IIMB-WP No. 668/2022

WORKING PAPER NO: 668

Land Misallocation and Industrial Development

Kunal Dasgupta
Associate Professor of Economics and Social Sciences
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
kunal.dasgupta@iimb.ac.in

Rahul Rao
Doctoral Student
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
rahul.rao18@iimb.ac.in

Year of Publication – July 2022

Land Misallocation and Industrial Development

Kunal Dasgupta* Rahul Rao†

March 2022

Abstract

In many poor and densely populated countries, a large fraction of land is devoted to agriculture. This makes industrialization and urbanization, both of which require land, a serious challenge. We argue that land can be re-allocated from agricultural to non-agricultural uses without any adverse effect on agricultural output when the existing land use in agriculture is sub-optimal. Using actual and potential crop yield data from Indian districts, we show that a planner who allocates agricultural land to its best possible use can release up to 40 percent of land without affecting aggregate agricultural output. The effect is analogous to land-augmenting productivity increase in the agricultural sector. Using a calibrated two-sector model, we find that re-allocation of the freed up land from agriculture to the manufacturing sector raises manufacturing output by 29 percent and real income by 9.4 percent.

JEL codes: Q15,O13,O21.
Keywords: Mis-allocation, Agriculture, Manufacturing, Land, GAEZ.

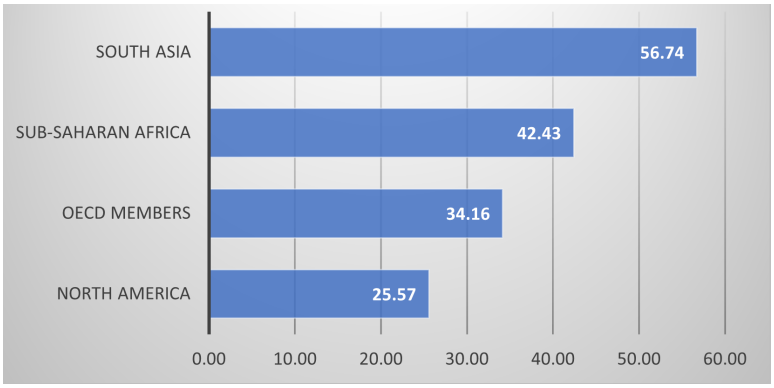
*Associate Professor, Economics & Social Sciences Area, Indian Institute of Management, Bangalore (E-mail: kunal.dasgupta@iimb.ac.in)

†PhD Student, Economics & Social Sciences Area, Indian Institute of Management, Bangalore (E-mail: rahul.rao18@iimb.ac.in)

1. Introduction

In densely populated and poor countries, a majority of land is used in the agricultural sector (see Figure 1). Accordingly, it is difficult to acquire land for non-agricultural uses, namely industrial development and urban expansion. A scarcity of land, then, becomes a constraint on industrialization and urbanization in these countries.

Figure 1: Percent of land under agriculture



Note: The data is average for years 2014-18.
Source: World Bank

Consider the case of India where land acquisition is frequently cited as a major hurdle in the development of critical infrastructure in the country.[1] According to the World Bank, in 2018, roughly 60 percent of total available land area in India was used for cultivation.[2] At the same time, a popular position among Indian policymakers is that only land that is not currently being used for cultivation should be made available for industrialisation.[3] This argument, however, ignores the fact that in a developing country like India, large swathes of agriculturally unproductive land are also cultivated, resulting in yields that are often extremely low. This is clearly displayed in Table 1 which shows the average yields for rice and wheat in BRICS nations, a group including India, that are at comparable levels of development.[4]

Our paper asks the following question: how much land can be freed up for non-agricultural usage in India if agricultural land is allocated to its best use? To answer this question, we rely on a novel agronomic data set that provides potential yields (a measure of land productivity) for a set

[1] For instance, (Mohanty et al., 2009) noted that 70 percent of delays in infrastructure and other development projects were caused due to issues related to land acquisition. Similarly, according to an estimate by ASSOCHAM (The Associated Chamber of Commerce and Industry of India), the country's leading business association, projects worth US$100 billion are at stake due to land acquisition – many of them critical infrastructure projects linked to railways, national highways, ports and power plants (https://www.orfonline.org/research/the-land-acquisition-stalematecontentions-solutions/).

[2] https://data.worldbank.org/indicator/AG.LND.AGRI.ZS

[3] Fertile land, it is argued, has many stakeholders with high opportunity costs (Ranganathan, 2010).

[4] BRICS consists of Brazil, Russia, India, China and South Africa.

Table 1: Yields of rice and wheat in BRICS countries (2014-18)

	Rice	Wheat
China	5.27	5.39
Brazil	4.58	2.50
South Africa	4.57	3.45
Russia	2.61	2.69
India	2.50	3.10

Note: Yields are measured as tonnes per hectare.
Source: UN Food and Agriculture Organization (FAO)

of crops in millions of fields that span the entire earth.[5] We use this data to compute the minimum amount of land that a central planner would require to achieve the actual district-level production of every crop if land is allocated to crops optimally. We then calculate inefficiency of a district in a particular season as the percentage difference between actual and optimal land usage – higher is this difference, greater is the inefficiency. We consider two main non-overlapping seasons: (i) Kharif, and (ii) Rabi.[6] The total area under cultivation during the Rabi season being significantly lower than that during the Kharif season, we focus on the Kharif season and report results for the Rabi season in the Appendix.[7]

We find that up to 20.7 million hectares (or 32 percent) of agricultural land cultivated during the Kharif season can be potentially released across all districts in India. On the other hand, if the planner meets targets for actual production at the level of states, which are higher administrative units than districts, 36.8 million hectares (or 47 percent) can be freed up. Intuitively, moving from the district-level to the state-level eases the planner's constraint in terms of which fields to cultivate. In fact, if the planner meets production targets at the country-level, then up to 45.7 million hectares (or 57 percent) of land can be freed up.

