Towns and Rural Land Inequality in India

Manaswini Rao, Juan Eberhard, Prashant Bharadwaj*

January 26, 2026

Abstract

Using the universe of land records from a large state in India, we document three empirical facts on rural land holding inequality at the village-level: 1) inequality is higher close to urban areas and decreases with distance, 2) this is due to fewer medium-sized farms, and 3) the distance to urban area-land holding inequality relationship depends on the size of the urban area. To explain these patterns, we build a parsimonious model where individual farmers face a U-shaped agriculture production function between land size and farm productivity, bequest considerations of their land among their children, and a significant urban opportunity cost of farm production. Our findings bring attention to spatial patterns in land holding inequality that could have implications for structural transformation.

JEL Codes: Economic Development (O1), Urban, Rural Regional, and Transportation Analysis (O18)

*Author affiliations: Rao - University of Delaware Department of Economics (email: manaswini.rao@gmail.com), Bharadwaj - UC San Diego Department of Economics (email: prbhardwaj@ucsd.edu), Eberhad - Universidad Andres Bello, Facultad de Economia y Negocios (email: juan.eberhard@gmail.com). We are grateful to Andrew Foster, Karthik Muralidharan, and Siddharth George for feedback.

1 Introduction

Why do individuals in rural areas of developing countries persist in small scale farming? Why are small farmers unable to expand, and why do they not migrate to urban areas? These questions have been the focus of research in development economics for decades and are intimately connected with questions regarding the existence of poverty traps, market frictions, and structural transformation.[1] A key empirical fact that has been proposed in the literature that is consistent with models of poverty traps is that there are many more small farms in developing countries than in developed countries (Rigg, Salamanca, and Thompson 2016). Large scale farming is more efficient than small scale farming (especially since large farms can take advantage of mechanization); hence, the existence of many small farms is often taken as *prima facie* evidence of the existence of poverty traps, which might arise due to various market frictions and frictions related to rural-urban migration in developing countries (Foster and Rosenzweig 2022; Lagakos 2020).

In this paper, we aim to provide a new insight by examining village-level land holding distributions using the universe of land records data from a large state in India. We characterize these distributions using measures of inequality such as the Gini coefficient. While there are several interesting questions about land holding patterns that we could ask, we first describe basic spatial patterns of rural land holding inequality in relation to urban areas. This is a novel exercise since land records for the universe of land holders is generally not accessible in rural areas of developing countries and any survey-related datasets are typically not representative of the population-level land-holding distribution. We argue that the observed spatial patterns, when viewed through the lens of our parsimonious yet intuitive model, shed light on the existence of preponderance of small farms and its association with urbanization. We want to be upfront that this is not a paper about whether urbanization *causes* these land holding patterns to emerge or vice-versa; instead, the paper first describes and then seeks

[1] This research includes some of the seminal theoretical work in this space such as Rosenstein-Rodan (1961), Murphy, Shleifer, and Vishny (1989), Banerjee and Newman (1993), and Banerjee and Newman (1994). The theoretical work has lead to a large empirical literature attempting to find evidence for poverty traps (see Kraay and McKenzie 2014 for an excellent review, Gollin, Lagakos, and Waugh 2014 who argue about importance of agricultural productivity differences and income inequality, and Balboni, Bandiera, Burgess, Ghatak, and Heil 2021 as a recent example of empirical work on poverty traps).

to understand through a model with minimal assumptions that could generate such patterns. We also want to be clear that there could be other models, such as those invoking specific market failures, that might generate our observed patterns. Our goal in putting forth this model was to focus on the role of urbanization in the tradeoff between preferences over land ownership and returns from migration, especially given the tremendous attention in the literature and policy focus on land redistribution and limited voluntary migration in low income countries (Rosenzweig 1978; Jayachandran 2006; Bryan, Chowdhury, and Mobarak 2014; Munshi and Rosenzweig 2016; Bazzi 2017; Bryan and Morten 2019).

We compute land inequality as a Gini coefficient at the village level using the land records data. An important feature of Indian agriculture, and in many other developing countries, is that a vast majority of households are small, land owning, farmers (approximately 80% own less than 5 acres of land in India); therefore, the Gini makes sense as an intuitive measure of how land is held at the village level.[2] Using the Gini as the basis for measuring land holding inequality, we highlight three empirical patterns: first, we note that land inequality is higher near towns and lower in villages further away from towns. Second, higher Gini coefficient near towns is driven by fewer mid-sized farmers (with 5-8 acre farms). In other words, villages near towns have relatively more small *and* large farms (compared to middle sized farms).[3] Third, the Gini-distance relationship is steeper for larger towns compared to smaller towns. These patterns are also observed among landholding distributions generated using household census dataset from the same state that enumerates household-level landholding (complementing the individual-level landholding distribution from the administrative dataset) as well as in an all-India representative household panel dataset (India Human Development Survey or IHDS dataset). These patterns are also agnostic to definitions of village or other administrative borders that could endogenously

[2] We also examine the robustness of our empirical findings using other parameters of the distribution including the mean and standard deviation as well as share of farmers by different size-bins. The Gini also correlates strongly with key parameters of skewed distributions such as Pareto distribution, typically used in analyzing asset-inequality as in Bazzi (2017). Importantly, the Gini measure is non-parametric and thus, captures the empirical feature of the distribution of land among population within an area. Furthermore, we calculate landholding gini agnostic of village boundaries using multiple distance bands surrounding urban centers to test for the robustness of our empirical facts against any endogenous changes to village borders.

[3] Our findings are robust to different cut-offs classifying landholding sizes into small, mid, and large. We use 5 and 8 acre cut-offs following official definitions of small and large landholdings.

respond to urban influence.

These observed spatial patterns are consistent with a model where agricultural productivity is a U-shaped function in farm size (as in Cornia 1985; Barrett 1996, and more recently revisited by Gaurav and Mishra 2015 and Foster and Rosenzweig 2022)[4], farmers face bequest considerations that depend on their number of children, and an urban wage premium that dissipates with distance from the city (due to migration-type costs that increase with distance). With these key ingredients of the model, we are able to simulate a data generating process that fit the empirical patterns discussed above. Hence, according to this model, the process of structural transformation (one of the features of this is the act of moving out of agriculture - see Syrquin 1988) is possible for mid-sized farmers, close to towns, but not for small farmers close or far away from towns. This generates the observed distribution of higher land inequality near towns. Our model provides a lens as to why, even in the presence of high urban wages and other urban advantages, small farmers who live nearby are unlikely to move or expand their farming activities.

We test the implications of the model using multiple datasets including the all-India household panel IHDS dataset that contains information from two time periods, 7 years apart. First, we examine transitions in landholding sizes among households by baseline land-holding size bins in the subsequent period. We find that over 90% of the smallholders remain small and only a small fraction of them upgrade to larger landholdings. Any land transition observed is concentrated among mid and large-sized landholders. Further, these transition patterns are correlated with distance to town such that large landholders become mid-sized with increasing distance from towns. Second, our model implies that large farms close to towns will get *larger*: we confirm this relationship by examining the ratio of landholding sizes at the 99th percentile relative to 25th percentile across smaller vs. larger towns. Third, we find that mid-sized landholding households are relatively more absent closer to towns when tracked at endline 7 years later (serving as a proxy for potential migration). Lastly, we find that higher land inequality near towns is associated with increased mechanization and higher agricultural yields.

[4] A large empirical literature has documented a negative farm-size relationship with agricultural productivity. Recent work has emphasized the presence of measurement error and heterogeneity in farm and production shocks to explain the non-monotonic and U-shape patterns (see Gollin and Udry 2021).

Finally, we examine alternate explanations including differences in skilling and financial frictions. We find no differential household composition (number of children) or skilling by landholding sizes *and* distance to town. We also rule out many other explanations, all of which do not recreate the observed patterns in the data. Further, we show that our model can incorporate complementary mechanisms with additional assumptions. For example, our model can incorporate financial frictions as additional constraints on expanding landholding sizes. In line with this, we observe the key distance patterns based on whether a village has a bank branch or other financial institution present. However, our preferred explanation does not need to invoke additional frictions - non-linearity in agricultural productivity by landholding size and land market frictions due to fixed supply of land are sufficient to generate the spatial patterns we observe.

Our paper is related to many different strands of the literature. First, it adds to the recent empirical literature on structural transformation and spatial frictions (Heise and Porzio 2021; Madhok, Noack, Mobarak, and Deschenes 2025), including rural-urban migration (Young 2013; Bryan, Chowdhury, and Mobarak 2014; Munshi and Rosenzweig 2016; Bazzi, Gaduh, Rothenberg, and Wong 2016), by bringing attention to selective exit from agriculture based on asset (land) ownership rather than skills. Our patterns support a model where farm-level productivity and individual life-cycle concerns make remaining in agriculture more attractive for small farmers but not for mid-sized farmers close to towns who earn relatively better returns (as also modeled by Lagakos, Mobarak, and Waugh 2023 requiring low relative productivity in rural areas for migration). The patterns we demonstrate is supported not just by variations in the land inequality Gini coefficient but also by the absolute and relative shares of mid-sized farmers with respect to small and large farmers.

Second, our paper models the relationship between land inequality and urban influence independent of the initial conditions, acknowledging that that these initial conditions could matter. In this regard, our paper extends the work by Bardhan, Luca, Mookherjee, and Pino (2014), where the authors examine the effect of land tenancy reform (leading to land redistribution) in West Bengal and subsequent demographic transitions generating land inequality through land subdivisions from inheritances. Our model is consistent with this empirical observation of lifecycle changes and subdivision of land due to inheritance. We show is that such lifecycle transitions apply to all farming households within a village, generating intra-village land distribution that is constantly changing. However, it motivates mid-sized farmers in villages close to town to either (a) exit agriculture (because it is suboptimal given their landholding size and urban wage premium), and/or (b) downsize or upsize their landholding for optimal farm production.

Third, it complements the vast literature on urban development (such as in China documented by Kahn, Sun, Wu, and Zheng 2018, and in Africa documented by Henderson and Kriticos 2018) by linking it with rural landholding patterns and the selective role of urban opportunities. Saiz (2010) and Hsieh and Moretti (2019) note that land restrictions generate scarcity in urban housing supply that drive up prices and wages, affecting labor allocation across space and subsequently, economic growth. Our work sheds light on the role of urban opportunity on land distribution in surrounding rural areas, opening a new area for additional research in determining the causal relationships in both directions in light of Glaeser and Henderson (2017) and Henderson and Turner (2020) - whether the ability to consolidate land in rural areas leads to higher urban development in such areas, and whether higher wages in urban areas motivate increasing consolidation in surrounding rural areas.

Finally and importantly, we add to development economics literature that has focused on land ownership patterns from the point of view of tenure security and incentives for investments to improve agricultural productivity (Banerjee, Gertler, and Ghatak 2002; Burchardi, Gulesci, Lerva, and Sulaiman 2019; Adamopoulos and Restuccia 2020) and its relationship to migration (de Janvry, Emerick, Gonzalez-Navarro, and Sadoulet 2015) by documenting novel correlations between rural land inequality and urbanization. Our research has direct relevance to understanding the existence of asset-based poverty traps in developing countries (Balboni, Bandiera, Burgess, Ghatak, and Heil 2021).

The rest of the paper is as follows. Section 2 describes the data sources. Section 3 documents the key empirical patterns. Section 4 presents a simple model describing the economic processes generating the observed correlations. Section 5 examines the implication of the model. Section 6 examines alternative theories and welfare implications and Section 7 concludes.

2 Data

We exploit multiple data sources from India - administrative, satellite imagery, and survey-based - to empirically identify some key facts on the relationship between rural land inequality and proximity to urban centers. We focus on a village as the basic unit of analysis to examine the spatial patterns. We construct the land inequality measures and exploit other outcomes at the village-level using disaggregated unit-level data (both at landholder and household levels), which we discuss in detail in the subsections below. Aggregating data at the village-level provides thousands of observations at varying distances from many distinct towns.

2.1 Definitions

Towns considered in this paper are what Census of India, 2011, defines as statutory towns. This definition is based on whether a census unit has a legally recognized urban local governance body such as a municipality, a town corporation, a cantonment board, or a notified town area committee. This definition makes towns in our analysis larger than population-density/cluster based definitions of towns (called census towns). Thus, all our spatial analysis is with respect to regions that are urban areas in an economic sense and not merely population clusters. We have a total of 49 distinct statutory towns in our sample.

Similarly, villages considered in our analysis are as per the definition of Census of India, 2011. Any administrative area that is not classified either as a statutory town (our definition of an urban center) or as a census town (based on population density threshold) are considered rural. A village may contain one or more hamlets (clusters of households living in close proximity), all of which are considered part of the same village. There are over 9000 census villages in our sample. Land records are maintained for every village for revenue administration by the respective state, which form one of the main datasets in our analysis.

2.2 Land Records

We use land revenue records that contain landholding data spanning the universe of rural land holders in one large state in India. The main fields in this dataset that we use for our analysis include land ownership details - unique ids with basic demographic detail, area owned, and village identifiers. These are based on recent field-level updates to land records to capture the most accurate status of land ownership since they were also used for income transfers under both state and national-level farmer income support programs.

We classify a farm as a small farm if the extent of land is less than 5 acres, as mid-sized if the land holding is between 5-8 acres, and as large for those more than 8 acres. This is similar to the categorization followed under the Agricultural Census by the Ministry of Agriculture and Farmer Welfare, Government of India.[5]

2.3 Census Data

In addition to the above-mentioned administrative data on the universe of land holdings, we also employ both the population census village-level aggregates and socio-economic and caste census household data for the particular state. The idea behind using two different sources of *population-level* data is to: (a) verify whether the patterns replicate in either datasets, and (b) address any measurement error issues based on whether land holding is recorded at an individual-level (one with the title) or the household-level. The population census data also includes the set of landless households, therefore, the Gini coefficients measured thus incorporates inequality in access to land as well.

2.3.1 Socio-Economic and Caste Census (SECC)

The Government of India administered a comprehensive socio-economic and caste census during the 2011 population census, and collected detailed information on education, occupation, income, and assets (including land) at the household-level.[6] We

[5]The Agricultural Census has a finer classification as follows: (a) marginal farms as those with below 1 hectare/2.47 acres of land; (b) small farms between 1-2 hectare or 2.47-5 acres; (c) semi-medium farms as those with 2-4 hectare or 5-10 acres; (d) medium farms as those with 4-10 hectare or 10-25 acre; and (e) large farms as those with over 10 hectare or 25 acres of land. Since the number of farms with over 10 acres is very small in our data, we chose to use our coarser classification which captures the land size cut-off between small and medium farmers. We test for robustness around varying definitions of these land-size bins.

[6]Some of this data is even reported at the individual-level but land is mainly reported at the household-level.

use this census micro-data for the corresponding state to construct village-level landholding Gini-index that includes the landless to verify the patterns generated using the land records data. We also use other variables in this dataset, such as the extent of farm mechanization by household, to test the implications of our model.

2.3.2 Population Census

From the 2011 population census, we use the village-level primary census abstract documenting village population and amenities, including distance from the nearest statutory town in kilometer (km). We focus on the following variables: village-level population; number of households; total geographic, cultivated, irrigated, and non-agricultural areas; primary crop produced; road connectivity via district, state, or national highways; presence of a bank branch; and the name and distance to the nearest town. We merge this with the town-level census by the nearest town, to obtain the corresponding town population. Note that this nearest town could be located in another district or another state, altogether.

2.4 Satellite Data

We use FAO GAEZ satellite data products to examine agricultural outcomes since productivity measures are not reported either in the land records data or in the population census data. Specifically, we use the yield achievement ratio (yield and production gap) thematic product to examine whether there is also a distance pattern in better agricultural outcomes and how this correlates with land holding inequality. The yield achievement ratio is our preferred productivity measure as it reflects the difference between realized yield and "predicted" yield based on soil and climatic suitability of a crop.

