is 501; for small–medium is 520; for medium–large, 520; and for large farms, 450.

Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Land-use Median Regression by Farm Size – Local Population

	Small		Small–Medium		Medium–Large		Large	
Cost to Port	0.0045	0.0059	-0.0053	-0.0056*	-0.0079**	-0.0081**	-0.0110***	-0.0109***
	(0.76)	(1.00)	(-1.58)	(-1.71)	(-2.07)	(-2.19)	(-2.91)	(-2.88)
Altitude	0.0010	0.0011	0.0000	0.0002	-0.0002	-0.0003	-0.0011	-0.0011
	(0.67)	(0.75)	(0.74)	(0.29)	(-0.27)	(-0.49)	(-1.48)	(-1.39)
Temp	-0.0383	-0.0414	-0.4680***	-0.4943***	-0.2649	-0.2883*	0.0229	-0.0013
	(-0.12)	(-0.14)	(-2.79)	(-2.92)	(-1.54)	(-1.68)	(0.11)	(-0.01)
Rain	-0.0038	-0.0034	-0.0094***	-0.0095***	-0.0068**	-0.0067**	-0.0051*	-0.0054*
	(-0.97)	(-0.86)	(-3.53)	(-3.58)	(-2.55)	(-2.48)	(-1.65)	(-1.73)
Slope	-0.0846	-0.0188	-0.1015	-0.1023	-0.2076	-0.2026	-0.1965	-0.1949
	(-0.23)	(-0.05)	(-0.50)	(-0.50)	(-1.04)	(-1.02)	(-0.85)	(-0.85)
Soil 2	-2.1386***	-2.1002***	-1.8197***	-1.7818***	-1.5805***	-1.5816***	-1.6984***	-1.7085***
	(-2.83)	(-2.79)	(-4.25)	(-4.10)	(-3.06)	(-3.08)	(-2.62)	(-2.65)
Soil 3	-0.9227	-0.8473	-0.8729*	-0.9393**	-0.9222*	-0.9181**	-0.9942**	-0.9793**
	(-0.84)	(-0.78)	(-1.93)	(-2.10)	(-2.27)	(-2.39)	(-2.28)	(-2.25)
Soil 4	-1.4279**	-1.281**	-1.4894***	-1.4712***	-1.5151***	-1.5439***	-1.6160***	-1.6292***
	(-2.41)	(-2.15)	(-4.78)	(-4.75)	(-4.91)	(-5.23)	(-4.50)	(-4.53)
Soil 5	-0.0730	0.1017	-0.8029**	-0.7753**	-0.5831*	-0.5919**	-0.6423*	-0.6309*
	(-0.11)	(0.15)	(-2.38)	(-2.32)	(-1.84)	(-1.97)	(-1.70)	(-1.68)
Dist. to PA	-0.0059	-0.0063	0.0005	0.0006	0.0028	0.0030	0.0055*	0.0055*
	(-0.90)	(-0.97)	(0.16)	(0.20)	(1.13)	(1.22)	(1.78)	(1.75)
Mining	-0.0136	-0.0133	0.0061	0.0125	-0.0432**	-0.0477**	0.0075	0.0053
	(-0.31)	(-0.31)	(0.15)	(0.29)	(-1.99)	(-2.19)	(0.28)	(0.20)
Power Plant	-0.5107	-0.4962	-0.2998	-0.3123	-0.1086	-0.172	-0.0445	0.0201
	(-1.29)	(-1.24)	(-1.07)	(-1.09)	(-0.33)	(-0.56)	(-0.14)	(0.06)
Power Plant Neighbor	-0.4703	-0.5181	-0.3035	-0.3184	-0.3089	-0.2960	-0.1575	-0.1136
	(-1.33)	(-1.47)	(-1.25)	(-1.34)	(-1.37)	(-1.29)	(-0.63)	(-0.45)
Population	-0.0008	-	-0.0005	-	0.0002	-	0.0003	-
	(-0.89)	-	(-0.54)	-	(0.14)	-	(0.42)	-
Prop. Land Title	-0.3951	-0.3485	0.4597	0.4670	0.5894	0.6071	0.6563	0.6556
	(-0.44)	(-0.39)	(1.02)	(1.02)	(0.83)	(0.88)	(0.74)	(0.78)
Fines	-0.0009	-0.0021	-0.0012	-0.0014	-0.0015	-0.0012	-0.0015	-0.0012
	(-0.48)	(-1.40)	(-1.07)	(-1.61)	(-0.81)	(-1.52)	(-0.94)	(-0.81)
Dist. to Ibama	-0.0005	-0.0013	-0.0009	-0.0009	0.0013	0.0012	0.0005	0.0006
	(-0.24)	(-0.58)	(-0.74)	(-0.76)	(1.14)	(1.05)	(0.37)	(0.43)
Constant	5.6863	5.5114	16.317***	17.005***	10.267**	10.888**	3.4441	2.9096
	(0.65)	(0.64)	(3.41)	(3.53)	(2.08)	(2.21)	(0.60)	(0.50)