In the baseline scenario, actual land usage is less than the optimal in several districts. At first glance, this result is puzzling. On further inspection, we find that the actual yields in these districts are higher than even the *maximum* potential GAEZ yield across fields within those districts. One possible explanation is that the higher actual yield is due to application of more/higher quality inputs than what we have assumed.[8] Alternately, the higher actual yield could result from

[5]This data has been prepared under the Global Agro-Ecological Zones (GAEZ) project by the Food and Agricultural Organization (FAO) and the International Institute for Applied Systems Analysis (IIASA).

[6]Kharif includes summer and whole year crops which are grown from early April to September. Rabi includes winter, autumn and whole year crops which are grown from late September until March.

[7]We solve the problem season-wise because farmers grow multiple crops throughout a year, while GAEZ data reports potential yield for one-time production only.

[8]The GAEZ potential yields are computed for different water supply and input usage intensities. In the baseline

only selected areas within a field being used in agriculture. To see this, note that the potential yield of a GAEZ field is an average over several, presumably heterogeneous, sub-fields that lie inside it. If the planner does not use the entire area within a GAEZ field, the effective yield for land that is actually cultivated could be different from the average potential yield. For example, the planner could be using the top 25th percentile of sub-fields in terms of productivity. In this case, he would require less land to grow a given amount of a crop relative to what he would require if, instead, he had used sub-fields with average productivity.

Next, we examine this selection hypothesis whereby only a subset of plots within a field are being cultivated. Observe that selection could be either positive or negative. While positive selection, as discussed above, is a result of farmers using the most productive plots of land, negative selection arises when such productive plots are not available for cultivation.[9] Accounting for both positive and negative selection reduces the number of districts in which actual land usage is less than the optimal. Furthermore, the planner can free up to 25.2 million hectares (or 39 percent) of agricultural land, which is more than the benchmark scenario.

The constrained optimization problem that we solve involves minimizing the land required for producing a given amount of output. Freeing up agricultural land by re-allocating it to its best use, however, is analogous to land-augmenting productivity improvement in the agricultural sector. In the last part of the paper, we examine the welfare implications of such a productivity increase. To do this, we develop a two-sector model where production is carried out in both sectors using land and labor. Both factors face mobility barriers; for labor, these could be barriers to physical mobility while for land, these could be legal impediments to re-allocating land from agricultural to non-agricultural usage. We calibrate the model using data from the agricultural and manufacturing sector of India. Following an increase in agricultural productivity in the model, we find that (i) both land and labor move from agriculture to the manufacturing sector, (ii) land price declines while wage goes up, (iii) both agricultural and manufacturing output go up, (iv) agricultural prices go down, and (v) real income in the economy increases by 9.5 percent.

This paper adds to the literature that emphasizes the role of lower agricultural productivity in keeping countries poorer (Gollin et al., 2002; Caselli, 2005; Restuccia et al., 2008). These papers argue that higher employment share along with low labor productivity in agriculture is central to keeping poor countries' aggregate productivity at low levels. These papers, however, do not explicitly account for the role of geography in contributing to low agricultural productivity. Lately, with the availability of the GAEZ dataset, some papers have attempted to incorporate geography into the analysis as well. Adamopoulos and Restuccia (2018) perform a cross-country analysis to examine the role of geography in lowering agricultural productivity. They find that the lower

scenario, we assume rain-fed/irrigated water supply and intermediate input usage.

[9] This could be due to the presence of buildings, infrastructure, forests, water bodies, etc.

agricultural productivity in poor countries is partly due to non-availability or ignorance about scientific methods of crop production, and partly due to misallocation of heterogeneous land across crops.

Our paper uses the GAEZ dataset to document the extent of misallocation of heterogenous land across different crops in the Indian agricultural sector. To the best of our knowledge, we are the first ones to do so. Somewhat related work is Bolhuis et al. (2021) who report the misallocation across heterogeneous farms or farmers within and between different Indian states due to distortions in land rental markets. But, we study a different land misallocation problem in our paper. By combining potential yields data from GAEZ with actual yields data, we ask the following question: how much agricultural land can be released for non-agricultural use if the current level of output is produced by a planner who grows crops by allocating them optimally across land? This is the first contribution of our paper.

The second contribution is to identify a potential methodological shortcoming in the literature which uses GAEZ dataset to answer other economic questions. Papers such as Costinot and Donaldson (2016) assume that the entire GAEZ field is available for agricultural use. The GAEZ field, at roughly 100 square kilometers, is large. As we argue in this paper, use of only a fraction of a GAEZ field is more realistic than assuming the usage of the entire field for cultivation. Now, if the land *within* a GAEZ field is heterogeneous in terms of productivity, there could be selection. Although Costinot et al. (2016) and Adamopoulos and Restuccia (2018) assume heterogeneous land parcels within a GAEZ field to derive some of their theoretical results, they do not account for a possible selection on these parcels. Selection would drive a wedge between the average yield of a field and the average yield of parcels within that field that end up being cultivated. While we begin our analysis with the same (entire field being available) assumption, we relax it later to account explicitly for selection.

The rest of the paper proceeds as follows. In Section 2, we describe the data sets used in the analysis. We define the planner's problem in section 3. The results are presented in section 4. Section 5 sets up the two-sector model while section 6 quantitatively analyses the model. Section 7 concludes.

2. Data

In this section, we describe the data sets used in our analysis. We describe the various procedures to clean the data in the Appendix.