We download these products as raster images from FAO GAEZ v4 data portal. We then overlap these raster images with the population census administrative boundary shapefiles of the villages in the state we study to obtain the mean values of potential yields and yield achievement ratio, respectively, across pixels contained within each village polygon.

We also use this data to carry out multiple robustness checks including to examine whether the distance pattern we observe is due to any sorting by productivity of land for specific crops (mainly wetland rice and cotton).

2.5 Household Panel Survey: IHDS

We also use the pan-India India Human Development Survey (IHDS) data, where surveys were conducted in 2005 (Desai, Vanneman, and National Council of Applied Economic Research 2005) and again in 2012 (Desai, Vanneman, and National Council of Applied Economic Research 2012).[7] This dataset consists of a panel of a representative sample of households across India, using a multi-stage random sampling design: villages (for rural sample) were first randomly selected to form the primary sampling unit (PSU), and within the sampled villages, the surveys were administered to a random sample of households. A critical feature of this dataset is that the sample of households can be tracked across the two waves. We use this dataset to examine life cycle changes at a household-level over time - attrition from the first survey round to the next, share of children, household splits, and land inheritance as transitions between the two time periods. We note however that this survey data is representative of households at the national-level but not representative across the land holding distribution within a village.

2.6 Merging Datasets

We fuzzy merge land records data with census 2011 village amenities data using district, sub-district, and village names using the Hungarian Method, minimizing the string distance between the triple (village name, sub-district name, district name) in the two datasets.[8] The algorithm yields 75% match rate, which we further improve by manually matching the unmatched villages. We also merge other secondary datasets including economic census of all enterprises in our study villages using primary keys (SHRUG ID) from Asher and Novosad (2020) and Asher, Lunt, Matsuura, and Novosad (2021).

[7] This is limited by number of panel rounds available with latest round currently in the field.

[8] This method has also been used in merging census datasets with crime and labor market records in the United States in Norris, Pecenco, and Weaver (2021).

2.7 Constructing Land Inequality Measure

We first construct measures of land inequality as a Gini coefficient at the village-level using the administrative land records data. This is as fine grained as one could get since we are interested in landholding distribution at the village-level, which is also the basic unit of administration in India. We construct this empirically as:

$$Gini_v = 2 * \frac{1}{N * 100} \sum_{i \in v} (p_i - l_i)$$

where p_i is the percentile rank of farmer i in village v based on the size of their land holding, and l_i is the cumulative share of land held by all farmers ordered by their percentile rank below i. Figure 1 shows the spatial distribution of the village-level landholding inequality and urban areas.

2.7.1 Robustness of the constructed measure

We address potential issues with the land records dataset as follows. To account for land registered under different household members, we also generate a household-level measure of land ownership using the SECC dataset by aggregating land ownership within a household. This, complementary survey-based dataset, also helps account for landless individuals and households in the same village while computing the overall village-level Gini coefficient. Figure A1 shows that examining landholding inequality among landed individuals (farmers) is a lower-bound on the household-level land inequality and strongly correlated with the Gini calculated using household-level data.

In addition to constructing the landholding gini at a village-level using other datasets, we also construct the gini parameter pooling all landholdings in each 5 km distance bin from urban centers in our sample. We use this to generate empirical patterns that are agnostic to administrative definitions of village boundaries, which could endogenously respond to urban influence.

2.8 Summary Statistics on Land Inequality

A summary of the land records data in Figure 2 shows that there are too many small and marginal farmers, who own less than their share of land within these villages. An overwhelming majority - over 85 percent - are small farmers, with average land holding of about 2 acres. Mid-sized farmers, with average land holding of 6 acres, constitute about 10 percent of farmers whereas only 5 percent of all farmers are large farmers with average land holding of 12 acres. Despite being a small share, large farmers own close to a fifth of all land. On the other hand, mid-sized farmers constitute less than 20 percent of all farmers in a majority of the villages and in over 10 percent of the villages, there are no mid-sized farmers.

A more equal distribution, albeit with some inequality, should imply that each group of farmers own the same share of land as their share among all farmers. Instead, what we see in Figure 2 suggests significant land concentration. Further, the ratio between land size at the 99th percentile and that at the 25th percentile shows wide variation in our data (see Panel B Figure 2). Not surprisingly, land inequality - measured as a village-level Gini coefficient as described above - also shows substantial variation across villages in our study sample (Figure A1).

3 Key Empirical Facts

We document three key empirical patterns relating rural land inequality and proximity to towns, which we detail below. First, landholding inequality (Gini index) is higher in villages close to towns, which stabilizes after a certain distance (roughly 35-40 km from towns). Second, villages close to towns have fewer mid-sized farmers relative to either small or large farmers. Finally, the relationship between distance to town and landholding inequality becomes steeper as town size increases. We discuss each of these facts in greater detail below.

Fact 1: Increasing Rural Land Inequality Towards Towns

We find a steep urban distance gradient in rural landholding inequality, where the village-level landholding Gini index is strongly and negatively correlated with the distance of the village to its nearest urban center. To see this, we run the following regression specification as shown in Equation 1 to document the distance-inequality correlation in a non-parametric form using 5 km distance-bins between the study villages and their nearest towns:

$$Gini_{v(t)m} = \delta_t + \delta_m + \sum_{j \neq 35-40 \text{ km}} \beta_j D_{jv(t)m} + \epsilon_{v(t)m} \tag{1}$$

$D_{jv(t)m}$ is a dummy that takes value 1 when village $v(t)$ in the vicinity of town t is within a specific 5 km distance bin. The leave-out group is villages that are in the $35 - 40$ km distance bin from the nearest town and the farthest bin includes all distances greater than 75 km. A distance of 0-5 km implies that the village is outside the town limits but is within the "peri-urban" area. Note that our choice of the leave-out group is just for the purpose of illustration and we could replace this with other distance bins as the reference group. δ_t and δ_m are nearest town and sub-district fixed effects, respectively. We also estimate the correlations without any fixed effects to document similar patterns. Standard errors are clustered by the nearest town, to account for spatial correlation in the data.[9]

Panel A Figure 3 and Column 1 Table 2 show the distance correlations. The correlation is positive and significant in villages within 35 kms from a town. We also note that this correlation dissipates with distance. Panel A of Table 2 presents the parametric estimates of this relationship modeled using OLS with fixed effects as in Equation 1. For ease of readability, we transform the Gini index into standard deviation (σ) units relative to the "baseline" group of villages that are 35-40 km away. The coefficients imply around 0.1 σ increase in land inequality for every 10 km reduction in a village's distance from its closest town.

The observed patterns exist even as a binary correlation without the nearest town and sub-district fixed effects (Column 1 Panel B Table 2). Our intention behind adding these fixed effects is to account for any omitted variables that could be driving the correlations, even though we are cautious not to interpret these correlations as causal effects.

[9] Given the population-level data and no causal parameter to estimate, we needn't cluster the standard errors per se (Abadie, Athey, Imbens, and Wooldridge 2022). Clustering increases the size of the standard errors, which suggests that the distance patterns we observe is not sensitive to clustering.

Fact 2: Fewer Mid-Sized Farmers Near Towns

A second fact that we observe is that there are fewer mid-sized farmers (individuals or households with 5-8 acres of landholding) closer to towns relative to small and large farmers. We estimate Equation 1 using the ratio of the number of mid-sized farmers to the number of small and large farmers, respectively as the dependent variable. Both specifications yield a negative correlation between the share of mid-sized farmers relative to the share of small and large farmers, respectively, and village's distance from its nearest town (Panel B Figure 3 and Columns 2-3 Table 2). Parametric estimates imply that every 10 km increase in distance from town is associated with a 0.65 percentage point increase in the share of mid-sized farmers relative to small farmers and nearly 16 percentage points increase relative to large farmers in a village.

Fact 3: Distance Correlation Steeper Near Larger Towns

The third fact is about the intensity of urban influence. We observe a higher urban-distance gradient of rural landholding inequality surrounding bigger towns by town population. That is, villages in the vicinity of top population quintile towns have a much higher landholding inequality relative to villages in the similar vicinity of bottom quintile towns. We modify the distance specification above to include an interaction between the distance bins and the nearest town population as shown in Equation 2 below.

$$\begin{aligned} Gini_{v(t)m} &= \delta_t + \delta_m + \sum_{j \neq 35-40 \text{ km}} \gamma_j D_{jv(t)m} \text{ x Town Size}_t + \sum_{j \neq 35-40 \text{ km}} \beta_j D_{jv(t)m} \\ &+ \epsilon_{v(t)m} \end{aligned} \tag{2}$$

Town Size$_t$ is the census population count of the nearest statutory town t. We also estimate a slightly different specification by including an interaction term based on town population quintiles rather than population counts. Furthermore, we generate a dummy variable that takes a value 1 when the nearest town is in the top quintile of town population distribution and compare the urban-distance gradient with towns in the bottom quintile such that the reported coefficients reflect relative urban-distance gradient in village-level gini in regions around large towns relative to those around smaller towns.

Panel C Figure 3 presents these results on urban-distance gradients of surrounding rural land inequality by town population quintiles. We note that the urban distance gradient of rural land inequality is steeper closer to larger (top quintile) towns relative to the distance gradient in the vicinity of smaller (bottom quintile) towns, which is statistically and economically significant. We also examine this interaction using distance and town population as continuous variables (Columns 4 and 5 Table 2). Every 100,000 additional population in towns is associated with an additional 0.02 σ units higher rural land inequality Gini index on average, which decreases by a tenth of this effect size for every 10 km increase in distance to town. So, a town with one million additional population is associated with an average rural landholding Gini 0.2 σ units higher relative to smaller towns, and this decreases by a tenth of this effect size for every 10 km increase in distance.[10]

3.1 Robustness of Observed Empirical Patterns

These observations are robust to a variety of tests. First, the same patterns are observed using SECC - a completely different dataset of household surveys administered by the national statistical office (see Figure A2). This, census-based dataset, also helps account for land parcels registered under different household members to calculate land inequality at the farm household-level. Further, the Gini coefficient measured using SECC data also captures landless households within a village. This test addresses two important concerns regarding the construction of the Gini-index: (a) measurement error due to not including landless households, and (b) measurement error due to using landholding ownership data at an individual-level rather than at a household-level (particularly if farmstead land is registered under different household members).

Second, we notice a similar urban-distance gradient in the extent of landless households in the villages (see Figure A3), suggesting that the landholding inequality patterns also reveal insights about the extent of landlessness and its correlation with ur-

[10]Some of the large urban metropolitan areas in India are over 10 million in population. For example, Delhi metropolitan area had a population of around 16 million as of 2011 census. Mumbai had over 18 million. The 2011 population census classifies over 52 metropolitan areas with population over a million.

banization. Further, the patterns are also visible when using first and second moments of landholding size distribution at the village-level instead of using the Gini-index.

Third, the empirical patterns remain unaffected if we are village-agnostic, i.e. if we generate landholding distribution by pooling all farmers in 5 km distance bins from their nearest town (Figure A4) or after dropping villages surrounding large metropolitan cities, as shown in Figure A5. This ensures that the observed patterns are not driven by endogenous changes to village boundaries or by outlier urban areas.

Fourth, the distance correlation is also evident using IHDS, an all-India representative household panel data (see Figure A6). We use both the available waves of data within IHDS - 2005 and 2012, respectively - to examine the correlations. Columns 1-2 Table A1 presents this relationship. As noted, village-level landholding Gini and farm size compositions are negatively correlated with distance to nearest town across both survey waves, suggesting that the observed facts are not specific to our study state or a specific time but is more likely a general phenomena as observed across a large country like India.[11]

Fifth, the composition results (Fact 2) are also robust to alternate landholding size partitions into small, mid, and large. That is, our patterns are not sensitive to our preferred definition of small as those with landholding under 5 acres, mid-sized as 5-8 acres, and large as over 8 acres in landholding. In Figure A7, we show that the farm-size-distance correlations remain consistent even when using different cut-offs for defining small, mid, large landholding sizes. We also find that the composition of farm size patterns remain evident even when using percentage of total farmers belonging to different farm-size bins rather than using ratios of different farm-size bins (see Figure A8).

Lastly, we try to address biases that could arise due to omitted variables correlated with both village-level landholding distributions and distance to towns. In Table A2, we examine if any of the village-level amenities vary by distance to town, which could be driving the landholding distribution patterns. For example, crop suitability patterns could lead some villages (closer to urban markets) to consolidate land for better

[11]Since the IHDS sample is not representative of the village-level land size distribution (specifically, this dataset over-represents larger land-owning households relative to the population-level distribution), our test of farm-size correlation is weak. Further, this dataset also does not have identifying information on the nearest town to be able to obtain the town population to test the town-size correlations.

productivity. We summarize the mean and standard deviation of these variables in the data to help interpret the size of the reported coefficient and confidence intervals. We find a precise 0, suggesting that the observed correlation is not driven by crop suitability, amenities, or local non-agricultural activities and employment.

However, there could still be other unobserved omitted variables such as village-level population or other amenities such as banks, schools, or access to good roads. We account for village-level population when comparing farm size composition by distance to town (it appears on the LHS denominator in Figure A8). We focus on amenities such as banks, schools, and access to good roads as potential alternate explanations later in the paper. While some of these factors explain one or two observed facts, none of them explain all three facts together.

A Note on Potential Endogeneity of Distance to Town We want to be clear that our intention with the robustness tests is not to infer a causal interpretation of spatial distance between rural areas and towns as driving the observed variation in landholding distributions. To further allay concerns about omitted variables that could be biasing the correlation coefficients of the distance bins, we use access to national and state-level highway as instruments to distance to town, after conditioning on sub-district and nearest town fixed effects. National highways are constructed by the National Highway Authority of India (NHAI), a federal department, using federally allocated funds. Similarly, state public works departments construct state highways using state-level funds. Highways connect major urban and trading centers. We carry out over-identification tests by using the presence of both national and state highways as instruments to distance between a village and its nearest town and fail to reject the null. The first stage is strong - instruments are strongly and negatively correlated with distance (see Table A3). With the assumption that the changes in landholding distributions occur only through changes in distance to urban areas due to highway connectivity (which proxy access to markets), then it is plausible that the observed correlations are less likely to be driven by omitted confounders.[12]

[12]There are a few considerations to take into account when thinking about potential omitted variables: (a) we examine the patterns using most recent land holding data available, (b) the persistence in land holding patterns and inequality vis-a-vis the timing of urban settlement and growth is likely to be broken by a series of land reforms implemented in the 1960s and 1970s, and (c) infrastructure (road) construction and presence of statutory towns predate the landholding data. Thus, all the RHS variables are predetermined relative to the timing of measuring LHS variables (landholding

Our objective behind presenting these facts is to develop a simple theory of structural transformation where exit from agriculture is selective based on land size available for agricultural production and urban opportunity cost. We describe this in detail in the next few sections of the paper.

4 A Model of Selective Exit from Agriculture

We build a parsimonious model that can describe the observed empirical patterns. Our model posits that urbanization encourages mid-sized farmers to exit cultivation for better wage opportunities in urban areas relative to their returns to agricultural production with their existing land endowment. At the same time, intra-village land distribution is constantly changing due to the division of land holdings for the use of farmers' descendants.

Our multi-period dynamic model starts with a farmer-level optimization problem which involves maximizing a value function that incorporates the trade-off between returns from investing in agriculture given their current land endowment, a future bequeathing of the land, returns from migrating to the nearest urban area (which depend on the characteristics of that town, i.e. urban wages, which in turn, are a function of town population), and the cost of migration (which depends on distance to the town). Next, we aggregate individual farmer decisions over the distribution of farmers within a village to show how aggregating these individual decisions affects village-level landholding inequality Gini index.