Notes: Land-use IVQR regression results for the median. For each farm size, the column in the left presents the main specification, and the column in the right excludes "Population" among the regressors. Spatially lagged regressors are omitted from the table.
t-statistics in parentheses, *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Table 3: Land-use Median Regression by Farm Size – Prop. of Land Title

	Small		Small–Medium		Medium–Large		Large	
Cost to Port	0.0045	0.0055	-0.0053	-0.0049	-0.0079**	-0.0088**	-0.0110***	-0.0119***
	(0.76)	(0.92)	(-1.58)	(-1.45)	(-2.07)	(-2.27)	(-2.91)	(-3.13)
Altitude	0.0010	0.0009	0.0000	0.0001	-0.0002	-0.0003	-0.0011	-0.0009
	(0.67)	(0.62)	(0.74)	(0.10)	(-0.27)	(-0.38)	(-1.48)	(-1.27)
Temp	-0.0383	-0.0284	-0.4680***	-0.4294**	-0.2649	-0.3019*	0.0229	-0.0265
	(-0.12)	(-0.09)	(-2.79)	(-2.54)	(-1.54)	(-1.76)	(0.11)	(-0.13)
Rain	-0.0038	-0.0037	-0.0094***	-0.009***	-0.0068**	-0.0072***	-0.0051*	-0.0062**
	(-0.97)	(-0.93)	(-3.53)	(-3.38)	(-2.55)	(-2.73)	(-1.65)	(-2.00)
Slope	-0.0846	-0.0877	-0.1015	-0.1799	-0.2076	-0.1874	-0.1965	-0.2875
	(-0.23)	(-0.24)	(-0.50)	(-0.85)	(-1.04)	(-0.95)	(-0.85)	(-1.26)
Soil 2	-2.1386***	-2.1276***	-1.8197***	-1.9908***	-1.5805***	-1.7028***	-1.6984***	-1.8049***
	(-2.83)	(-2.92)	(-4.25)	(-4.89)	(-3.06)	(-3.31)	(-2.62)	(-2.84)
Soil 3	-0.9227	-0.8689	-0.8729*	.-0.8805**	-0.9222*	-1.0041**	-0.9942**	-1.0573**
	(-0.84)	(-0.78)	(-1.93)	(-1.96)	(-2.27)	(-2.52)	(-2.28)	(-2.50)
Soil 4	-1.4279**	-1.4054**	-1.4894***	-1.5509***	-1.5151***	-1.5449	-1.6160***	-1.6795***
	(-2.41)	(-2.37)	(-4.78)	(-4.95)	(-4.91)	(-4.98)	(-4.50)	(-4.68)
Soil 5	-0.0730	-0.0478	-0.8029**	-0.9223***	-0.5831*	-0.6658	-0.6423*	-0.6728*
	(-0.11)	(-0.07)	(-2.38)	(-2.79)	(-1.84)	(-2.11)	(-1.70)	(-1.80)
Dist. to PA	-0.0059	-0.006	0.0005	-0.0001	0.0028	0.0027	0.0055*	0.0056*
	(-0.90)	(-0.92)	(0.16)	(-0.03)	(1.13)	(1.07)	(1.78)	(1.81)
Mining	-0.0136	-0.0166	0.0061	0.0054	-0.0432**	-0.0442**	0.0075	0.0097
	(-0.31)	(-0.39)	(0.15)	(0.12)	(-1.99)	(-2.00)	(0.28)	(0.36)
Power Plant	-0.5107	-0.4563	-0.2998	-0.3146	-0.1086	-0.1318	-0.0445	-0.0315
	(-1.29)	(-1.15)	(-1.07)	(-1.11)	(-0.33)	(-0.38)	(-0.14)	(-0.10)
Power Plant Neighbor	-0.4703	-0.5514	-0.3035	-0.356	-0.3089	-0.3057	-0.1575	-0.007
	(-1.33)	(-1.57)	(-1.25)	(-1.49)	(-1.37)	(-1.34)	(-0.63)	(-0.03)
Population	-0.0008	-0.0007	-0.0005	-0.0006	0.0002	0.0001	0.0003	0.0001
	(-0.89)	(-0.88)	(-0.54)	(-0.62)	(0.14)	(0.08)	(0.42)	(0.14)
Prop. Land Title	-0.3951	-	0.4597	-	0.5894	-	0.6563	-
	(-0.44)	-	(1.02)	-	(0.83)	-	(0.74)	-
Fines	-0.0009	-0.0009	-0.0012	-0.0012	-0.0015	-0.0015	-0.0015	-0.001
	(-0.48)	(-0.46)	(-1.07)	(-0.97)	(-0.81)	(-0.81)	(-0.94)	(-0.68)
Dist. to Ibama	-0.0005	-0.0008	-0.0009	-0.001	0.0013	0.0008	0.0005	0.0005
	(-0.24)	(-0.37)	(-0.74)	(-0.85)	(1.14)	(0.75)	(0.37)	(0.41)
Constant	5.6863	4.9876	16.317***	15.783***	10.267**	11.999**	3.4441	4.432
	(0.65)	(0.57)	(3.41)	(3.28)	(2.08)	(2.48)	(0.60)	(0.77)