2.1 GAEZ Data

We use micro-geographical field level potential yield data compiled by the Global Agro-Ecological Zones (GAEZ) project.[10] This project has been developed by the Food and Agricultural Organization (FAO) in collaboration with the International Institute for Applied Systems Analysis (IIASA). GAEZ fields are grid cells at 5 arc-minute resolution. These cells are not uniform in size because the mapping from arc-minutes to square kilometers depends on the distance from the equator (latitude).[11] At such high resolution, GAEZ captures the following geographic attributes of each field that are important for agricultural production: (1) soil quality, which includes depth, fertility, drainage, texture, chemical composition; (2) climate conditions, which include temperature, sunshine hours, precipitation, humidity, and wind speed; and (3) terrain and topography, which include elevation and slope.

These "natural inputs", along with assumptions about level of "human inputs" such as fertilisers, irrigation and cultivation practices, are then fed into a state-of-the-art crop-specific agronomic model. The model has several crop-specific parameters that govern how a given set of growing characteristics map into crop yields based on the crop's biophysical requirements for growth. It calculates the potential yield (in tonnes per hectare) for each crop in each field, *not just for the ones which are being actually grown there.* This potential yield is the maximum output that can be attained in the field given the crop's production requirements, the field's geographic characteristics, and assumptions about human input levels.

GAEZ reports potential yields for two types of human inputs: (i) water supply conditions, and (ii) level of complementary inputs. Three categories of the former (irrigated, rain-fed and total) and three categories of the latter (low, intermediate and high) are considered. The key observation is that because the human inputs and crop-specific model parameters are the same for each field, the potential yield for a crop varies across fields due to field-level geographic attributes only. See Adamopoulos and Restuccia (2018) for a detailed discussion of this data.

To provide a graphical preview of the data, figure 2 shows the potential yield for growing rice crop across all the GAEZ fields that span India. The yields correspond to intermediate input usage and mixed water supply sources.[12] As visible, there is a lot of variation in the natural productivity of different plots of land in growing rice across India. Also, we can see many regions which are totally unproductive for growing rice like almost the entire Western and North-Western parts of the country. In the results section, we also provide quantitative assessment of such variation for different crops at different administrative levels.

[10] http://www.gaez.iiasa.ac.at

[11] At the equator, these cells are roughly 10 kilometers by 10 kilometers fields.

[12] Rather than uniformly assigning the same water supply source across all the fields, we take into account the fact that water supply source varies across fields, with some having access to irrigation while others relying on rain. The Appendix provides the details.

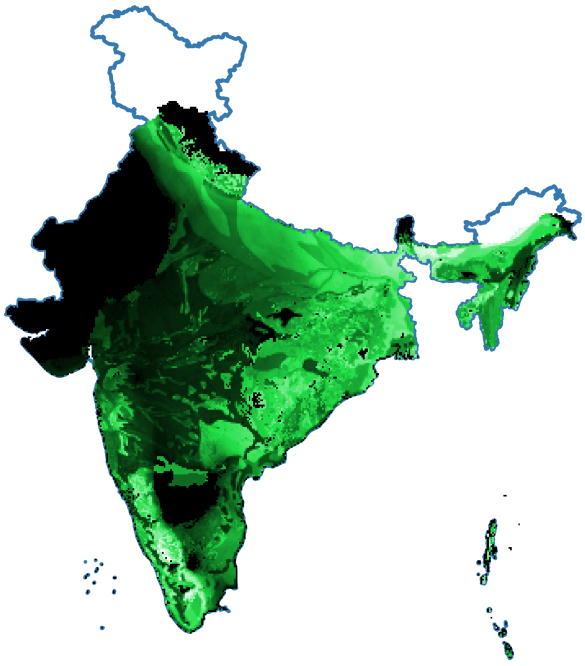


Figure 2: Potential Yield of growing rice in various parts of India. Brighter regions have higher yield while darker regions are naturally unproductive for growing rice.

We spatially merge these fields with boundaries for the districts,[13] the most dis-aggregate level at which production data is available. We have a total of 618 districts, with the median district covering 107 fields. We include 21 crops in our analysis. Table 5 in the Appendix lists all the crops along with some descriptive statistics.

2.2 Production Data

We use district and season-wise crop production statistics compiled by the Directorate of Economics and Statistics, Government of India.[14] For our analysis, we use production data averaged over 2007-2011 for the crops that were grown in 2011.[15] We are primarily interested in the area and the volume of production for each crop.[16]

Before the potential and actual production data can be merged, however, an adjustment

[13] District boundaries are obtained from the 2011 Census of India.

[14] https://eands.dacnet.nic.in

[15] Potential yields do not vary over time, although see Costinot and Donaldson (2016).

[16] The unit for area of production is hectares for all crops, while the unit for volume of production is tonnes except for coconut and cotton. Coconut production is measured in number of nuts produced. We assume the average weight of a green coconut to be 1.3 kilograms and convert into tonnes (https://www.researchgate.net/figure/Weight-of-dry-and-green-coconut_tbl3_309596220). Similarly, cotton output, which is measured in bales, is converted into tonnes as well.

needs to be made.[17] A large fraction of farmland in India witnesses multi-cropping whereby multiple crops are grown on the same piece of land within a year (Frolking et al., 2006). To see why this could be an issue, suppose a field is used to grow X tonnes of rice in one season and Y tonnes of wheat in a different season. If we do not take multi-cropping into account, the planner would require twice the land than actually used to produce X tonnes of rice and Y tonnes of wheat. To deal with this issue, we consider the allocation problem season-by-season.