We make three simple assumptions: First, agricultural productivity follows a U-shape function with respect to farm-size following recent literature (Cornia 1985; Barrett 1996; Gaurav and Mishra 2015; Foster and Rosenzweig 2022) that has demonstrated such non-linearities in productivity in various contexts. This implies higher productivity among small and large farmers but lower productivity among mid-sized farmers. These non-linearities can occur due to scale economies either on the intensive margin (intensive use of family labor and minimizing span of control problem as discussed in Caunedo and Kala 2021) or on the extensive margin through mechanization that is feasible only for large farm sizes (Foster and Rosenzweig 2022). We examine correlations between mechanization/farm productivity with distance to town

gini or composition ratios).

as well as village-level Gini index later in the paper to provide further evidence of this assumption.

Second, the decision of land holdings is affected by household composition. In particular, we assume that there is some probability of dividing land among farmers' children[13]. This probability is monotonically increasing in farmer's age and the number of their children. This assumption takes into account the standard life cycle changes that predominantly includes transfer of wealth and assets from older to younger generations.

Our third assumption is that towns offer an opportunity cost to farmers through the average wage offered in the town. This average wage is increasing with town population, but the perceived wage at the village is a decreasing function of the distance to town. This last feature captures the increasing cost of migration when distance to town increases. Alternatively, this could also be interpreted as costs imposed on transportation of agricultural goods to urban markets. Either way, these costs appear only as a function of distance to town and town characteristics and is independent of landholding size since all farmers in the same village experience the same urban opportunities and costs.

4.1 Farmer-level Optimization

We first begin with farmer-level optimization. Consider a representative farmer, who lives for infinite periods, starting at age A, number of children $N \geq 0$, and a landholding size L. At any point in time, the farmer faces the decision of continuing with agriculture by optimally adjusting their land endowment, where land has an exogenous price P_L (relative to the farmer's problem), or permanently migrating to the nearest town. A farmer continuing with agriculture would divide the land among his children with probability λ. This probability is a function of the age of the farmer and the number of children. The land is divided equally among children and the farmer keeps an equal portion for themself. If the farmer doesn't have any children (N=0), the probability of dividing the land is 0.

As discussed above, agricultural productivity follows a U-shape function in land size following the recent literature that has found that agricultural productivity is not

[13]We also assume that the infinitely-lived farmer keeps a fraction of the land for their own use and also include the possibility of no bequest.

monotonic with respect to land size (Cornia 1985; Barrett 1996; Gaurav and Mishra 2015; Foster and Rosenzweig 2022). We calculate agricultural income by scaling productivity by corresponding landholding size.

Farmer's optimization problem when their income is mainly from agricultural production is as follows:

$$
\begin{aligned}
V(L, A, N, P_L) &= \max_I u(c) + \beta[\lambda(A, N)V(\frac{L'}{N+1}, A+1, 0, P_L) \\
&\quad +(1-\lambda(A, N))V(L', A+1, N, P_L)] \\
c + P_L I &\leq f(L) \\
L' &= L + I \\
f(L) &= L(\alpha_1 L^{\alpha_2} + \alpha_3 L + \alpha_4) \\
\lambda(A, N) &= \begin{cases} 0 & \text{if } N = 0 \\ g(A, N) & \text{if } N > 0 \end{cases}
\end{aligned} \tag{3}
$$

where I is the extent of new land purchased, and c is current consumption. This set up implies that the farmer's value function incorporates current period consumption and the discounted present value of future returns from cultivating land. This future return incorporates an optimal time for adjusting their land endowment to L' to scale production as well as consider implications on landholding size due to bequest. That is, the probability of holding a certain land endowment decreases over time due to potential subdivision of land from bequest. This probability that land is subdivided is $g(A, N)$, which is an increasing function of A and N, and satisfies

$$\lim_{N \to 0} g(A, N) = 0 \quad \wedge \quad \lim_{A \to \infty} g(A, N) = 1$$

The function $f(L)$ is total agricultural income which is agricultural productivity scaled by landholding size. We introduce non-linearities in the productivity term, which is composed of three elements. First, we have a non-linear gain in production, captured by the parameters $\alpha_1 > 1$ and $\alpha_2 > 1$. Second, we introduce span of control type negative production term that varies linearly with landholding size (captured by parameter $\alpha_3 < 0$). Third, we include fixed production gains from one time capital investments (like machinery or irrigation equipment) that does not vary with landholding size (α_4).

The farmer compares returns from their agricultural enterprise with the returns from migrating to the nearest town of size s at distance d after selling their land, which earns them $P_L L$. We denote $M(w, L, P_L)$ as the value of permanently migrating to the nearest town (and starting to work in the next period) with perceived urban wage $w(s, d)$, or equivalently

$$M(w, L, P_L) = u(P_L L) + \sum_{t=1} \beta^t u(w) \tag{4}$$

Therefore, the full problem for a farmer with land L, age A, number of children N, facing current land prices P_L, who lives near a town that offers a perceived urban wage $w(s, d)$ (after accounting for the cost of migration as a function of distance) is

$$F(L, A, N, w, P_L) \quad = \quad \max\{V(L, A, N, P_L), M(w, L, P_L)\} \tag{5}$$

This set up also implies that the dynamics of the distribution of land-holdings would depend on both the initial land-holding distribution, and the distributions of farmer demographics - age and number of children. In our simulations, we create heterogeneity within and across villages using random initial distribution of land (that is right-skewed), age and number of children, and we study how this joint distribution evolves over time.[14]

Finally, we assume that the farmer takes current urban wage as given even though migration from other nearby villages would likely affect urban wages in equilibrium. We believe that this is a plausible assumption since in order to forecast changes in urban wages due to migration of other farmers from other villages in the vicinity, a farmer would need a lot information - joint distributions of landholding sizes, farmer age, and their number of children in each village within the influence of the urban area. Relaxing this assumption would mean that the farmer would account for the expected value of future urban wage rather than consider a deterministic current wage.

[14]In this set up, the distribution of land is changing while converging to the long-run steady state. We can alter this initial distribution to reflect different policies to shed light on how the subsequent dynamic would emerge.

4.2 Generating Village-Level Land Inequality

We aggregate the optimal response for each farmer at the village-level, subject to a constraint of fixed supply of land within a village, to compute the village-level landholding Gini-index over time. We do this through a simulation exercise rather than analytically computing a steady state Gini-index, given the challenge of the large set of changing parameters that make such exercise intractable.

We start with 1500 villages with 550 farmers per village. We assign initial values for land, age and number of children for each farmer, with the village-level distributions following random normal distribution with a right skew for land. The villages are scattered around 150 towns with random normal distribution of population (with parameters $\mu = 100,000$ and $\sigma = 32,000$) and distance (with parameters $\mu = 45$ km and $\sigma = 30$ km), similar to parameters observed in data. This ensures that the initial conditions imply no correlation between village-level Gini-index and the distance between the village and its nearest town.

We calibrate our agricultural income function $f(L)$ using parameter values $\alpha_1 = 1.15, \alpha_2 = 2.52, \alpha_3 = -97, \alpha_4 = 600$, such that the landholding size corresponding to minimum productivity per acre is between 8 and 10 acres, as estimated by (Foster and Rosenzweig 2022) in a similar context.[15]

We assume standard functional forms for utility (log) and perceived wages as $w = (s)^{0.5}/d^{0.18}$, where s is the nearest town population and d is the distance between the village and the town. We assume the relative price of land within a village to be an exogenous parameter. This is because land markets are extremely thin within villages (Foster and Rosenzweig 2022) and around 95% of the households report acquiring their land from inheritance in India (Desai, Vanneman, and National Council of Applied Economic Research 2005). Finally, we assume our discount factor, $\beta = 0.95$.

[15] Note that our definition of mid-sized is 5-8 acres, so the model does not mechanically generate the minima of the value function at 5-8 acres. In fact, we calibrate our model such that the minima corresponds to 8-10 acres as found in the literature. The shape of the value function without any constraints in Equation 3 is actually monotonic. The non-linearities in the value function mainly stem from the constraints (bequest and inelastic land supply) rather than from the assumption of the functional form of $f(.)$.

4.3 Model Simulations

Figure 4 shows the distribution of landholding Gini-index, its correlation with the distance to the nearest town, and the relative distance gradient by town population quintiles using simulated data following the data generating process as per our model. Panel A shows the distance correlation subsequent to farmer-level optimization given their initial endowments of land, age and number of children. Panel B shows farm-size composition relationship and Panel C shows the town-size correlation.

These patterns are similar to what we observe in the data (see Figure A9 where we overlay the distance gradients generated by the model relative to data, and see Figure A10 for the model fit). This suggests that the observed patterns in landholding distributions could be driven by either or both of the following: (a) mid-sized farming households sell their agricultural land and migrate to the nearest town, and/or (b) downsize or increase their landholding so that they either become small farmers or large farmers themselves to avoid farming a sub-optimal size of land. Interestingly, bequest motives imply a continually changing landholding distribution. As farmers' age and the number of their children are independent of farmer's initial landholding endowment as well as are independent of their village's distance from its nearest town, any movements in landholding distributions from bequest are unlikely to produce the observed empirical patterns that arise from differences in (urban) opportunity costs of agricultural production within a village.

In our simulations, we find that the urban pull is small for villages far from a town, and we observe fewer farmers leaving agriculture including those with mid-sized landholding. We observe more farmers increasing their land holdings under bequest motives, but because land supply is fixed at the village level, the supply of land for purchase is very limited. For villages closer to a town, the urban pull is large, and we observe more mid-sized farmers exiting agriculture due to relatively lower returns from agriculture. Farmers also seek to increase their landholding size both for productivity reasons and for bequest motives in villages close to town as they do in villages farther away from towns. However, due to land supply constraint, only very few small and mid-sized farmers expand their landholdings to become a large farmer. Only those who have fewer children, who are young, and who live in villages with less binding land supply constraint are able to transition to becoming a large landholder. On the other hand, the few large farmers become even larger, particularly when the land supply constraints are relaxed due to the exit of mid-sized farmers closer to towns.

4.4 Model Assumptions and Caveats

We verify and find support for the main model assumptions in our data. Note from Section 2.7 that only a small fraction (under 10%) of farmers are mid-sized. We assume that selective exit of *some* of these farmers is plausibly unlikely to dramatically relax the land supply constraints to alter the land price or change the demographics of the nearby towns. We also find support for non-linearities in agricultural productivity (yield) by land size in our context (see Figure A11). Finally, we find no systematic differences in household demographics such as the number of children or probability of inheritance by distance to town and its interaction with initial landholding size (see Table A4).

It is important to note that our model does not need to assume credit market frictions that could prevent smallholders from taking advantage of economies of scale or for mid-sized farmers from far away villages to borrow to fund the cost of migration. In fact, we will show later that our model is also consistent with such frictions, although these are not required to generate the key empirical facts. We show that non-linearities in agricultural production function and frictions in the land markets are sufficient to generate the patterns we observe in the data.[16]

Lastly, we caveat that our model is based on partial equilibrium conditions arising from farmers' optimization problem and we do not explicitly model general equilibrium, steady-state conditions, which would internalize the perceived urban wage w due to changing town size s, and changing relative price of land P_L. There are two key reasons for this. First, modeling general equilibrium in this context is computationally very complex requiring distributions of land, age, and the number of children for each village, town size distribution, and urban wages in equilibrium, in addition

[16]One way in which land market friction could be addressed is if farmers owning fewer acres could rent land for cultivation from larger farmers. However, tenant farming is rare in the state we study. In general, tenant farming is rare in states that had *ryotwari* system of land revenue during British India (Banerjee and Iyer 2005). On the other hand, this is common in erstwhile *zamindari* areas, where subsequent tenancy reforms led to improved tenure security for tenant farmers through land redistribution (for e.g., Operation Barga of West Bengal discussed in Banerjee, Gertler, and Ghatak 2002). Further, land rental markets also have severe frictions as documented in Bolhuis, Rachapalli, and Restuccia (2021).

to endogenizing the relative price of land. Second and more importantly, the general equilibrium response could require a long time horizon to emerge given market frictions. Since our empirical patterns are based on a cross-section of rural landholding inequality in current time and not how they evolved over time, our model incorporates the path to equilibrium rather than the steady-state itself. We examine the rich implications of our model in our data that we discuss below.

5 Model Implications

5.1 Land Consolidation

A plausible implication of our model is that large farmers could buy agricultural land sold by mid-sized farmers to increase their scale of production even more. Therefore, landholding sizes at the top of the distribution should be relatively large compared to landholding sizes at lower percentiles in villages close to towns compared to those farther away. Indeed, we note this in Panel A in Figure 5 where the ratio of farm size at the 99th percentile relative to farm size at the 25th percentile shows a negative urban-distance gradient. The figure suggests that farms at the 99th percentile of size distribution is nearly 20 times as large as farms at the 25th percentile of the distribution within a village.

5.2 Land-size Transitions By Household Landholding

Another related implication is that mid-sized farmers are the main economic agents for whom the opportunity costs bind. They react either by downsizing or upsizing their landholding, or through completely exiting agriculture and migrating to urban areas to earn wages from the non-agricultural sector. Testing this is hard mainly due to data challenges: (a) panel data on the entirety of land holding distribution is hard to come by, (b) data on migration at the household or individual-level is hard to come by; even aggregate data on migration by landholding distribution is unavailable.

We overcome these challenges by exploiting an all-India representative household panel data from IHDS (2005, 2012) to examine: (a) whether there is differential attrition by ex-ante landholding size in the subsequent survey rounds to proxy for potential migration, and (b) compute a transition matrix of landholding between the two survey rounds, both based on the households' initial landholding size. Analysis of attrition investigates whether the entire household is missing in the second round, which is a likely (although not a conclusive) indicator of migration. The transition matrix also allows us to examine whether mid-sized farmers downsize or upsize over time.

Table 3 presents the landholding transition matrix by initial landholding distribution between the two survey rounds of IHDS panel. Over 92% of the initial small farm households (< 5 acres) continue to remain small in 2012 and only a small fraction increase their landholding. Among mid-sized farms in 2005, 42% become small and about 28% become large in 2012 and only 30% remain mid-sized. Among large farm households, the majority remain large although a substantial fraction downsize. Importantly, a greater share among these erstwhile large and mid-sized farms become small farms compared to being mid-sized in 2012. Perhaps a reason for the downsizing we observe could be due to lifecycle events such as inheritance, but we show that such events are not systematically correlated with landholding distributions and distance to nearest town to generate the empirical patterns we observe.

Table 4 documents rural household land transition patterns between the two survey waves in 2005 and 2012, respectively, recording correlations between landholding sizes in 2012 and distance to town interacted with baseline landholding sizes (i.e., size in 2005) as follows:

$$\begin{aligned} Size2012_{hv(t)} &= \delta_v + \beta_1 Dist_{hv(t)} \text{ x Large2005}_{hv(t)} + \beta_2 Dist_{hv(t)} \text{ x Small2005}_{hv(t)} \\ &+ \alpha_1 \text{Large2005}_{hv(t)} + \alpha_2 \text{Small2005}_{hv(t)} + \alpha_3 Dist_{hv(t)} + \varepsilon_{hv(t)} \qquad (6) \end{aligned}$$

The leave-out group is households h with mid-sized landholding in 2005 in village v at a specific distance from its closest town t. We structure the two-period data into wide format, with corresponding variables for each survey wave, to study transition between one period to the other. To account for time invariant, unobserved confounders between village-level landholding distribution and distance to town, we include village fixed effect.