Notes: Land-use IVQR regression results for the median. For each farm size, the column in the left presents the main specification, and the column in the right excludes "Proportion of Land Title" among the regressors. Spatially lagged regressors are omitted from the table.
t-statistics in parentheses, $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Land-use Median Regression by Farm Size – Fines and Distance to Ibama

	(1)	(2)	(3)	(4)
Small				
Costs to Port	0.0038	0.0042	0.0042	0.0045
	(0.0060)	(0.0060)	(0.0060)	(0.0060)
Fines	-	-0.0010	-	-0.0009
	-	(0.0019)	-	(0.0019)
Dist. IBAMA	-	-	-0.0004	-0.0005
	-	-	(0.0022)	(0.0022)
Small–Medium				
Costs to Port	-0.0055*	-0.0055	-0.0048	-0.0053
	(0.0033)	(0.0034)	(0.0034)	(0.0034)
Fines	-	-0.0012	-	-0.0012
	-	(0.0012)	-	(0.0012)
Dist. IBAMA	-	-	-0.0008	-0.0009
	-	-	(0.0012)	(0.0012)
Medium–Large				
Costs to Port	-0.0101***	-0.0093**	-0.0094**	-0.0079**
	(0.0037)	(0.0038)	(0.0038)	(0.0038)
Fines	-	-0.0012	-	-0.0015
	-	(0.0018)	-	(0.0019)
Dist. IBAMA	-	-	0.0011	0.0013
	-	-	(0.0011)	(0.0011)
Large				
Costs to Port	-0.0126***	-0.0116***	-0.0109***	-0.0110***
	(0.0036)	(0.0037)	(0.0038)	(0.0038)
Fines	-	-0.0016	-	-0.0015
	-	(0.0015)	-	(0.0016)
Dist. IBAMA	-	-	0.0008	0.0005
	-	-	(0.012)	(0.0013)