The production data is reported for six seasons – Summer, Kharif, Autumn, Rabi, Winter and Whole Year.[18] For our analysis, we combine mutually non-exclusive seasons into two broad categories – (i) Kharif, which includes Summer and Whole Year, and (ii) Rabi, which includes Winter, Autumn and Whole Year. We assume that there is no overlap between these two seasons. For simplicity, we also assign half of the area and volume of production of Whole Year crops to both seasons.[19]

Table 5 in the Appendix reports (5-year average) yields for different crops across the two seasons. All the 21 crops are produced during Kharif season, while all crops except soybean are produced during Rabi season. According to the table, average yields are lower in the Kharif season for most crops. Combined with a higher area under cultivation, this suggests a possibly higher misallocation of land during the Kharif season.[20] Accordingly, we focus on the Kharif season and report the corresponding results for the Rabi season in the Appendix.

3. Planner's Problem

In this section, we discuss the planner's problem. The planner's objective is to minimize the land required for producing the actual output of crops at the level of an administrative region/unit, where the latter could be a district, a state or the entire country. One could think of the planner as solving a two-stage problem. In the first stage, the planner chooses how much of each crop to produce. In the second stage, the planner chooses how much land to use for meeting the production targets solved in the first stage. We are implicitly assuming that the first stage has already been solved and focus on the second stage.

We divide the entire country into I administrative units which are indexed by $i \in \mathcal{I} \equiv$

[17] The actual production data for each crop measures fresh weight after harvest whereas GAEZ uses dry weight to calculate potential yields. Accordingly, we convert the actual production data by using standard conversion factors provided by GAEZ.

[18] Kharif season starts at the onset of monsoon, between late May and mid July, and ends by September. Autumn crops have their production cycle starting in September and ending late November. Rabi and Winter seasons are synonymous and their cycle lasts from October until March. Summer crops are grown between Kharif and Rabi, and their cycle is from March until late June. Finally, Whole Year crops are produced throughout the year with multiple sowing and harvesting cycles distributed across various seasons (Source: http://www.arthapedia.in/).

[19] There could be instances where a crop is reported under both Winter and Autumn seasons, for example. In such cases, we only consider the season for which higher production is reported.

[20] Overall, 79.90 million hectares of agricultural land are used in the Kharif season, while the corresponding number for the Rabi season is 68.40 million hectares land.

$\{1,2,...,I\}$. Each administrative unit has heterogeneous GAEZ fields indexed by $f \in \mathcal{F}_i \equiv \{1,2,...,F_i\}$, with the area of field f in unit i being denoted by L_i^f. Each field can potentially grow K crops, where a crop is indexed by $k \in \mathcal{K} \equiv \{1,2,...,K\}$. Let A_i^{fk} denote the potential productivity or yield of field f in producing crop k under a given intensity of input usage and water supply technology. Note that A_i^{fk} is exogenous and varies with both f and k within administrative region i.

Let π_i^{fk} denote the fraction of land allocated to crop k in field f by the planner. The quantity of crop k produced in field f is then given by

$$Q_i^{fk} = A_i^{fk}\pi_i^{fk}L_i^f.$$

π_i^{fk} is the choice variable of the planner. We can then write the planner's problem as

$$\min_{\pi_i^{fk}} \sum_k \sum_f \pi_i^{fk} L_i^f, \tag{1}$$

subject to

$$\sum_{f\in\mathcal{F}} A_i^{fk}\pi_i^{fk}L_i^f = \hat{Q}_i^k, \ \forall k \in \mathcal{K} \tag{2}$$

$$0 \leq \pi_i^{fk} \leq 1, \tag{3}$$

$$\sum_{k\in\mathcal{K}} \pi_i^{fk} \leq 1. \tag{4}$$

In (2), $\hat{Q}_i^k$ denotes the actual production (in tonnes) of crop k in region i. The inequality in (3) allows for π_i^{fk} to be equal to zero – a field may not grow every crop. The inequality in (4) suggests that an entire field may not be used for crop production – parts of a field could be used for non-agricultural purposes.

We solve for the allocations $\{\pi_i^{fk}\}_{f\in\mathcal{F}_i;k\in\mathcal{K}}$ for each region i. It entails solving a large-scale linear programming problem.[21] The solution gives the optimal area required to meet the actual production of all crops within region i. We then define the *inefficiency index* of region i in allocating land to agricultural crops as follows:

$$\text{Inefficiency} = \frac{\text{Actual Area} - \text{Optimal Area}}{\text{Actual Area}}. \tag{5}$$

Inefficiency in region i denotes the percentage of actual area under cultivation that can be released if the crops are optimally allocated across fields.

[21] We use the *linprog* MATLAB function to solve this problem.

4. Results

In this section, we report the results of the planner's constrained optimization problem. A necessary condition for the planner to achieve a reduction in total land usage is that there is variation in potential yield across fields for most, if not all, crops. For example, to reduce the amount of land required to grow, say, wheat in a given district, the planner should be able to re-allocate wheat production from fields with lower potential yield to those with higher potential yield. Table 2 shows the variation in potential yields at different administrative levels for the five largest crops in terms of land usage. There exists substantial variation in the potential yield of all crops that can be exploited by the planner for reallocating land within a given administrative region. This variation is highest at the country level for all crops.