This exercise has two takeaways. First, we observe non-linearity in household attrition by landholding size that varies with distance to town. Overall attrition is around 23%, which varies both by distance to town and farmsize bins. Large and small landholders are 1-1.5 percentage points more likely to be absent relative to midsized landholders for every 10 kms increase in distance to closest town. Conversely, mid-sized landholders are more likely to be absent during the follow-up survey in villages closer to towns. Second, both small and large landholding households are more likely to converge towards becoming mid-sized with increasing distance from their nearest town (Column 2). On the other hand, households in villages close to towns are less likely to converge to mid sized landholding, with some suggestive evidence that smallholders even become landless. These are consistent with the increase landholding inequality we observe close to towns.

5.3 Plausible Welfare Implications

While commenting on the general equilibrium welfare implications of rural land inequality and proximity to towns is outside the scope of this paper, empirically, we find strong positive associations with increased mechanization and village-level agricultural productivity closer to towns (see Panels B and C Figure 5).[17]

There are likely multiple, potentially interacting, economic mechanisms behind these observed welfare effects. First, the exit of mid-sized farmers may leave behind more productive farms (i.e. small and large), which increases overall productivity at the village level. Second, land consolidation by large farmers, for which we find suggestive evidence, could lead to increased mechanization that support increasing returns to scale. Finally, this observed correlation is also consistent with the plausible urban influence through access to markets and higher agricultural wages that increases overall income and consumption levels in the surrounding villages (we find negative urban-distance correlation with agricultural wages in Columns 2-3 Table A1).

5.4 Path to Equilibrium

The model suggests a dynamic process in the evolution of rural landholding inequality over time as a function of distance to urban centers. Instead of modeling the steady state and general equilibrium conditions, we model partial equilibrium conditions as a

[17] We measured agricultural productivity as average yield achievement ratio defined by FAO as the ratio of actual yields to potential yields.

path to general equilibrium, focusing on farmer-level optimization over their choice of landholding size. This optimization process could lead to varying land inequality due to urban opportunity that dominates only for a certain type of farmers. One of the key reasons for our modeling choice is that what we observe in cross-section may not necessarily be a steady state. Indeed as Bardhan, Luca, Mookherjee, and Pino (2014) document, landholding inequality is likely continually evolving due to subdivision of land due to bequest, land market transactions, and demographic dynamics, including migration, differential fertility, which can affect landholding distribution at any given point in time.

6 Discussion

Policy recommendations to address wealth inequality, particularly land, is hard because it requires anticipating many dynamic variables including family life-cycle changes, understanding the current landholding distribution, and reducing frictions across multiple markets (land, labor, and capital) that could interact with these dynamics. Given the history of land redistribution and land ceiling legislations in the past, further land redistribution may be politically infeasible and economically unclear as the best policy response.[18] Land gets fragmented over time due to lifecycle events such as inheritance, resulting in more small farms that can't take advantage of economies of scale. On the other hand, land consolidation policies without taking into account the constraints facing small farmers in switching to better opportunities (such as migrating to urban areas to earn better wages) would not be welfare enhancing.

In this paper, we bring to fore some key empirical facts about rural land inequality that is shaped by who remains in agriculture. This choice is based on the relative attractiveness of farming enterprise with respect to urban wage income and whether or not binding constraints keep households tied to sub-scale agriculture. What we tried to demonstrate in this paper is present a simple economic framework with minimal assumptions that could explain the observed patterns in the data. It is certainly plausible that more than one specific mechanism could be at play. We tested a variety

[18]There are very few 100 acre farms in the state. There are no longer large farms and the current "large farms" are only 12 acres on average, with the largest farm being no larger than 60 acres.

of alternate explanations, which are unable to explain all the observed patterns. We present the results from examining some of these explanations below.

6.1 Role of Financial Constraints

A leading alternate mechanism behind the rural land inequality and urban distance patterns could be financial constraints. Financial constraints could prevent small or mid-sized farmers from acquiring more land to reach economies of scale and optimal productivity at a larger farm size. One metric to determine presence of financial constraints is whether or not a village has a bank branch (including regular commercial/retail banks, cooperative banks, and agricultural credit societies) that increases access to credit even among small farming households.

Agriculture sector loans, particularly targeting small and marginal farmers, are considered priority sector as per national policy and banks particularly target such borrowers to meet their lending quota under the policy. Further, the state and national governments typically provide repayment support by underwriting such loans through budgetary provisions to compensate the banks in the event of non-payment (Phadnis and Gupta 2015). Access to finance can enable some farmers to acquire more farmland. This should lower inequality if both mid-sized and small farmers become large or exit entirely to migrate, respectively, generating a more homogeneous land-holding distribution among the remaining farmers.

An empirical test of this hypothesis from our model would be to examine how inequality evolves over time in the presence of financial constraints (given the dynamic nature of the farmer optimization problem). Since we only have cross-sectional data, we test the hypothesis using variation in the nearest town population to capture the evolution of rural land inequality with increasing urbanization. We estimate Equation 2 separately for villages with banks and those without banks.

Figure A13 juxtaposes the town-size and distance correlation with and without financial constraints, showing a clear distance-town size interaction gradient in the presence of financial constraints (no banks) and no such gradient in its absence (villages with banks). We notice this when using village-level Gini-index, farm-size compositions, or the ratio of farm sizes across the distribution. This is not inconsistent with our model which does not impose any assumptions on financial constraints. Indeed, additional constraints on access to capital or other factors of production within agriculture could further exacerbate inequality.

6.2 Alternate Explanations

Household composition: We examine multiple other explanations for the observed patterns. First, and most importantly, can the observed patterns be explained by family-size-choice based on landholding size and distance to town? For example, if mid-sized farmers choose to have more children near towns in order to have more family labor to replace costly hired labor, then the observed patterns could be explained by bequest implications of this initial condition. This would violate one of our assumptions where we take the number of children as exogenous to the model. However, Table A4 shows no differential patterns in the share of children (<15 years of age), composition over time (such as household splits), or land inheritance patterns by initial landhold size and distance to the nearest town.

Skilling: Second, mid-sized farmers may be differentially skilled relative to small farmers by distance to town, and thus more valued in urban areas. This explanation is based on the role of skill-based rural to urban migration in generating income inequality following rural-urban wage gap as discussed in Young (2013). This explanation suggests that the mid-sized farm households exit agriculture to invest in human capital due to differential return to skill rather than optimal farm size. However, we do not find strong distance correlation in differential skilling by household farm size. Specifically, we do not find that mid-sized farm households are: (a) relatively more educated (years of schooling) relative to small or large farm households with distance to the nearest town (Column 1 Table A5), (b) any more likely to have salaried employment that varies with distance (Column 2 Table A5), (c) any more likely to have a higher share of farm workers among family members that varies with distance (Column 3 Table A5), and (d) have higher total income (as a proxy for unobserved skills) relative to small or large farmers closer to towns (Column 4 Table A5, which shows the opposite correlation - mid sized farming households have higher household income farther away from towns). Further, we examine whether the patterns correlate with the presence of secondary school in a village and find no differential correlation between access to higher levels of schooling and rural land inequality (see Panel A Figure A14, where both types of villages - those with and those without secondary schools - display similar distance correlation in the presence of varying extent of urban pull).

Crop-choice: Third, we examine whether proximity to towns influences crop-choice, where certain crops are more suitable for increasing return to scale. The dominant crops in the region we study are paddy and cotton. We note no differential pattern in crop-choice by distance to town with varying degrees of urban influence (see Panel B Figure A14 and Column 4 Table A1).[19] This suggests that crop-choice is plausibly driven by the large share of small, subsistence farmers (rather than crop-choice driving the choice of farm size).

Market Access: Finally, having a good road - defined as national, state, or district highways or even rural roads as examined by Asher and Novosad (2020) - through a village could bring about changes in the local village economy and reduce migration costs. Therefore, placement of roads could also drive the observed distance correlations between rural land inequality and urbanization. This explanation does not contradict our model but rather, complements by providing potential variation in the cost of migration, which can be modeled as a function of road quality-adjusted distance to town. While there is some evidence that inequality is worse when rural roads are of better quality (by lowering the cost of migration), we cannot reject the similarity in distance gradient by access to quality road (Panel C Figure A14).

7 Conclusion

This paper is one of the first attempts at documenting patterns in land inequality in rural India using the universe of land records in a large state as well as validating some of the key implications of a parsimonious model using multiple data sources.

We find a very clear spatial pattern in rural land inequality based on the distance between the rural area and its nearest town. These patterns are consistent with a

[19] Vegetables and perishables, while important from urban consumption point of view, are not among the main three crops reported in the census data for our context and therefore, unlikely to influence the choice of farm-size.

model where mid-sized landholders find it more attractive to exit agriculture near urban areas. While small and large landholders do not face this trade-off, small landholders could face constraints such as access to finance to acquire more land to exploit economies of scale in agricultural production enjoyed by large landholders.

Our model, along with the empirical observations, point to the role of urbanization and non-linearities in farm productivity by farm size in creating asset-size traps and incomplete structural transformation where only a specific group is able to take advantage of urban opportunities, leaving many to rely on subsistence agriculture.

We are cautious about making any specific policy recommendations surrounding land policies. However, policies that aim at reducing market frictions, such as credit, could help in mitigating some of the inequality in asset distributions. We believe that this paper lays down an important research agenda to study additional questions relating landholding-based poverty traps and structural transformation using different data sources, research designs, and structural modeling tools.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2022): "When Should You Adjust Standard Errors for Clustering?*," *The Quarterly Journal of Economics*, 138(1), 1–35.

Adamopoulos, T., and D. Restuccia (2020): "Land Reform and Productivity: A Quantitative Analysis with Micro Data," *American Economic Journal: Macroeconomics*, 12(3), 1–39.

Asher, S., T. Lunt, R. Matsuura, and P. Novosad (2021): "Development Research at High Geographic Resolution: An Analysis of Night Lights, Firms, and Poverty in India using the SHRUG Open Data Platform," World Bank Economic Review.

Asher, S., and P. Novosad (2020): "Rural Roads and Local Economic Development," *American Economic Review*, 110(3), 797–823.

Balboni, C., O. Bandiera, R. Burgess, M. Ghatak, and A. Heil (2021): "Why Do People Stay Poor?*," *The Quarterly Journal of Economics*, 137(2), 785–844.

Banerjee, A., P. Gertler, and M. Ghatak (2002): "Empowerment and Efficiency: Tenancy Reform in West Bengal," *Journal of Political Economy*, 110(2), 239–280.

Banerjee, A., and L. Iyer (2005): "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India," *American Economic Review*, 95(4), 1190–1213.

Banerjee, A. V., and A. F. Newman (1993): "Occupational Choice and the Process of Development," *Journal of Political Economy*, 101(2), 274–298.

Banerjee, A. V., and A. F. Newman (1994): "Poverty, incentives, and development," *The American Economic Review*, 84(2), 211–215.

Bardhan, P., M. Luca, D. Mookherjee, and F. Pino (2014): "Evolution of land distribution in West Bengal 1967–2004: Role of land reform and demographic changes," *Journal of Development Economics*, 110, 171–190.

Barrett, C. B. (1996): "On price risk and the inverse farm size-productivity relationship," *Journal of Development Economics*, 51(2), 193–215.

Bazzi, S. (2017): "Wealth Heterogeneity and the Income Elasticity of Migration," *American Economic Journal: Applied Economics*, 9(2), 219–55.

Bazzi, S., A. Gaduh, A. D. Rothenberg, and M. Wong (2016): "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *American Economic Review*, 106(9), 2658–98.

Bolhuis, M. A., S. R. Rachapalli, and D. Restuccia (2021): "Misallocation in Indian Agriculture," Discussion Paper w29363, National Bureau of Economic Research.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014): "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh," *Econometrica*, 82(5), 1671–1748.

Bryan, G., and M. Morten (2019): "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia," *Journal of Political Economy*, 127(5), 2229–2268.

Burchardi, K. B., S. Gulesci, B. Lerva, and M. Sulaiman (2019): "Moral Hazard: Experimental Evidence from Tenancy Contracts*," *The Quarterly Journal of Economics*, 134(1), 281–347.

Caunedo, J., and N. Kala (2021): "Mechanizing Agriculture," Working Paper 29061, National Bureau of Economic Research.

Cornia, G. A. (1985): "Farm size, land yields and the agricultural production function: An analysis for fifteen developing countries," *World development*, 13(4), 513–534.

de Janvry, A., K. Emerick, M. Gonzalez-Navarro, and E. Sadoulet (2015): "Delinking Land Rights from Land Use: Certification and Migration in Mexico," *American Economic Review*, 105(10), 3125–49.

Desai, S., R. Vanneman, and N. D. National Council of Applied Economic Research (2005): "India Human Development Survey (IHDS)," .

——— (2012): "India Human Development Survey (IHDS)," .

Foster, A. D., and M. R. Rosenzweig (2022): "Are There Too Many Farms in the World? Labor Market Transaction Costs, Machine Capacities, and Optimal Farm Size," *Journal of Political Economy*, 130(3), 636–680.

Gaurav, S., and S. Mishra (2015): "Farm size and returns to cultivation in India: Revisiting an old debate," *Oxford Development Studies*, 43(2), 165–193.

Glaeser, E., and J. V. Henderson (2017): "Urban economics for the developing World: An introduction," *Journal of Urban Economics*, 98, 1–5.

Gollin, D. (2019): "Farm size and productivity: lessons from recent literature," *IFAD RESEARCH SERIES 34*, (281567).

Gollin, D., D. Lagakos, and M. E. Waugh (2014): "Agricultural Productivity Differences across Countries," *American Economic Review*, 104(5), 165–170.

Gollin, D., and C. Udry (2021): "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture," *Journal of Political Economy*, 129(1), 1–80, Publisher: The University of Chicago Press.

Heise, S., and T. Porzio (2021): "The Aggregate and Distributional Effects of Spatial Frictions," *Working Paper*.

Henderson, J. V., and S. Kriticos (2018): "The Development of the African System of Cities," *Annual Review of Economics*, 10(1), 287–314.

Henderson, J. V., and M. A. Turner (2020): "Urbanization in the Developing World: Too Early or Too Slow?," *Journal of Economic Perspectives*, 34(3), 150–73.

Hsieh, C.-T., and E. Moretti (2019): "Housing Constraints and Spatial Misallocation," *American Economic Journal: Macroeconomics*, 11(2), 1–39.

Jayachandran, S. (2006): "Selling Labor Low: Wage Responses to Productivity Shocks in Developing Countries," *Journal of Political Economy*, 114(3), 538–575.

Kahn, M. E., W. Sun, J. Wu, and S. Zheng (2018): "The Revealed Preference of the Chinese Communist Party Leadership: Investing in Local Economic Development versus Rewarding Social Connections," NBER Working Papers 24457, National Bureau of Economic Research, Inc.

Kraay, A., and D. McKenzie (2014): "Do poverty traps exist? Assessing the evidence," *Journal of Economic Perspectives*, 28(3), 127–48.

Lagakos, D. (2020): "Urban-rural gaps in the developing world: Does internal migration offer opportunities?," *Journal of Economic perspectives*, 34(3), 174–92.

Lagakos, D., A. M. Mobarak, and M. E. Waugh (2023): "The Welfare Effects of Encouraging Rural–Urban Migration," *Econometrica*, 91(3), 803–837.

Lobell, D. B. (2013): "The use of satellite data for crop yield gap analysis," *Field Crops Research*, 143, 56–64.

Madhok, R., F. Noack, A. M. Mobarak, and O. Deschenes (2025): "Internal Migration and the Spatial Reorganization of Agriculture," Working Paper 33854, National Bureau of Economic Research.

Munshi, K., and M. Rosenzweig (2016): "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap," *American Economic Review*, 106(1), 46–98.