Notes: Land-use IVQR regression results for the median. The first column excludes "Fines" and "Distance to IBAMA;" the second column excludes only "Distance to IBAMA;" the third column excludes only "Fines;" and the last column presents the main specification. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Monitoring Efforts and Policy Implications

	Tax (US$/ha)	LS Tax (US$)	LS 80% Rule (US$)	Emissions at US$1/tCO$_2$
Actual Monitoring				
Specification				
Main	42.50	479 million	4.01 billion	-4.17 billion tC
No Dist. IBAMA	42.50	482 million	4.81 billion	-4.19 billion tC
No Fines	40.00	453 million	3.08 billion	-4.20 billion tC
No Fines/No Dist. IBAMA	37.50	425 million	3.89 billion	-4.21 billion tC
Fines at pre-2004 Level				
Specification				
Main	47.50	582 million	4.17 billion	-4.18 billion tC
No Dist. IBAMA	47.50	585 million	4.97 billion	-4.20 billion tC

Notes: This table reports the counterfactual policy implications for different model specifications. The first column presents the value of the uniform tax (US$/ha) that leads to 80% of the forested area. The second column presents the lost surplus (LS) of a uniform tax. The third column presents the lost surplus (LS) of the 80% rule. And the fourth column shows the avoided carbon emissions (in tons of carbon) associated with a US$1/CO$_2$ carbon tax. The top panel refers to the “actual monitoring” counterfactual; the bottom panel refers to the counterfactual in which I hold the number of fines after 2004 at pre-2004 levels. The different rows correspond to different estimated models: (i) the main specification; (ii) the model that excludes “Distance to IBAMA” among regressors; (iii) the model that excludes “Fines” among regressors; and (iv) the model that excludes both “Distance to IBAMA” and “Fines” among regressors. All monetary values are in dollars.

Table 6: Land-use Median Regression by Farm Size – Spatial Dependence

	(1)	(2)	(3)	(4)
Small				
Costs to Port	0.0040	0.0054	0.0045	0.0049
	(0.0057)	(0.0058)	(0.0060)	(0.0058)
Spatial Lag Population	-	0.0003	-0.0022	-0.0019
	-	(0.0014)	(0.0017)	(0.0025)
Spatial Lag Mining	-	-0.0066	0.0147	0.0420
	-	(0.0523)	(0.0541)	(0.0751)
Spatial Lag Power Plant	-	0.1972	-0.2428	-1.2371
	-	(0.6879)	(0.8007)	(1.4002)
Spatial Lag Dist. PA	-	0.0084	0.0101	0.0042
	-	(0.0053)	(0.0076)	(0.0063)
Small–Medium				
Costs to Port	-0.0051	-0.0049	-0.0053	-0.0059*
	(0.0032)	(0.0032)	(0.0034)	(0.0034)
Spatial Lag Population	-	0.0016**	0.0015	0.0012
	-	(0.0007)	(0.0013)	(0.0016)
Spatial Lag Mining	-	-0.0142	-0.0193	-0.0214
	-	(0.0302)	(0.0338)	(0.0361)
Spatial Lag Power Plant	-	-0.1016	-0.0574	-0.1236
	-	(0.3846)	(0.0019)	(0.6845)
Spatial Lag Dist. PA	-	0.0048	0.0010	-0.0011
	-	(0.0031)	(0.0034)	(0.0033)
Medium–Large				
Costs to Port	-0.0073**	-0.0062*	-0.0079**	-0.0056
	(0.0037)	(0.0037)	(0.0038)	(0.0037)
Spatial Lag Population	-	0.0011	0.0011	0.0017
	-	(0.0007)	(0.0013)	(0.0015)
Spatial Lag Mining	-	-0.0369	-0.0536*	-0.0285
	-	(0.0274)	(0.0322)	(0.0489)
Spatial Lag Power Plant	-	-0.1519	-0.1882	-0.6725
	-	(0.3755)	(0.5804)	(0.9669)
Spatial Lag Dist. PA	-	0.0029	-0.0022	-0.0020
	-	(0.0025)	(0.0029)	(0.0029)
Large				
Costs to Port	-0.0100***	-0.0098***	-0.0110***	-0.0115***
	(0.0038)	(0.0038)	(0.0038)	(0.0037)
Spatial Lag Population	-	0.0022	0.0002	0.0010
	-	(0.0040)	(0.0016)	(0.0022)
Spatial Lag Mining	-	-0.0660*	-0.0768**	-0.0848*
	-	(0.0395)	(0.0369)	(0.0461)
Spatial Lag Power Plant	-	-0.4322	0.0995	0.0149
	-	(0.4938)	(0.5199)	(0.7785)
Spatial Lag Dist. PA	-	0.0015	-0.0034	-0.0032
	-	(0.0031)	(0.0033)	(0.0032)
Average # of Neighbors	-	3.5	7.3	12