Table 2: Variation in potential yields across fields

Crops	# of districts	Districts		States		Country	
		$75^{th}/25^{th}$	$90^{th}/10^{th}$	$75^{th}/25^{th}$	$90^{th}/10^{th}$	$75^{th}/25^{th}$	$90^{th}/10^{th}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rice	540	2.40	4.63	2.26	4.23	3.40	8.64
Cotton	254	1.92	3.24	1.90	2.99	2.93	4.7
Soybean	215	2.12	3.76	1.91	3.05	3.00	5.57
Pearl Millet	284	2.55	4.79	1.93	3.16	3.34	4.88
Maize	523	2.11	3.70	1.89	3.20	3.18	5.11

Note: The table reports the mean value for the ratio of 75^{th} to 25^{th} percentile potential yields (columns 2, 4 and 6) and the ratio of 90^{th} to 10^{th} percentile potential yields (columns 3, 5 and 7) at different administrative levels.
Source: GAEZ data set

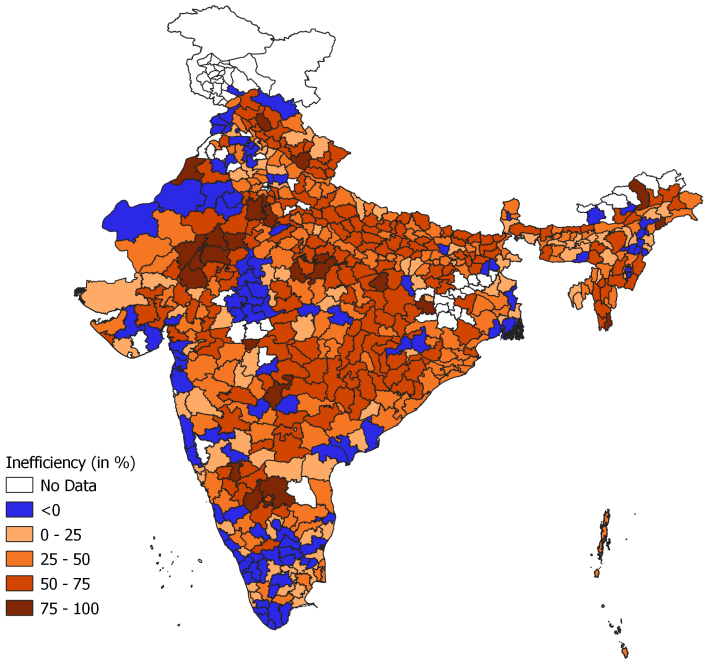
4.1 Baseline

For our baseline scenario, we assume that the planner uses mixed water supply sources and intermediate level of agricultural inputs. We report our baseline results at different administrative levels – districts, states and country.

4.1.1 Districts as administrative units

In this exercise, the planner solves the problem district-by-district, minimizing the land required to attain the actual output of all crops in a district, using the average potential GAEZ yields in that district. Figure 3 plots the inefficiency index in each district in allocating the Kharif crops.

Figure 3: Inefficiency across districts for Kharif season



The districts shaded blue use less area relative to the planner. About 18 percent of all districts fall in this category (104 districts out of a total of 565[22]). Because the planner's allocation is the most efficient in terms of minimizing total land required for cultivation given a certain level of human inputs, one possibility is that these districts use higher level/quality of human inputs than what we have assumed in the benchmark scenario. In fact, the share of such districts drops to 8 percent when we solve the planner's problem assuming irrigated water conditions everywhere (see Appendix). The other possibility is some sort of selection, a possibility we examine in the next section.

The remaining, that is around 82 percent of the districts, use more land than the planner. Because our assumption about the planner's access to inputs is conservative, it is more likely that these districts have access to similar inputs as the planner, but allocate land inefficiently.[23] The most inefficient amongst these districts are located in and around Central India, in the south-eastern parts of Rajasthan, and also in the northern hilly regions of Himachal Pradesh and Uttarakhand. Most districts in the North-Eastern part of the country are inefficient as well.

[22]The GAEZ data covers 618 districts out of which 575 are left after data cleaning. Because there exists no feasible planner's solution for 10 of these districts, we are left with 565.

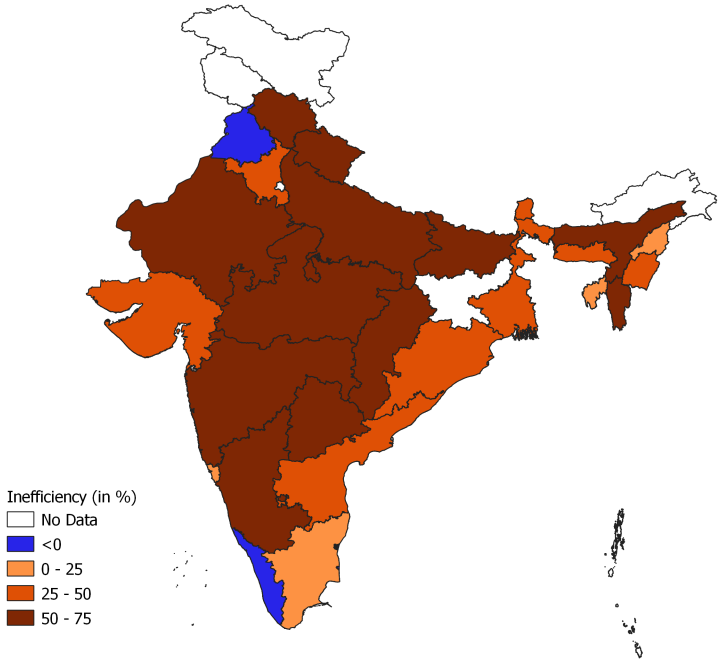
[23]The amount of land that we assume to be irrigated when solving the planner's problem is significantly less than what has been documented elsewhere (Jain et al., 2019).

Which states are the most and least efficient in allocating crops across heterogeneous fields? If we sum up the actual and optimal areas at the state level and re-calculate inefficiency, Kerala in the southern part of the country turns out to be the most efficient, while the least efficient is Himachal Pradesh in the northern part of the country. Table 6 in the Appendix contains the result for all states.