Murphy, K. M., A. Shleifer, and R. W. Vishny (1989): "Industrialization and the big push," *Journal of political economy*, 97(5), 1003–1026.

Norris, S., M. Pecenco, and J. Weaver (2021): "The Effects of Parental and Sibling Incarceration: Evidence from Ohio," *American Economic Review*.

Phadnis, A., and A. Gupta (2015): "The Politics of Farm Loan Waivers," *Economic and Political Weekly*, 54(23), 7–8.

Rigg, J., A. Salamanca, and E. C. Thompson (2016): "The puzzle of East and Southeast Asia's persistent smallholder," *Journal of Rural Studies*, 43, 118–133.

Rosenstein-Rodan, P. N. (1961): "Notes on the theory of the 'big push'," in *Economic Development for Latin America*, pp. 57–81. Springer.

Rosenzweig, M. R. (1978): "Rural Wages, Labor Supply, and Land Reform: A Theoretical and Empirical Analysis," *The American Economic Review*, 68(5), 847–861.

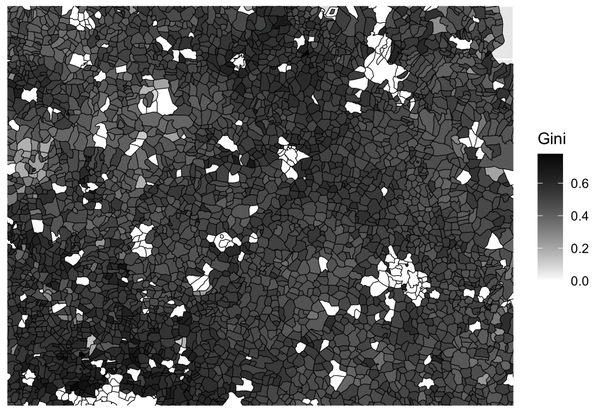
Saiz, A. (2010): "The Geographic Determinants of Housing Supply*," *The Quarterly Journal of Economics*, 125(3), 1253–1296.

Syrquin, M. (1988): "Patterns of structural change," *Handbook of development economics*, 1, 203–273.

Young, A. (2013): "Inequality, the Urban-Rural Gap, and Migration*," *The Quarterly Journal of Economics*, 128(4), 1727–1785.

Figures

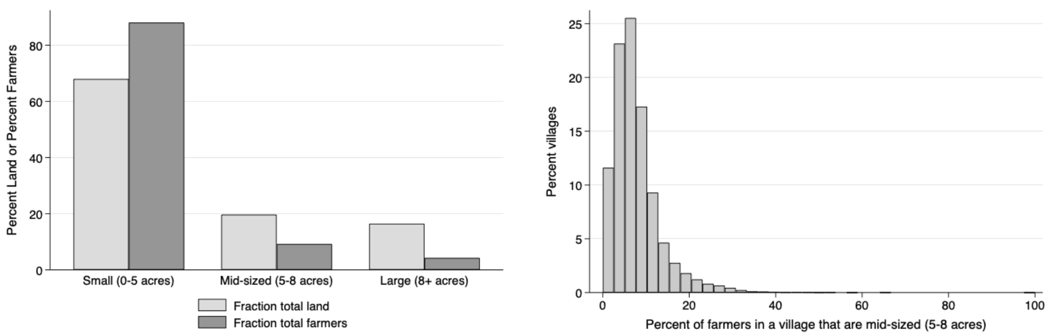
Figure 1: Spatial distribution of rural landholding inequality in India



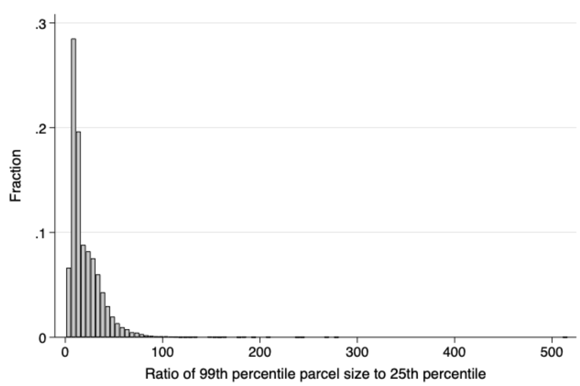
Notes: This figure plots the raw spatial distribution of village-level landholding inequality, measured as Gini coefficient, using the individual-level landholding administrative data. Large white polygons represent urban areas, including statutory towns (definition used for our analysis) as well as census towns and urban outgrowths. Gray-scale polygons represent various degrees of landholding inequality in a village defined by a polygon. Darker shades imply higher inequality. We compare the Gini-coefficient constructed using individual-level landholding administrative using measures of land-concentration (ratio of landholding sizes at different percentiles of the distribution) and also correlate with Gini-index constructed using household census data (which also includes landless households) in Figure A1. We also correlate the constructed measure of inequality with the composition of the landholding distribution, both as a fraction of total land within a village as well as the fraction of individuals in specific land size-bins in Figure 2.

Figure 2: Land size distribution

Panel A: Distribution of farmers by landholding size



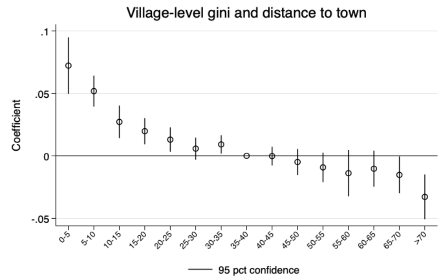
Panel B: Land holding size at 99th percentile relative to 25th percentile



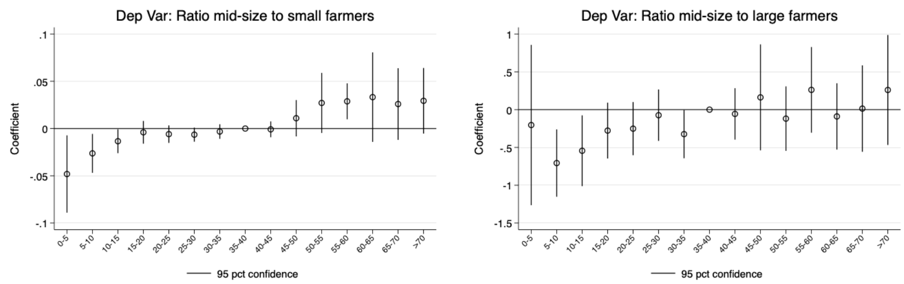
Notes: The figures in Panel A present the summary of farmer size groups in the land records data, which includes the universe of landholders by village. Note that this data does not include the landless. The figure on the right shows the distribution of mid-sized farmers (those with landholdings between 5 and 8 acres) across the study villages. Panel B presents the distribution of the ratio between 99th percentile land holding size and 25th percentile size. This ratio is about 10-20 in about half the villages and more than 20 in another half. For example, if 25th percentile landholding size is 0.5 acre, then landholding size at the 99th percentile is upwards of 10 acres in half the villages.

Figure 3: Key Facts

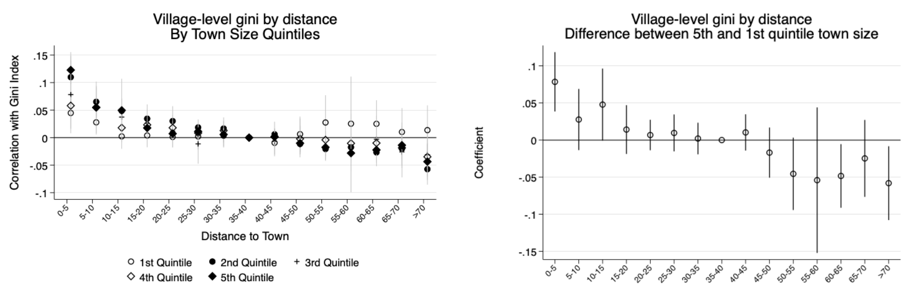
Panel A: Distance Correlation



Panel B: Farm-Size Correlation



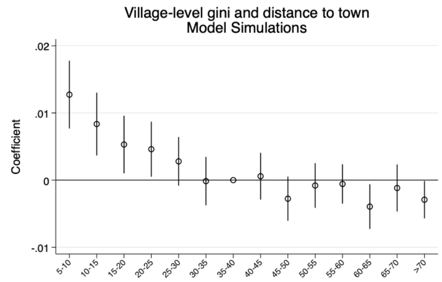
Panel C: Town-Size Correlation



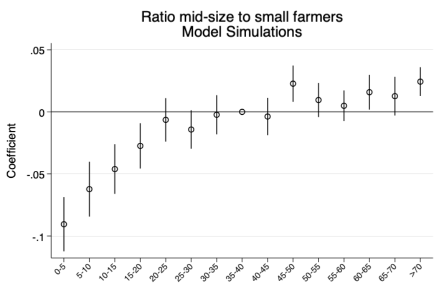
Notes: The panels above document empirical facts about rural landholding inequality based on the distance of the rural area (i.e., a village) and closest town using individual-level landholding administrative data. The x-axes record 5 km distance bins, where the 35-40 km bin serves as the leave-out group. The y-axes correspond to regression coefficients on the specific distance bins when either gini coefficient is the dependent variable (Panels A and C) or the ratio of the number of mid to small or the number mid to large farmers based on their landholding sizes (Panel B). Small and marginal farmers are those with less than 5 acres of land. Mid-sized farmers are those with 5-8 acres of land. Large farmers are those with more than 8 acres of land. Panel C plots the coefficients of the interaction terms between distance bins and the size of the nearest town (based on population quintiles). The regression specifications control for the nearest town and sub-district fixed effects and cluster standard errors by the nearest town. We replicate these patterns using household-level landholding data from household census in the same state in Figure A2, calculating gini by aggregating landholdings within specific distance bands from towns (being agnostic to village boundaries) in Figure A4, and from using an all-India representative household panel survey in Figure A6.

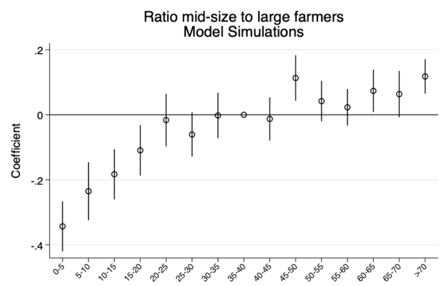
Figure 4: Model Simulations

Panel A: Simulated Distance Correlation

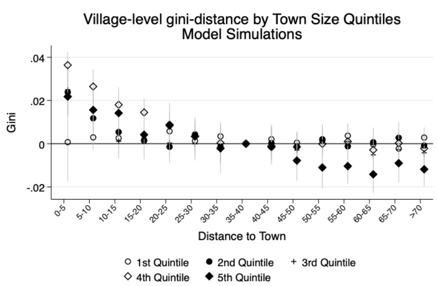


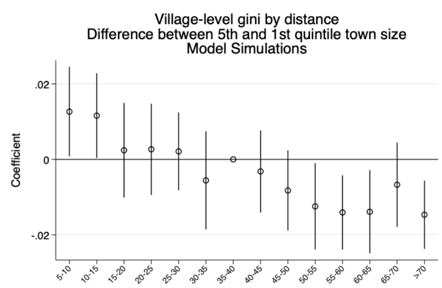
Panel B: Simulated Farm-Size Correlation





Panel C: Simulated Town-Size Correlation

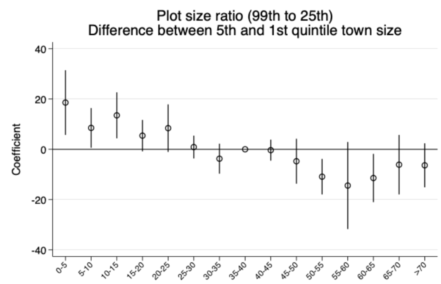




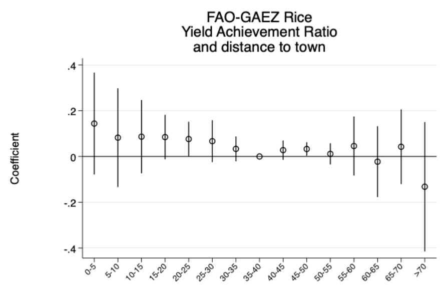
Notes: Above graphs are based on simulated data, following the data generating process as per our model. Simulations include 1500 villages, distributed at random distances from 150 towns of varying population sizes. The distance and town population (town-size) distributions are drawn from random normal distributions with $\mu_d = 45, \sigma_d = 30$ and $\mu_s = 100,000, \sigma_s = 32,000$, respectively. Each village starts with 550 farmers, each with initial land endowment drawn from skewed distributions. We assume log utility functional form, fixed relative price of (agricultural) land within a village (numeraire), and a net urban wage following $w = \frac{s^{0.5}}{d^{0.18}}$. We also assume that a village represents a market and that prices are determined only at the market level (i.e., we do not allow price differences between farmers within a village).The above graphs show the resulting Gini-index/farm size compositions and distance to town correlations subsequent to farmer-level optimization over multiple time-periods/iterations of the model.

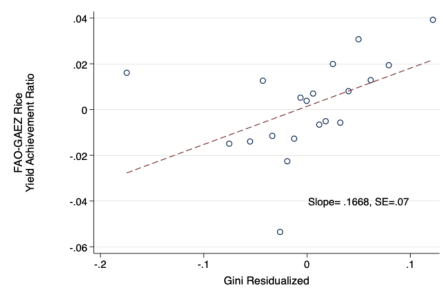
Figure 5: Model Implications

Panel A: Land Consolidation

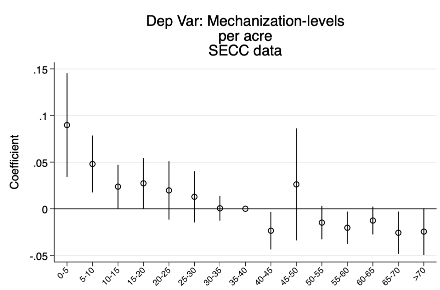


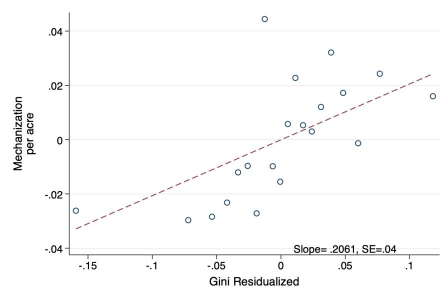
Panel B: Productivity Implications





Panel C: Factor Intensity and Economies of Scale





Notes: Panel A plots the interaction term between distance bin and town size with the ratio of farm sizes at 99th and 25th percentile as the dependent variable. Panel B presents the distance correlation in yield achievement ratio for wetland paddy using FAO GAEZ data (left) and its correlation with rural land inequality (right). We compute the pixel average of the GAEZ data within each village polygon. Panel C presents the distance correlation in mechanization per unit area (capital intensity) using SECC household census data (left) and its correlation with rural land inequality (right). All specifications include the nearest town and sub-district fixed effects and cluster standard errors by the nearest town.

Tables

Table 1: Data Sources

Variable	Source	Obs	Year	Scope
Farmer-level Gini	Land Records	12,843	2017	One State (Universe)
Household-level Gini	SECC	9,984	2012	One State (Universe)
Dist. to Town (km)	Census	10,686	2011	One State (Universe)
Town Size	Census	49	2011	One State (Universe)
Village Bank	Census	10,686	2011	One State (Universe)
Village Road	Census	10,686	2011	One State (Universe)
Village Sec. School	Census	10,686	2011	One State (Universe)
Village Water Src	Census	10,686	2011	One State (Universe)
Agri Outcomes	FAO GAEZ	NA	2010	Raster Image (All India)
HH Panel	IHDS	21919	2005, 2012	All India Sample
Village Module	IHDS	15627	2005, 2012	All India Sample

Notes: This table describes all sources of data used in our analysis in this paper. Between 2011 population census and individual-level landholding data from 2017, the state we study created new villages from existing villages. As a result, all our analysis is at the village unit defined by the 2011 population census, and this is also the key unit when merging the administrative data with other secondary datasets. The population census also provides detailed village-level data on amenities. There are 49 unique towns within the neighborhood of the villages we study, which are reported to be the nearest towns for each of the village in the 2011 population census for our state. Socio-Economic and Caste Census (SECC) was also carried out within a similar time period as the population census and the village units have population census codes to enable merging. The sampling frame in the India Human Development Survey (IHDS) is the entire country, from which villages were randomly selected, stratified by state, to form the primary sampling units (PSU). From each PSU, a sample of households were randomly selected. The original sample consists of 41554 households. Please refer to IHDS documentation for further details. The final observations used for analysis in this paper is restricted based on household-level panel among landed households in 2005.