Notes: IVQR regression results for the median. Column (1) excludes spatially lagged regressors; in column (2) the cutoff distance is 50 km; in column (3) the cutoff distance is 75km; in column (4) the cutoff is 100 km. The spatial weight matrices are of the power functional type with a distance decay parameter that equals 1. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Land-use Regression Model by Farm Size – Standard Errors

	Quantiles				
	10	25	50	75	90
Small					
Coefficient on Cost to Port	0.0080	0.0027	0.0045	-0.0003	-0.0121
IVQR SE	(0.0045)*	(0.0049)	(0.0060)	(0.0083)	(0.0249)
Cluster-Bootstrap SE, IR	(0.0051)	(0.0047)	(0.0056)	(0.0145)	(0.0362)
Cluster-Bootstrap SE, C = 50	(0.0055)	(0.0049)	(0.0057)	(0.0175)	(0.0361)
Cluster-Bootstrap SE, C = 150	(0.0050)	(0.0043)	(0.0055)	(0.0142)	(0.0351)
Small–Medium					
Coefficient on Cost to Port	-0.0049	-0.0023	-0.0053	-0.0063	-0.0008
IVQR SE	(0.0035)	(0.0035)	(0.0034)	(0.0038)*	(0.0069)
Cluster-Bootstrap SE, IR	(0.0044)	(0.0041)	(0.0039)	(0.0058)	(0.0129)
Cluster-Bootstrap SE, C = 50	(0.0047)	(0.0045)	(0.0043)	(0.0063)	(0.0121)
Cluster-Bootstrap SE, C = 150	(0.0044)	(0.0040)	(0.0036)	(0.0055)	(0.0115)
Medium–Large					
Coefficient on Cost to Port	-0.0184	-0.0112	-0.0079	-0.0042	-0.0010
IVQR SE	(0.0038)***	(0.0037)***	(0.0038)**	(0.0040)	(0.0049)
Cluster-Bootstrap SE, IR	(0.0050)***	(0.0047)**	(0.0050)	(0.0064)	(0.0109)
Cluster-Bootstrap SE, C = 50	(0.0057)***	(0.0050)**	(0.0055)	(0.0066)	(0.0095)
Cluster-Bootstrap SE, C = 150	(0.0053)***	(0.0042)***	(0.0046)*	(0.0057)	(0.0089)
Large					
Coefficient on Cost to Port	-0.0110	-0.0109	-0.0110	-0.0107	-0.0156
IVQR SE	(0.0042)***	(0.0040)***	(0.0038)***	(0.0042)**	(0.0050)***
Cluster-Bootstrap SE, IR	(0.0059)*	(0.0043)**	(0.0042)***	(0.0054)**	(0.0083)*
Cluster-Bootstrap SE, C = 50	(0.0077)	(0.0052)**	(0.0051)**	(0.0060)*	(0.0099)
Cluster-Bootstrap SE, C = 150	(0.0062)*	(0.0042)***	(0.0041)***	(0.0053)**	(0.0081)*