In total, the actual area used for producing Kharif crops is around 64.3 million hectares, whereas the optimal area is around 43.6 million hectares. Hence, roughly 20.7 million hectares (32 percent) of agricultural land can be freed up if crop allocation is done optimally at the district level.

4.1.2 States as administrative units

Figure 4: Inefficiency across states during Kharif season



In the next exercise, we let the planner minimize the land required to produce the state-level output of each crop. Figure 4 plots the inefficiency index in each state in producing and allocating the Kharif crops. Compared to the previous scenario, the planner no longer needs to meet production targets at the district level; this allows the planner to use less land. To see this, suppose a district A has non-zero production of a crop X. Furthermore, assume that the productivity

of X in A is low. In the previous exercise, the planner would be forced to allocate some fields in A towards producing X. In the current exercise, however, the planner could allocate relatively more productive fields in other districts to X.

The above exercise can potentially free up 36.8 million hectares (47 percent) of agricultural land[24], 77 percent more than the 20.7 million hectares in case of the district-level optimization. In this exercise, the state of Punjab is the most efficient followed by Kerala, while Himachal Pradesh is still the least efficient.

4.1.3 India as a single administrative unit

Finally, we allow the planner to minimize the land required to produce the total output of each crop at the country-level. The total actual area is 79.9 million hectares while the total optimal area is 34.2 million hectares. Accordingly, 45.7 million hectares (57 percent) of agricultural land can be freed up by a planner who can potentially reallocate the production of any crop to any field across the country.

Observe that our baseline scenario assumes the same intensity of inputs and the same water supply source across fields. In reality, there could be some variation in input usage as well as water sources across fields. Furthermore, we use the entire area of a region to solve the planner's problem. Again, in reality, every field in a region may not be used for agriculture; while some fields may not be available for agriculture, farmers may optimally choose not to farm in parts of other fields. Next, we turn to the selection issue.

4.2 Positive Selection

The potential yield of a GAEZ field is an average over 100 sub-fields that are contained in it.[25] Given that each GAEZ field is large with an average area of around 80 sq. kilometers (in India), the possibility of potential yields varying across sub-fields within a given field cannot be ruled out. With heterogeneous potential yields within a GAEZ field, if the planner does not use the entire field, the effective yield for land that is *actually* cultivated could be different from the average potential yield.[26] We explore this possibility next.

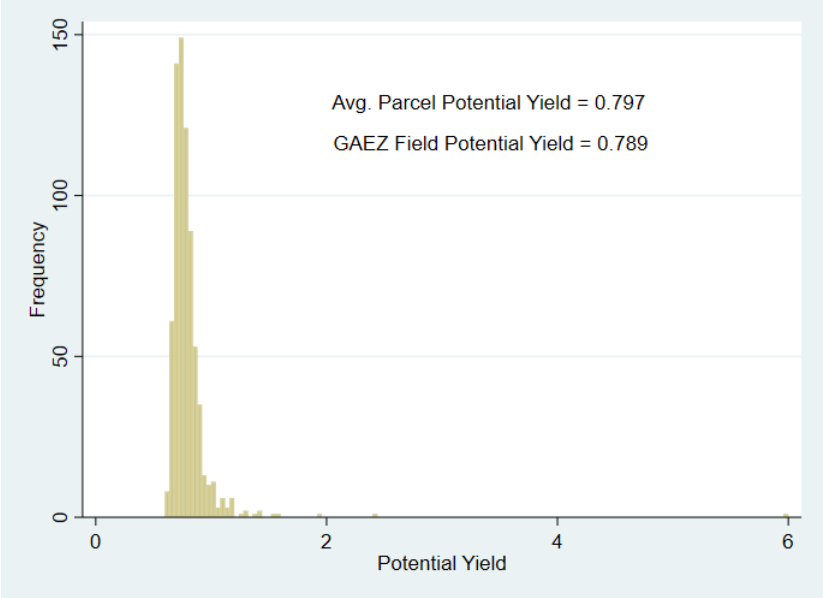
We model selection by assuming that each GAEZ field f in region i consists of a continuum of land parcels indexed by ω. The corresponding productivity of such a parcel is $A_i^{fk}(\omega)$. Following Sotelo (2020), we assume that the productivity of these parcels is i.i.d Frećhet such that the

[24] Total actual area = 77.6 million hectares, Total optimal area = 40.8 million hectares.

[25] A GAEZ field has a resolution of 5 arc-minutes while a sub-field has a resolution of 30 arc-seconds. An arc-second being $\frac{1}{60}$th of an arc-minute, thus a GAEZ field subsumes 100 such sub-fields.

[26] It is reasonable to assume that only a fraction of land within each GAEZ field is used for cultivation. Based on the figure reported in the introduction, an average of 40 percent land within each GAEZ field in India is *not* used for cultivation.

Figure 5: Frequency distribution of parcel potential yields for barley in a GAEZ field



distribution of $A_i^{fk}(\omega)$ is given by

$$Prob(A_i^{fk}(\omega) \leq a) = exp\Big(-\gamma^{\theta}(A_i^{fk})^{\theta} a^{-\theta} \Big). \tag{6}$$

Here, A_i^{fk} is the GAEZ reported average potential yield of field f in producing crop k, θ is an inverse measure of land heterogeneity amongst parcels[27], and γ is a normalization such that $\mathbb{E}[A_i^{fk}(\omega)] = A_i^{fk}$.[28]

The above specification suggests that parcels within a field could be heterogeneous in terms of their potential yields. As a result, a farmer or, in our case, the planner may not use the entire field to grow a given crop. To solve this problem, we divide each GAEZ field f into equal-sized discrete parcels of land. Then we build a vector of crop productivity for each parcel by averaging over 500 independent draws. Figure 5 shows the frequency distribution of parcel-level potential yields for the barley crop within a GAEZ field. The field-level potential yield (0.789) is very close to the mean (0.797) of parcel-level potential yields.[29] At the same time, there is substantial variation within the field, with the ratio of 90^{th} to 10^{th} percentile yield being equal to 1.34.