Table 2: Key Facts: Table

	(1)	(2)	(3)	(4)	(5)
	Gini	Mid-Small Ratio	Mid-Large Ratio	Gini (All)	Gini 1st vs. 5th Quintile
Panel A: Main Specification					
Dist (10 km)	-0.0973***	0.00651*	0.159***	-0.0957***	-0.0256
	(0.0265)	(0.00351)	(0.0473)	(0.0284)	(0.0423)
Dist (10 km) x Town Pop				-0.000314	-0.00187**
				(0.000463)	(0.000752)
Observations	9917	9914	9262	9917	3924
Adj R2	0.44	0.41	0.13	0.44	0.36
Within R2	0.012	0.004	0.001	0.012	0.005
Sub District Fixed Effect	Y	Y	Y	Y	Y
Town Fixed Effect	Y	Y	Y	Y	Y
Panel B: Without FE					
Dist (10 km)	-0.0583***	0.00401***	0.0434**	-0.0603***	-0.0374***
	(0.00457)	(0.000542)	(0.0218)	(0.00501)	(0.00676)
Town Pop (100,000)				0.0202***	0.0205***
				(0.00207)	(0.00211)
Dist x Town Pop				-0.00170***	-0.00231***
				(0.000409)	(0.000417)
Observations	9918	9915	9264	9918	3928
R2	0.019	0.007	0.0007	0.07	0.11
Sub District Fixed Effect	N	N	N	N	N
Town Fixed Effect	N	N	N	N	N

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: Panel A in the table above corresponds to the regressions behind Figure 3. Panel B represents simple bi-variate/interaction specification without any fixed effect. Gini coefficient is transformed into standard deviation (SD) units relative to villages in the 35-40 km distance bin. Distance to the nearest town is measured in multiples of 10 km and the town population is in multiples of 100,000 to enable better readability. For example, the coefficient reported in Panel A, Column 1 should be read as: village-level landholding gini reduces by $\approx$ 0.1 SD with every 10 km increase in distance from the closest town. Columns 2 and 3 present the farm-size composition facts, where mid-size to small farm ratio increases by 0.65 percentage points and mid-size to large farm ratio increases by $\approx$ 16 percentage points for every 10 km distance to town. Columns 4-5 test the town-size interaction where town size is measured as multiples of 100,000 population. Column 5 subsets the data to include village-level gini surrounding only towns in the top and bottom population quintiles to enable the comparison between the gradients surrounding really large and small towns. The coefficient on the interaction should be read as $\approx$ 0.002 SD additional decrease in village gini with every 10 km increase in distance to town in regions surrounding towns with 100,000 more population. Standard errors are clustered by the nearest town to account for spatial correlation within the geography of urban influence.

Table 3: Transition Matrix of Land Ownership Changes: Household Panel Survey

Farm Size (2005)	<= 5 Acre (2012)	5-8 Acre (2012)	>8 Acre (2012)
< 5	0.925	0.0393	0.0361
5-8	0.424	0.302	0.275
8-10	0.265	0.209	0.525
10-20	0.205	0.136	0.659
>20	0.171	0.0642	0.765

Notes: This table presents a simple transition matrix by recording the fraction of households in land size-bins based on their 2005 reported data that are in the respective land size-bins as reported in 2012. Land sizes are measured in acres. Size bin ≤ 5 includes landless households. Bins denoting intervals, such as 5-8, includes households in the upper limit. We classify landholders with landholding size <= 5 acres as small landholders, those in $5 - 8$ acre range are mid-sized landholders, and those in > 8 acre range are large landholders. For example, 92.5% of smallholders continue to remain smallholders in 2012 whereas 30% of mid-sized landholders continue as mid-sized in 2012. We further break-up the large land size-bin in 2005 to examine transitions within each of the subsets of $8 - 10$ acres, $10 - 20$ acres, and those with > 20 acres. Over three-fourths of really large landholders with > 20 acres continue to remain large in 2012. While we note an overall downsizing, we also note a smaller share of transitions to higher land size-bins. For example, 3.6% of smallholders become large in 2012 and over 27% of mid-sized landholders become large. We model both of these forces - downsizing plausibly due to bequest - and upsizing from land acquisition - in the farmer value function from agriculture.

Table 4: Land Transition and Distance to Town: Household Panel

	(1) Absent 2012	(2) Mid-Size 2012	(3) Small 2012	(4) Large 2012	(5) Landless 2012
Large x Dist (10 km)	0.0154***	0.0207*	-0.0170*	-0.00387	0.000221
	(0.00566)	(0.0107)	(0.0102)	(0.00750)	(0.000337)
Small Farm x Dist (10 km)	0.0104**	0.0120	-0.0156*	0.00410	-0.000453
	(0.00477)	(0.00971)	(0.00883)	(0.00372)	(0.000334)
Large Farm (2005)	-0.0312***	-0.352***	-0.105***	0.457***	0.0000193
	(0.0103)	(0.0195)	(0.0175)	(0.0134)	(0.000841)
Small Farm (2005)	-0.0101	-0.489***	0.514***	-0.0273***	0.00215***
	(0.00965)	(0.0180)	(0.0159)	(0.00742)	(0.000835)
Observations	17873	16462	16462	16462	16462
Mean	.229	.054	.508	.067	.371
SD	.42	.226	.5	.25	.483
Fixed Effect	Village	Village	Village	Village	Village

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: This table documents distance-landholding size patterns using household-panel data (IHDS waves 2005 and 2012) from the rural sector. Related to our model, we examine response rates of tracking the same household between the two rounds (absent in 2012) and transition between land size-bins between 2005 (on the RHS) and 2012 (on the LHS of regressions). The sampling frame is all land-owning households in the rural household panel sample. The dependent variable in Column 1 is a dummy variable if a household from IHDS-1 (2005) is not found/surveyed in IHDS-2 (2012). The dependent variables in Columns 2-5 are household land size-bins in 2012. The regressors are household land-size bins based on their 2005 landholding status (small is $<$ 5 acres in 2005, mid-sized is 5-8 acres in 2005, and large is $>$ 8 acres in 2005), and their interaction with distance to the nearest town. The leave-out group are households with mid-sized landholding in 2005. The fixed effect is at the primary sampling unit level, which is the village in the rural sector. Including PSU fixed effect should account for all time invariant unobserved confounders at a very local level. We use Eicker-Huber-White robust standard errors to account for any heterogeneity since villages are randomly sampled. The main coefficients of interest are the distance interaction terms.

Online Appendix for "Towns and Rural Land Inequality in India"

A1 Additional Data Sources

Other measures of village-level agricultural yields include NDVI product made available through BHUVAN NOEDA (by India Space Research Organization or ISRO) that uses Oceansat-2 Ocean Color Monitor (OCM2) sensor data. Lobell (2013) highlights the importance of using satellite data products such as NDVI and other vegetation indices for measuring yields when traditional field-based estimates are not available or are inaccurate. The satellite Oceansat-2 collects images every 2 weeks, covering a wide 1420 km area at a high radiometric resolution mapping to a 1 km x 1 km spatial resolution. The NDVI product generated by Oceansat-2 is highly correlated to that generated using MODIS sensor data, a more commonly used satellite data source. Using the NDVI raster images from the main growing season for each year from 2010 onward, we overlay the 2011 census village shapefiles for our state and compute the average NDVI value contained within each village boundary. The correlations look similar to the results from FAO GAEZ yield achievement ratio if we use NDVI based measure instead.

A2 Model Appendix

A2.1 Empirical Support for Model Assumptions

The two key assumptions in our model are: (a) non-monotonicity of agricultural production as a function of land size (the "U-shaped" yield function) and (b) presence of financial frictions also a function of existing land size that limits expansion of land through debt/borrowing.

A2.1.1 Land-Size and Agricultural Productivity

This assumption is largely based on recent advances in the literature on agricultural productivity and economic development such as Foster and Rosenzweig (2022) with the larger literature summarized by Gollin (2019). While a replication exercise is beyond the scope of this paper, we provide support for this assumption in our data.

In order to show this, we use village-level average NDVI as a proxy for agricultural productivity in the village and examine whether this follows a U-shape relationship based on the percentage of small, mid-sized, and large farmers within a village, in the absence of farm-level data on yields.

Panel A Figure A11 shows that the data supports non-monotonicity where larger percentage of mid-sized farmers reduce overall village-level productivity compared to villages with fewer mid-sized farmers relative to small or large farmers.

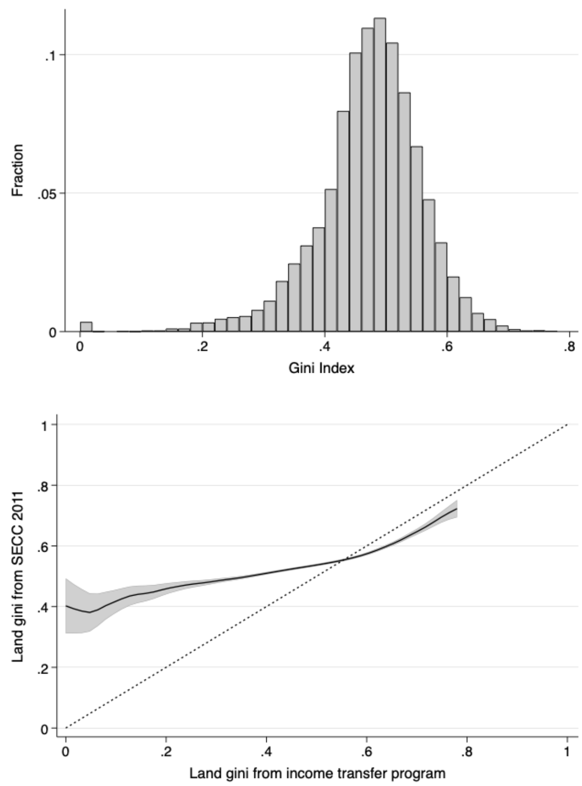
A2.1.2 Financial Constraints

In order to find support for the assumption of financial frictions, we examine land holding changes using household-level panel data from IHDS (rounds 2005 and 2012). We find that households exhibit positive, negative, and no change in the amount of land owned between the two rounds. Among households that report acquiring additional land, we find that the median share of new land acquired relative to initial land holding is 0.69 and close 60% of the households report a change less than the land their owned in the prior period (see Panel B Figure A11).

Further, we find that more mid-sized farmers sell land compared to small farmers or large farmers. The transition matrix Table 1 shows that over 92% of small farmers remain small whereas only 30% mid-sized farmers remain mid-sized between the two survey rounds. Similar to small farmers, a greater share of large farmers continue to remain large (over 76% of farmers with over 20 acres of land continue to remain large).

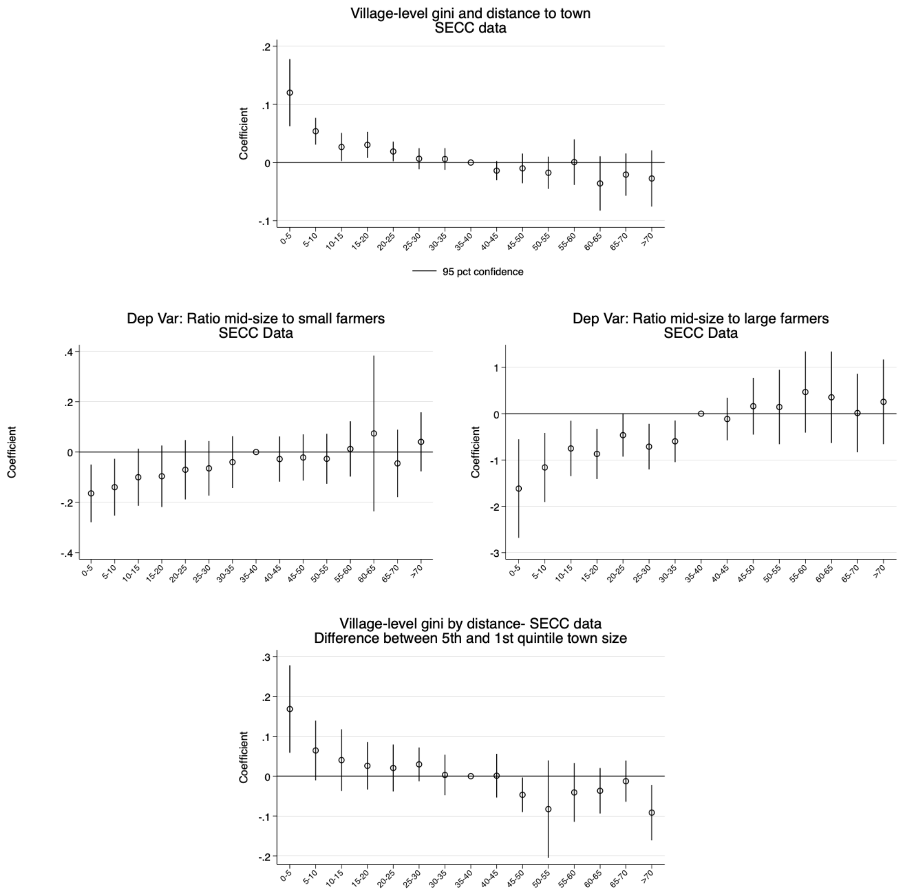
Appendix Figures

Figure A1: Land inequality and land concentration



Notes: The figures presents the distribution of village-level Gini coefficient using farmer-level data (land records used for income transfer) and the correlation between Gini constructed using SECC survey data (self reported landholding at the household level). SECC data also includes the landless and thus the Gini-coefficient computed only using the landholding data is a lower bound on the extent of village-level land inequality.

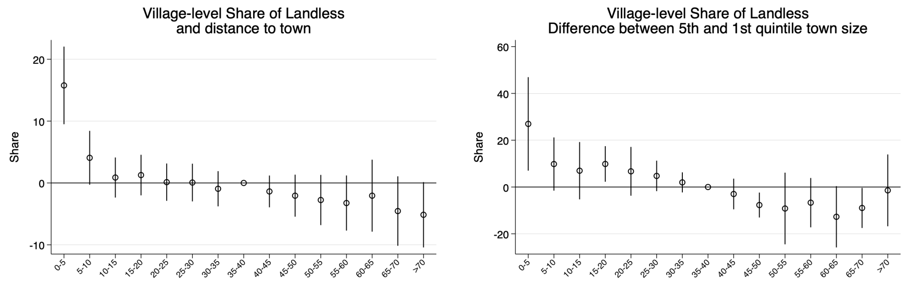
Figure A2: Robustness: SECC data on land ownership by household



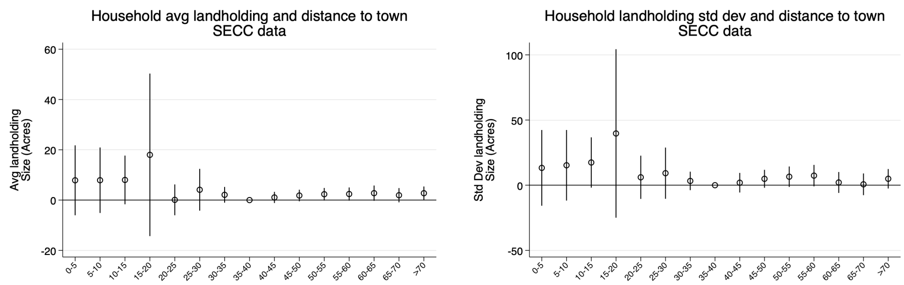
Notes: The figures above recreate Figure 3 using household-level census data from the Socio-Economic and Caste Census (SECC) for the study state, which also includes landless households.