Notes: This table reports the estimated coefficients on transportation costs based on the IVQR estimators. For each farm size, the dependent variable is the log odds ratio of the share of deforestation. The "IVQR SE" denotes the standard errors of the IVQR model as developed by Chernozhukov and Hansen (2008). The "Cluster-Bootstrap SE" denotes the standard errors calculated using the "pairs cluster bootstrap," as explained in the text. "IR" refers to the official division of the Brazilian territory into "immediate regions," corresponding to 82 clusters in the data. The terms "C = 50" and "C = 150" refer to the clusters generated by an unsupervised machine learning algorithm in ArcGIS, with 50 and 150 clusters, respectively. The number of bootstrap samples is 500. The unit of observation is a municipality in the Amazon.
The number of observations for small farms is 501, for small-medium farms is 520,
for medium–large farms is 520, and for large farms is 450.
Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Land-use Regression Model by Farm Size – Choice of Instrument

		Quantiles				
	2SLS	10	25	50	75	90
Small						
IV – Dist to Ports	-0.0015	0.0067	0.0023	0.0050	-0.0003	-0.0068
	(0.0089)	(0.0045)	(0.0049)	(0.0060)	(0.0084)	(0.0236)
IV – Both Instruments	-0.0024	0.0080*	0.0027	0.0045	-0.0003	-0.0121
	(0.0089)	(0.0045)	(0.0049)	(0.0060)	(0.0083)	(0.0249)
Small–Medium						
IV – Dist to Ports	-0.0066	-0.0050	-0.0024	-0.0062*	-0.0064*	-0.0008
	(0.0051)	(0.0035)	(0.0035)	(0.0034)	(0.0038)	(0.0069)
IV – Both Instruments	-0.0063	-0.0049	-0.0023	-0.0053	-0.0063*	-0.0008
	(0.0052)	(0.0035)	(0.0035)	(0.0034)	(0.0038)	(0.0069)
Medium–Large						
IV – Dist to Ports	-0.0062	-0.0190***	-0.0115***	-0.0084**	-0.0041	-0.0024
	(0.0037)	(0.0039)	(0.0038)	(0.0038)	(0.0040)	(0.0054)
IV – Both Instruments	-0.0058	-0.0184***	-0.0112***	-0.0079**	-0.0042	-0.0010
	(0.0039)	(0.0038)	(0.0037)	(0.0038)	(0.0040)	(0.0049)
Large						
IV – Dist to Ports	-0.0074	-0.0108***	-0.0108***	-0.0110***	-0.0106**	-0.0142***
	(0.0056)	(0.0042)	(0.0041)	(0.0038)	(0.0043)	(0.0050)
IV – Both Instruments	-0.0077	-0.0110***	-0.0109***	-0.0110***	-0.0107**	-0.0156***
	(0.0058)	(0.0042)	(0.0040)	(0.0038)	(0.0042)	(0.0050)

Notes: This table reports the estimated coefficients on transportation costs based on the 2SLS and IVQR estimators. For each farm size, the dependent variable is the log odds ratio of the share of deforestation. The specification "IV – Dist to Ports" uses distance to ports as the only instrumental variable for costs to ports; and "IV – Both Instruments" uses both distance to ports and distance to capital as instruments. The unit of observation is a municipality in the Amazon. The number of observations for small farms is 501, for small-medium farms is 520, for medium–large farms is 520, and for large farms is 450. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Choice of Instrument and Policy Implications

	Tax (US$/ha)	LS Tax (US$)	LS 80% Rule (US$)	Emissions at US$1/tCO$_2$
Specification				
Main	42.50	479 million	4.01 billion	-4.17 billion tC
IV – Distance to Ports	42.50	483 million	4.11 billion	-4.22 billion tC