Next, we solve the planner's problem with parcels instead of fields. Specifically, the planner chooses how much of each parcel to use for production so as to meet the district level actual

[27] We use $\theta = 1.658$ as calibrated in Sotelo (2020).

[28] $\gamma = \left[\Gamma\left(1 - \frac{1}{\theta}\right)\right]^{-1}$ where $\Gamma(.)$ is the Gamma function.

[29] It is important that the parcel-level potential yields are generated in a way such that the average across the parcels for any crop is close to the corresponding average for that field obtained from GAEZ. This ensures that the productivity of sampled parcels is representative of the GAEZ field-level productivity in each crop. We find this to be true in our simulations.

output of every crop. We see improvement both in terms of: (i) lesser number of districts having higher efficiency than the planner, and (ii) lesser area used by the planner, relative to the baseline case. In particular, the fraction of districts where the actual area is less than the optimal drops from 18 percent to 13 percent. At the same time, the total area that can be freed up is 28.4 million hectares,[30] 37 percent more than the baseline scenario. Intuitively, with the optimization problem at the level of parcels, the planner can choose to cultivate only the most productive parcels within a field.

4.3 Negative Selection

In the above exercise, we assumed that every GAEZ field can potentially be used by the planner. That may not be the case, however, as some of the fields may not be available for a variety of reasons. From a policy perspective, accounting for such negative selection is crucial. This is the issue that we turn to next.

Realistically, a planner would not be able to use parts of or entire GAEZ fields which have non-farm uses such as residential and commercial clusters, or are covered with forests, water bodies, etc. Figure 7 in the Appendix shows such areas[31] carved out of each GAEZ field.

Accounting for such negative selection reduces total available land area from 297 million hectares to 236 million hectares, i.e., almost a 20 percent reduction.[32] We proceed to solve the planner's problem as above (with positive selection) with the additional constrain that parts of certain fields may not be available. In this scenario, 25.2 million hectares (39 percent) of land can be released.[33] Inefficiency falls for each district (due to higher optimal area) and each state (see Table 6). The relative ranks also change as negative selection is not uniform across different parts of the country. At the state level, Mizoram is the least efficient although Kerala continues to be the most efficient.

In the next section, we describe a two-sector general equilibrium model based economy. Later, we also compute the new competitive equilibrium of this economy where the agricultural productivity increases by an amount equivalent to reallocating land to optimal crops.

5. Two Sector Model

In this section, we consider an economy with 2 production sectors – agriculture and manufacturing. These use land and labor as input factors. There is a representative household which

[30] Total actual area = 65.3 million hectares and Total optimal area = 36.9 million hectares

[31] Non-farm uses include allotments, cemetery, commercial, forest, grass, meadow, health, industrial, military, nature reserves, orchard, park, quarry, recreation ground, residential, retail, scrub, vineyard.

[32] The total non-farm land use area is around 90 million hectares but some of it falls under regions such as Jammu & Kashmir which are not included in the GAEZ dataset.

[33] In this case, total actual area = 64 million hectares and total optimal area = 38.8 million hectares.

consumes agricultural as well as non-agricultural goods, and rents out both land and labor to the 2 sectors. There exists barriers in the mobility of both factors – land and labor – from agriculture to manufacturing. This gets reflected in the factor price differentials between the 2 sectors. In the next section, we will evaluate how the factor shares and prices change in both sectors if there is a change in the productivity of agriculture sector. Also, we report the GDP and real income changes in the whole economy.

5.1 Production Technologies

We adopt Cobb-Douglas production functions for both sectors. In the agriculture sector, it is given by:

$$Y_a = X^{\alpha}[(A_a Z_a)^{1-\sigma} {L_a}^{\sigma}]^{1-\alpha}, \;\; 0 < \sigma < 1, 0 < \alpha < 1, \tag{7}$$

where the subscript a denotes agriculture (whereas m denotes manufacturing). Y_a is the agricultural output, X is the intermediate input (a manufacturing good), Z_a and L_a denote land and labor respectively, and A_a is the land-augmenting productivity.

The production in manufacturing sector follows

$$Y_m = A_m {L_m}^{\beta} {Z_m}^{1-\beta}, \;\; 0 < \beta < 1, \tag{8}$$

where Y_m, L_m, Z_m denote the output, labor and land used in manufacturing sector, respectively. Our formulation of production in both sectors excludes capital as an input factor which means that it is considered a part of the sectoral productivity.