Figure A3: Household Landholding Moments

Panel A: Share of Landless Households Within a Village

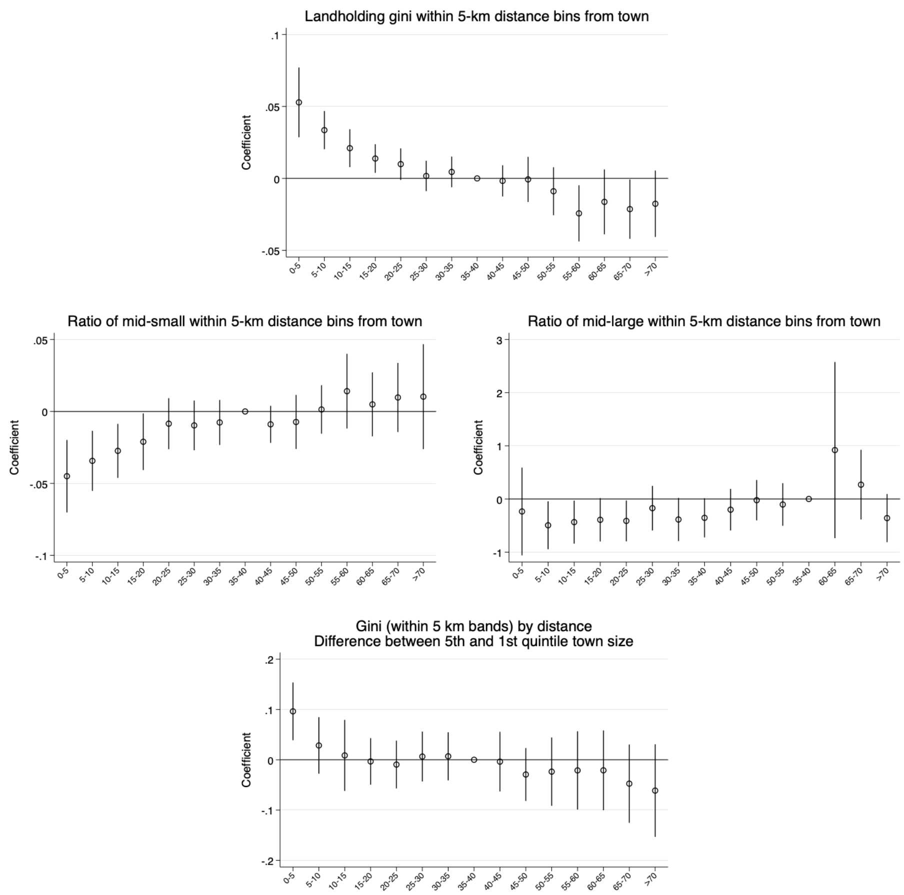


Panel B: Mean and Standard Dev of Household Landholding Within a Village



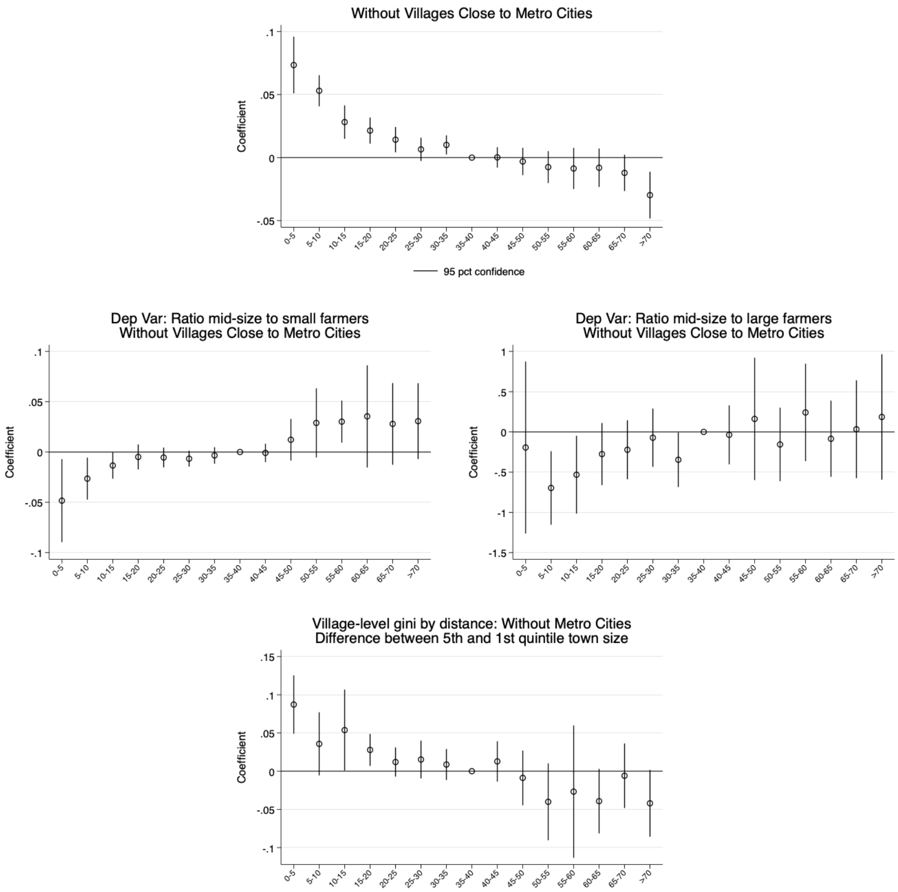
Notes: The figures in Panel A above plot the distance gradients using the share of landless households within a village instead of gini. The figures in Panel B plot the first and the second moments of the household landholding size within a village.

Figure A4: Robustness: 5-km Distance Bins without Village Borders



Notes: The figures above recreate Figure 3 after calculating landholding inequality within every 5 km distance bin from an urban center. That is, we pool all landholdings within each 5 km bins from the closest town and calculate the two dependent variables - landholding gini and farm size ratios.

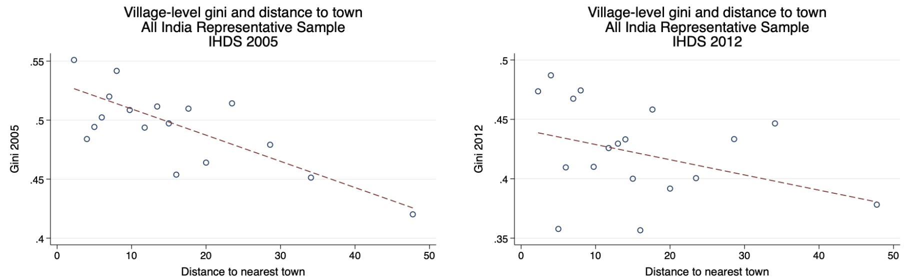
Figure A5: Robustness: Without Metro



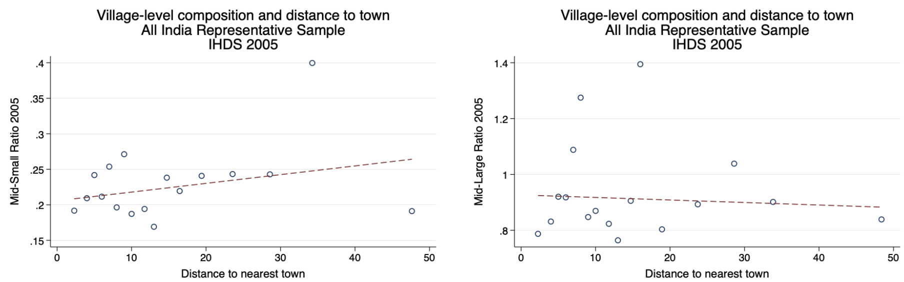
Notes: The figures above recreate Figure 3 after dropping villages in the vicinity of large, metropolitan cities.

Figure A6: Robustness: All India Representative Sample

Panel A: Distance Correlation

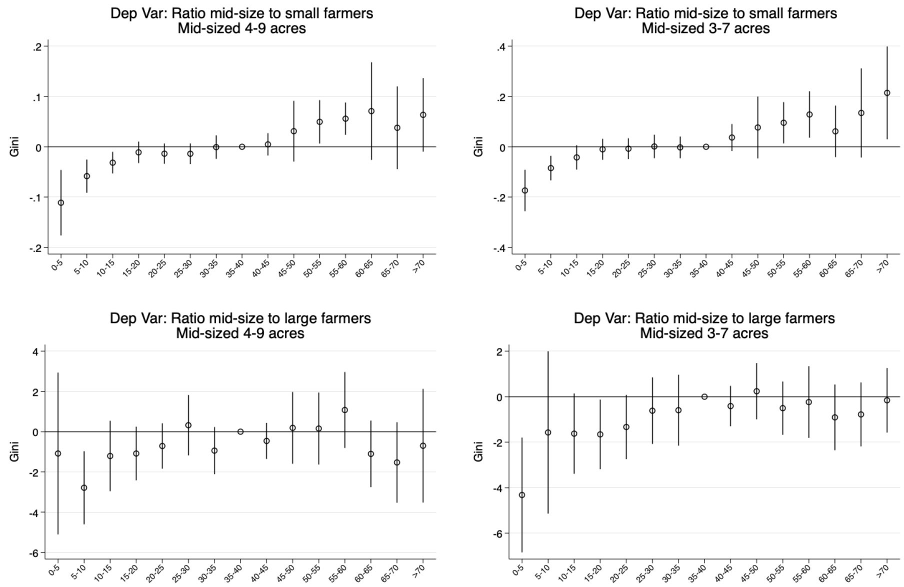


Panel B: Farm-Size Correlation



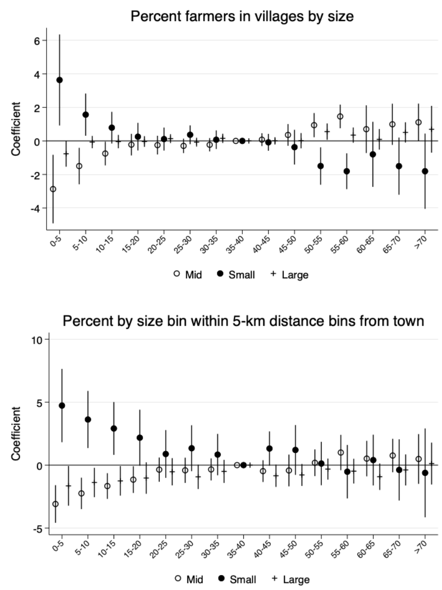
Notes: The panels above recreate the key facts figures using an all India representative sample survey (IHDS) data. Since households are randomly sampled within a village, we do not have accurate information on the full landhold distribution within a village at an all-India-level nor do we have information on the attributes of the nearest town, including its population. However, the random sampling ensures orthogonality with distance to town and we are able to recreate similar urban-distance correlation with village-level landholding inequality.

Figure A7: Robustness: Sensitivity to Landholding Size Cut-offs



Notes: The figures above recreate Fact 2 in Figure 3 using different landholding size cut-offs to classify farmers as small, mid, and large. Our preferred definition classifies small as those with < 5 acres, mid as those with $5 - 8$ acres and large as those with > 8 acres land. The graphs above tests for sensitivity using two different cut-offs. The left panel defines small as < 4 acres, mid as those with $4 - 9$ acres and large as those with > 9 acres. The right panel defines small as < 3 acres, mid as $3 - 7$ acres, and large as > 7 acres.

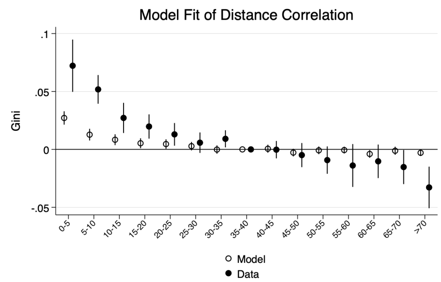
Figure A8: Robustness: Composition as Percentage of Total Farmers



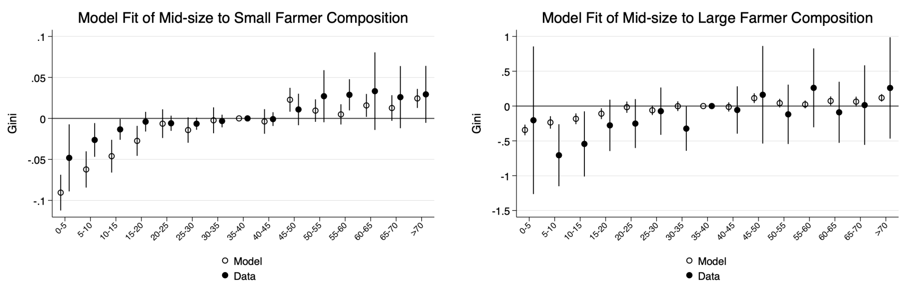
Notes: The figures above recreate Fact 2 in Figure 3 using village boundaries (top) and without village boundaries by including all landholdings within 5 km distance bins from urban centers (bottom).

Figure A9: Model vs. Data

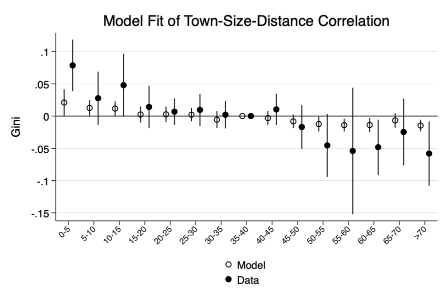
Panel A: Distance Correlation



Panel B: Farm-Size Correlation



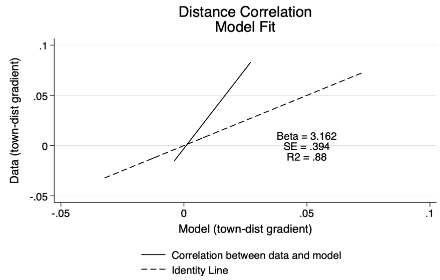
Panel C: Town-Size Correlation



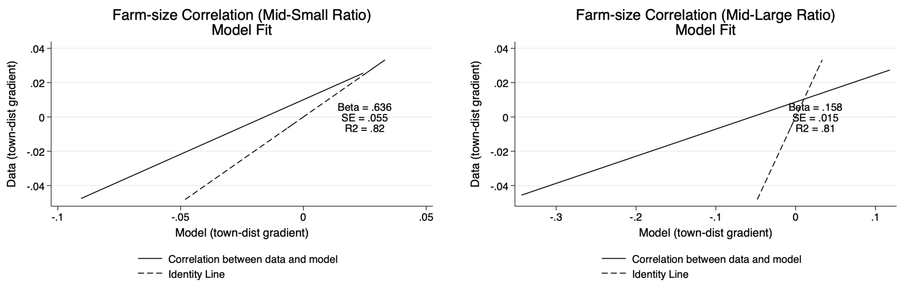
Notes: Above graphs overlay the Gini-distance to town correlation estimates from model with estimates from data across all three key facts.