Notes: This table reports the counterfactual policy implications for different model specifications. The first column presents the value of the uniform tax (US$/ha) that leads to 80% of the forested area. The second column presents the lost surplus (LS) of a uniform tax. The third column presents the lost surplus of the 80% rule. And the fourth column shows the avoided carbon emissions (in tons of carbon) associated with a US$1/CO$_2$ carbon tax. The first row, "Main," refers to the main specification presented in the paper, which makes use of both distance to port and distance to capital as instrumental variables for costs to ports. The second row, "IV – Distance to Ports," refers to the specification that uses distance to ports as the only instrument for costs to ports. All monetary values are in dollars.

Table 10: Comparison between the IVQR and the SPQIV at the Median

	Small-Medium		Medium-Large		Large	
	IVQR	SPQIV	IVQR	SPQIV	IVQR	SPQIV
Cost to Port	-1	-1	-1	-1	-1	-1
Altitude	0.0083	0.0083	-0.0233	-0.0233	-0.1012	-0.1012
	(0.1100)		(0.089)		(0.0815)	
Temp	-87.887**	-87.887	-33.687*	-33.687	-2.0766	-2.0766
	(43.834)		(19.301)		(18.416)	
Rain	-1.7606*	-1.8266	-0.8658**	-0.8658	-0.4625	-0.4855
	(1.0582)		(0.4182)		(0.2835)	
Slope	-19.06	-19.06	-26.398	-26.398	-17.802	-17.802
	(38.803)		(28.962)		(21.729)	
Soil 2	-341.71	-341.71	-201.02*	-201.02	-153.86*	-153.86
	(247.58)		(121.42)		(85.212)	
Soil 3	-163.93	-163.93	-117.28*	-117.28	-90.068**	-90.068
	(110.56)		(62.678)		(42.888)	
Soil 4	-279.69	-283.19	-192.7**	-199.93	-146.39**	-146.39
	(175.35)		(96.992)		(58.181)	
Soil 5	-150.77	-150.77	-74.165	-74.165	-58.186	-58.186
	(101.01)		(46.243)		(37.116)	
Dist. to PA	0.0867	0.0867	0.3558	0.3558	0.5017	0.5017
	(0.5368)		(0.3701)		(0.345)	
Mining	1.1417	1.1417	-5.4907	-5.4907	0.6757	0.6757
	(7.837)		(3.5109)		(2.4962)	
Power Plant	-56.307	-56.307	-13.811	-13.811	-4.0286	-4.0286
	(64.501)		(42.431)		(29.125)	
Power Plant Neighbor	-56.998	-56.998	-39.287	-39.287	-14.272	-14.272
	(59.719)		(36.445)		(23.867)	
Population	-0.0884	-0.088	0.0274	0.0274	0.025	0.025
	(0.1617)		(0.2004)		(0.0608)	
Prop. Land Title	86.32	86.32	74.961	74.961	59.457	59.466
	(102.41)		(98.393)		(82.761)	
Fines	-0.2332	-0.2332	-0.1907	-0.1907	-0.1316	-0.1316
	(0.2847)		(0.2741)		(0.1561)	
Dist. to IBAMA	-0.1617	-0.1617	0.1613	0.1633	0.0429	0.0429
	(0.2309)		(0.1746)		(0.1204)	

Notes: Land-use regression results for the median. For each farm size, the column in the left presents the logit model, where the constant is normalized to 0 and the coefficient to transport cost is normalized to -1; the column in the right the semiparametric model.
Spatially lagged regressors are omitted from the table.
Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: Yields Regression (2SLS) by Farm Size – Crop-Pasture Index

	Small	Small–Medium	Medium–Large	Large
Cost to Port	0.0038*	-0.0009	0.0010	0.0018
	(0.0015)	(0.0008)	(0.0011)	(0.0015)
Observations	496	517	513	446
Average Local Yields (q_m^{cp})	0.70	0.42	0.41	0.27

Standard errors in parentheses, *$p<0.05$.