Figure A10: Model Fit

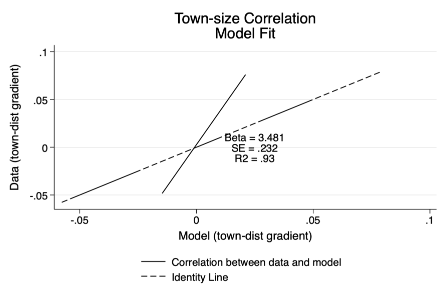
Panel A: Distance Correlation



Panel B: Farm-Size Correlation



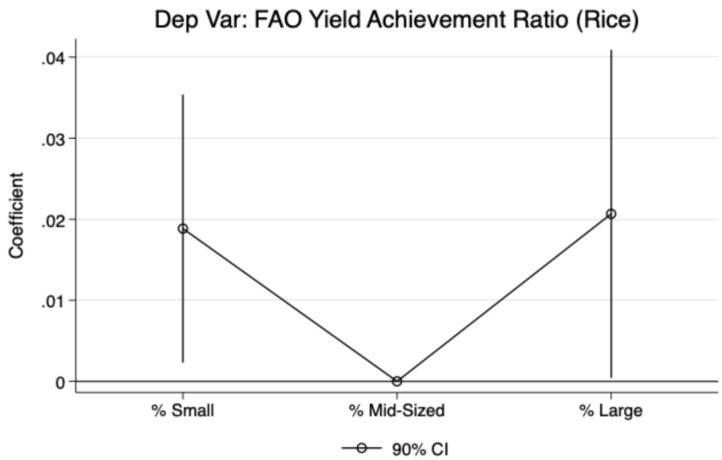
Panel C: Town-Size Correlation



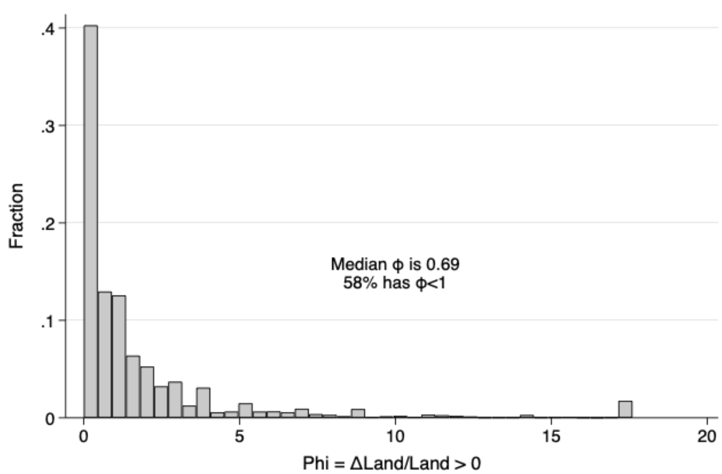
Notes: The figures plot town-distance correlations observed in data against town-distance correlations generated from the model, overlaid with the identity line. Regression coefficients, standard errors, and R^2 from linear regression of correlations observed in data on correlations observed in model simulations are presented to show the model fit.

Figure A11: Supporting Model's Assumptions

Panel A: Land-Size and Agricultural Productivity

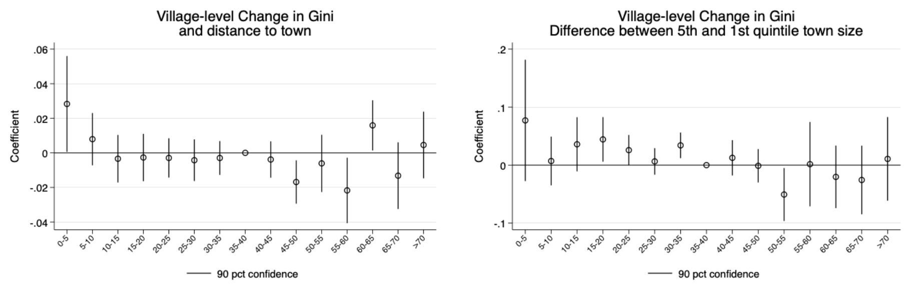


Panel B: Financial Constraints ($\phi < 1$)



Notes: Panel A is generated using village-level average FAO yield achievement ratio data for rice as the dependent variable and percentage of small and large farmers as the explanatory variables (with percentage of mid-sized farmers as the leave-out group). Panel B plots the distribution of ϕ using IHDS data, calculated as the percentage change in land holding between the two survey rounds in 2005 and 2012 when new land was acquired.

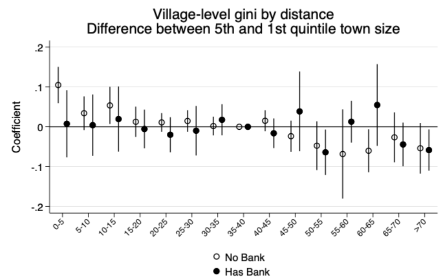
Figure A12: Path to Equilibrium: Not a Steady State



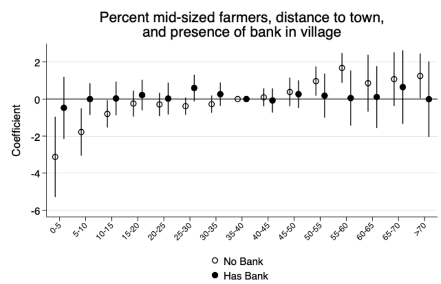
Notes: The graphs above are similar to Figure 3 except that the dependent variable is change in gini over time. The first time-period corresponds to 2011 when SECC data was collected. The second time-period corresponds to 2017-18 when land records were updated for the implementation of farmer income support program. We exclude all landless households from SECC to make it comparable to the land records data from later time-period.

Figure A13: Access to Finance

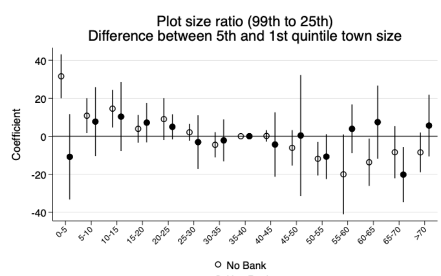
Panel A: Distance Correlation



Panel B: Share of mid-sized farmers



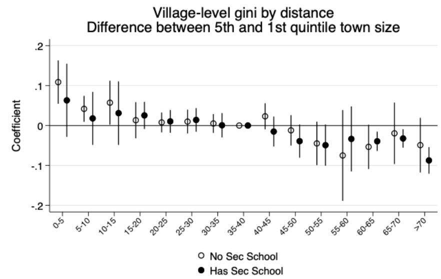
Panel C: Land Consolidation by Large Farmers



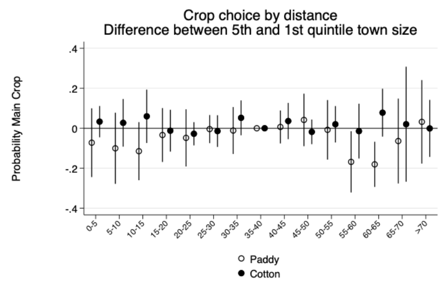
Notes: Panel A plots the interaction term between distance bin and town size with village-level gini as the dependent variable across two specifications - one where villages have a bank or formal financial institution for credit access in hollow circle and another where villages do not have any banks or formal financial institutions in solid black circles. Panel B is also constructed similarly but with the percent of mid-sized farmers as the dependent variable. Panel C examines the ratio of farm sizes at 99th and 25th percentile by town size and distance interaction.

Figure A14: Other Explanations

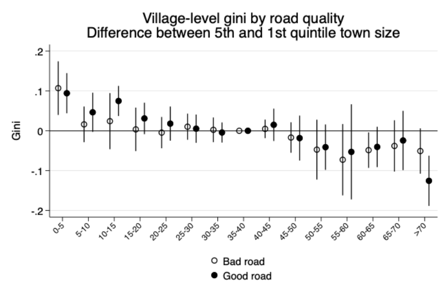
Panel A: Access to Skilling



Panel B: Crop-choice



Panel C: Migration Costs



Notes: The above figure examines whether village-level amenities - importantly, access to secondary schooling (Panel A), crop-choice (Panel B) and good quality roads (Panel C) help explain the dynamics of rural land inequality and urbanization.

Appendix Tables

Table A1: Key Patterns: Using All India Household Panel Data

	(1) Land Inequality 2005	(2) Land Inequality 2012	(3) Ploughing Wage	(4) Harvest Wage	(5) Urban Crop
Dist Town (km)	-0.0022***	-0.0013**	-0.182***	-0.0785***	0.000276
	(0.0006)	(.0006)	(0.0153)	(0.0126)	(0.000227)
Observations	1458	1440	14066	14141	15627
District Fixed Effect	N	N	Y	Y	Y

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: Data in this table include village-level measures of the outcome variables indicated in the column headers constructed using the rural sample of IHDS household data. Dist (km) is the distance in kilometers to the nearest town. The dependent variable in columns 1 and 2 is the village-level Gini index, constructed using the random sample of households. Since households are randomly sampled in IHDS, they are not representative of the village-level landholding distribution. In contrast, the population and census-level measure of Gini index for the state we study is by construction an accurate representation of the village-level landholding Gini index. Outcomes in columns 3-5 are village-level male agricultural wages for ploughing and harvesting, and whether the main crop grown could be classified as an urban crop (fresh fruits and vegetables). Since village identity is not disclosed, we are unable to estimate the town-size interaction since the nearest town identifier is not available in this dataset. We cluster standard errors by district in all the analysis and include fixed effect at the district-level for Columns 3-5 to account for any time invariant labor market unobserved confounders.

Table A2: Potential Omitted Variables

	(1) FAO Rice Suitability	(2) FAO Cotton Suitability	(3) Surface Water Availability	(4) Ground Water Availability	(5) Non-Agri Percent Vill Area	(6) Percent Change Non-Agri Employment
Dist (10 km)	3.044	-0.0467	0.000640	-0.00758	-0.293	12.87
	(2.019)	(0.0482)	(0.00665)	(0.00557)	(0.272)	(7.878)
Observations	10686	10686	10686	10686	10668	7148
Mean	9825.4	2.560	0.320	0.430	10.36	165.6
Std Dev	160.2	3.560	0.470	0.490	14.05	676.8
Sub-District FE	Y	Y	Y	Y	Y	Y
Town FE	Y	Y	Y	Y	Y	Y

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: The dependent variables are village-level characteristics, amenities, and non-agricultural opportunities, respectively, from data pertaining to the study state as described in Table 1. As before, distance to the nearest town is expressed in terms of multiples of 10 km. The specifications include sub-district and town fixed effect and standard errors are clustered by the nearest town.

Table A3: National and State Roads as Instruments

	(1)	(2)	(3)	(4)
	Village-Level Gini (SD units))	Distance Nearest Town (10 km)	Village-Level Gini (SD units)	Village-Level Gini (SD units)
	Reduced Form	First-Stage	2SLS	OLS
Has National HW	0.362***	-0.337***		
	(0.0339)	(0.0567)		
Has State HW	0.236***	-0.288***		
	(0.0256)	(0.0285)		
Dist (10 km)			-0.905***	-0.0973***
			(0.0909)	(0.0265)
Observations	9515	9514	9514	9917
F Stat (First Stg)		91.25		
Over-Id Test (p-val)			0.32	
Adj R-Squared	0.468			0.437
R-Squared	0.496			0.465
Within R2	0.0219			0.0119
Sub District Fixed Effect	Y	Y	Y	Y
Town Fixed Effect	Y	Y	Y	Y

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: The instruments are whether a village has a national highway ("Has National HW") or a state highway ("Has State HW") passing through it. Construction and maintenance of national highways are under the central government whereas the state highways are constructed and maintained by the corresponding state government. The specifications include sub-district and town fixed effect and standard errors are clustered by the nearest town.

Table A4: Testing the Family Size Explanation

	(1) Share Children (2005)	(2) HH Split (2012)	(3) Change HH Size (2012)	(4) Land Inherited (2012)
Large x Dist (km)	0.000617	-0.000883	0.0000351	0.000384
	(0.000648)	(0.00118)	(0.0000657)	(0.00108)
Small Farm x Dist (km)	0.00113*	-0.000585	0.0000130	-0.000130
	(0.000612)	(0.000909)	(0.0000517)	(0.000859)
Large Farm (2005)	-0.0233*	0.0719***	-0.00222	0.00689
	(0.0122)	(0.0217)	(0.00162)	(0.0185)
Small Farm (2005)	0.0142	-0.0393**	0.000512	-0.0166
	(0.0112)	(0.0181)	(0.00138)	(0.0169)
Observations	13170	13170	13170	10286
Village Fixed Effect	Y	Y	Y	Y

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: This table documents distance-landholding size patterns using household-panel data (IHDS waves 2005 and 2012) from the rural sector. The dependent (LHS) variables are from 2012 (later wave) whereas the explanatory (RHS) variables are from 2005. The sampling frame is all land-owning households in the rural household panel sample. We test for demographic factors as potential alternate explanation in this table instead of urban opportunity cost for mid-sized landholders closer to town. The dependent variable in Column 1 is the share of children within the family, who may come of age in 2012, and thus result in subdivision of land. The dependent variable in Column 2 is whether the household was split in 2012. Column 3 documents whether the household inherited any land in 2012. And lastly, Column 4 documents the change in family size between 2012 and 2005. The regressors are household land-size bins based on their 2005 landholding status (small is < 5 acres in 2005, mid-sized is 5-8 acres in 2005, and large is > 8 acres in 2005), and their interaction with distance to the nearest town. The leave-out group are households with mid-sized landholding in 2005. The fixed effect is at the primary sampling unit level, which is the village in the rural sector. Including PSU fixed effect should account for all time invariant unobserved confounders at a very local level. We use Eicker-Huber-White robust standard errors to account for any heterogeneity since villages are randomly sampled. The main coefficients of interest are the distance interaction terms. Although we note demographic changes by land-size bins, we find no distance correlation that could explain the empirical patterns.

Table A5: Testing the Differential Skilling Explanation

	(1) Schooling Years	(2) Share Salaried	(3) Share Family Farm Labor	(4) Total HH Income
Large x Dist (km)	0.00368	-0.000107	-0.000192	-497.7**
	(0.0125)	(0.000273)	(0.000690)	(240.3)
Small Farm x Dist (km)	0.00725	0.0000921	-0.000755	-14.20
	(0.0104)	(0.000219)	(0.000614)	(164.4)
Large Farm (2005)	1.381***	0.0115**	-0.0120	43049.5***
	(0.225)	(0.00487)	(0.0130)	(5013.8)
Small Farm (2005)	-1.531***	0.00519	-0.0266**	-17950.6***
	(0.192)	(0.00412)	(0.0118)	(3035.1)
Observations	15573	15573	15573	15572
Village Fixed Effect	Y	Y	Y	Y

Standard errors in parentheses
* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Notes: This table documents distance-landholding size patterns using household-panel data (IHDS wave 2005) from the rural sector. The sampling frame is all land-owning households in the rural household panel sample. We test for differential skilling as potential alternate explanation in this table instead of urban opportunity cost for mid-sized landholders closer to town. Column 1 reports the correlations using maximum years of schooling among the household members as the dependent variable. Columns 2 and 3 report correlation using share of household members engaged in salaried and own farm labor, respectively, as the dependent variables. Column 4 examines total household income's correlation with farm size and distance interaction. The regressors are household land-size bins based on their 2005 landholding status (small is $<$ 5 acres in 2005, mid-sized is 5-8 acres in 2005, and large is $>$ 8 acres in 2005), and their interaction with distance to the nearest town. The leave-out group are households with mid-sized landholding in 2005. The fixed effect is at the primary sampling unit level, which is the village in the rural sector. Including PSU fixed effect should account for all time invariant unobserved confounders at a very local level. We use Eicker-Huber-White robust standard errors to account for any heterogeneity since villages are randomly sampled. The main coefficients of interest are the distance interaction terms. Although we note skilling differences by land-size bins, we find no distance correlation that could explain the empirical patterns.