Table 12: Yields Regression (2SLS) for Crops by Farm Size – Coefficients for Costs to Ports

	Small	Small–Medium	Medium–Large	Large
Index – Only Crops $\left(q_m^c\right)$	0.0073*	-0.0021	-0.0052*	0.0098*
	(0.0022)	(0.0014)	(0.0026)	(0.0040)
Soy	-0.0001	-0.0007	-0.0002	-0.0022
	(0.0001)	(0.0009)	(0.0011)	(0.0018)
Corn	0.0033	-0.0012	-0.0019	-0.0012
	(0.0028)	(0.0019)	(0.0017)	(0.0028)
Rice	-0.0001	-0.0007	-0.0019	-0.0027
	(0.0011)	(0.0013)	(0.0015)	(0.0021)
Beans	0.0025	0.00003	-0.0003	-0.0012
	(0.0017)	(0.0010)	(0.0010)	(0.0014)
Manioc	0.0074	-0.0029	-0.0064	0.0052
	(0.0047)	(0.0039)	(0.0036)	(0.0032)
Observations	496	517	513	446
Average Local Yields $\left(q_m^c\right)$	1.03	1.04	1.16	0.87

Standard errors in parentheses, *p<0.05.

Table 13: Regress Census Deforestation on Satellite Deforestation and Covariates

	Full Sample	Restricted Sample
Satellite Prop. Deforestation	0.400***	0.330***
	(0.0260)	(0.0884)
Dist. to Port	-0.00000694	-0.0000595
	(0.0000126)	(0.0000833)
Dist. to Capital	0.0000664	-0.0000432
	(0.0000351)	(0.000187)
Altitude	0.0000316	-0.0000456
	(0.0000557)	(0.000154)
Temp	-0.0191	-0.0339
	(0.0149)	(0.0488)
Rain	-0.00113***	-0.00119
	(0.000279)	(0.000875)
Slope	-0.0337	0.000476
	(0.0177)	(0.0286)
Soil 2	-0.227***	1.351
	(0.0434)	(2.021)
Soil 3	-0.188***	-0.0343
	(0.0489)	(0.0780)
Soil 4	-0.250***	-0.0606
	(0.0356)	(0.0776)
Soil 5	-0.231***	0.0125
	(0.0347)	(0.0769)
Dist. to PA	0.000994***	0.000169
	(0.000206)	(0.000465)
Mining	-0.000816	-0.00847
	(0.00186)	(0.0266)
Power Plant	-0.000879	-0.00779
	(0.0187)	(0.1150)
Power Plant Neighbor	0.00828	0.0318
	(0.0147)	(0.0323)
Population	-0.0000426	0.000634
	(0.0000603)	(0.00110)
Prop. Land Title	0.101*	0.0651
	(0.0413)	(0.2310)
Fines	-0.000101	-0.000305
	(0.0000988)	(0.000332)
Dist. to IBAMA	0.000118	0.000166
	(0.000110)	(0.000583)
Prop. Clouds	-0.0381	-29.61
	(0.0321)	(24.56)
Constant	0.909*	1.537
	(0.4380)	(1.4170)
Observations	523	45
Adjusted-R^2	0.67	0.64

Notes: The second column presents the estimated coefficients for the full sample; and the third column, for the restricted sample. Both use the Eicker-Huber-White robust standard errors. The restricted sample includes municipalities in which the share of private land is greater than 80%, and the share of clouds/unobserved areas is less than 3%. Standard errors in parentheses, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: Unit Costs to Travel

Unit Cost	R$/ton.km	US$/ton.km
Paved Road	0.07678	0.0353
Unpaved Road – Outside Rainforest	0.0992	0.0457
Unpaved Road – Within Rainforest	0.15	0.069
Navigable River – Good Infrastructure	0.0444	0.0204
Navigable River – Poor Infrastructure	0.1139	0.0525
Railroad	0.0608	0.028
Land – Outside Rainforest	1.5	0.6912
Land – Inside Rainforest	3	1.3824