The first-order order conditions for the agriculture sector are given by:

$$X = (\alpha P)^{\frac{1}{1-\alpha}} (A_a Z_a)^{1-\sigma} {L_a}^{\sigma} \tag{9}$$

$$\frac{Z_a}{L_a} = \left[\frac{(1-\sigma)(1-\alpha)(\alpha)^{\frac{\alpha}{1-\alpha}} P^{\frac{1}{1-\alpha}} (A_a)^{1-\sigma}}{r_a} \right]^{\frac{1}{\sigma}} \tag{10}$$

$$P = \left[\frac{w_a {r_a}^{\frac{1-\sigma}{\sigma}}}{E {A_a}^{\frac{1-\sigma}{\sigma}}} \right]^{\sigma(1-\alpha)} \tag{11}$$

where P is the price of agricultural good. We treat the non-agricultural good as a numeraire, so all prices are measured in terms of non-agricultural output. r_a, w_a denote the land rent and labor wages in the agriculture sector, respectively. And, E is a function of parameters.[34]

In the manufacturing sector, profit-maximization yields the following:

[34] $E = \sigma(1-\sigma)^{\frac{1-\sigma}{\sigma}} (1-\alpha)^{\frac{1}{\sigma}} \alpha^{\frac{\alpha}{\sigma(1-\alpha)}}$

$$\frac{Z_m}{L_m} = \left(\frac{1-\beta}{\beta}\right)\frac{w_m}{r_m} \tag{12}$$

5.2 Barriers

As mentioned above, the movement of both land and labor is not frictionless from the agriculture to the manufacturing sector. Such that the factors are difficult to reallocate. These barriers are encapsulated in the factor price differentials between the 2 sectors as follows:

$$w_m = \frac{w_a}{1-\theta}, \;\; 0 < \theta < 1 \tag{13}$$

$$r_m = \frac{r_a}{1-\phi}, \;\; 0 < \phi < 1 \tag{14}$$

The wage gap between agriculture and manufacturing sectors can be very closely mapped into the spatial wage differences between rural and urban areas in India. For instance, as per NSSO 68th round data for 2011-12, only 56 male workers out of every 1000 in urban areas work in the primary sector whereas 594 are engaged in primary activities in rural areas. Munshi and Rosenzweig (2016) document the rural-urban wage gap at greater than 25 percent for decades after correcting for cost-of-living differences. They attribute this persistent spatial wage gap to workers' inability to move to urban areas. Because, (i) they do not have access to formal insurance – government safety nets and private credit – in urban areas, and (ii) they have access to well-functioning caste-based informal insurance networks in rural areas. So that male workers of households which face greater rural income risk migrate less as they can benefit more from the rural insurance networks.

In the land market in India, all states impose upper limits on the amount of agricultural land that a person or entity can buy. This is due to the Land Ceiling Acts enacted to implement land redistribution reforms after Independence Bolhuis et al. (2021). This effectively restricted sale of agricultural land in India. Additionally, in many states like Maharashtra, Gujarat, Karnataka etc., only agriculturists could buy agricultural land[35]. That is why, government intervention in the form of Land Acquisition Act was required (LARR Act, 2011). This raises the transactions costs and regulations for industrial development in the country.

5.3 Preferences

Following Restuccia et al. (2008), we assume that the household preferences are non-homothetic in nature and given by a Stone-Geary utility function as follows:

[35] Source: https://www.thehindu.com/features/homes-and-gardens/Land-laws-across-India/article14414630.ece

$$U = alog(c_a - \overline{a}) + (1-a)log(c_m), \;\; 0 \leq a < 1, \tag{15}$$

where c_a, c_m represent the consumption levels of agricultural and manufacturing goods, $\overline{a}$ denotes the subsistence level of agricultural consumption, and a is the utility weights of the 2 consumption goods.

The household's income gets derived from 2 sources – labor wage and land rent. So its budget constraint looks as follows:

$$y = Pc_a + c_m = w_a L_a + w_m L_m + r_a Z_a + r_m Z_m, \tag{16}$$

where y denotes the household income. The utility maximization subject to budget constraint gives the following:

$$c_a = \overline{a} + P^{-1} a[y - P\overline{a}], \tag{17}$$

$$c_m = (1-a)[y - P\overline{a}] \tag{18}$$

The above equations imply that after paying for $\overline{a}$ amount of agricultural good, the household allocates its remaining income $(y - P\overline{a})$ between the 2 goods proportional to their utility weights.

5.4 Market Clearing

All markets in the economy need to clear for the equilibrium to exist. The clearing conditions are given by:

$$N = L_a + L_m,$$

$$Z = Z_a + Z_m,$$

$$Y_a = Nc_a,$$

$$Y_m = Nc_m + X.$$

The first 2 equations represent the labor and land market clearing conditions such that N, Z denote the total labor and total land in the economy, respectively. Whereas, the last 2 equations ensure that both goods markets clear.

5.5 Competitive Equilibrium

The competitive equilibrium of the economy consists of: (i) allocations $(L_a, L_m, Z_a, Z_m, X, c_a, c_m)$, and (ii) prices (P, w_a, w_m, r_a, r_m), such that, (i) the first-order conditions of both sectors satisfy,

(ii) household is able to maximize its utility, and (iii) all the market clearing conditions hold. After doing some algebra, we get the following equilibrium conditions.

$$\frac{L_a}{N} = \frac{\frac{Z_a}{Z}}{G(1-\frac{Z_a}{Z})+\frac{Z_a}{Z}}, \text{ where } G = \frac{(1-\sigma)\beta(1-\theta)}{\sigma(1-\beta)(1-\phi)} \tag{19}$$

$$P^{1/(1-\alpha)} = \frac{CA_m}{A_a^{1-\sigma}}\left(\frac{Z_a}{Z}\right)^{\sigma}\left(\frac{L_a}{N}\right)^{-\sigma}\left(\frac{Z}{N}\right)^{\sigma-\beta}\left(1-\frac{L_a}{N}\right)^{\beta}\left(1-\frac{Z_a}{Z}\right)^{-\beta} \tag{20}$$

where $C = \frac{(1-\beta)(1-\phi)}{(1-\sigma)(1-\alpha)(\alpha)^{\alpha/(1-\alpha)}}$, and

$$y = r_a N\left[\frac{w_a}{r_a}\left(\frac{1-\theta\frac{L_a}{N}}{1-\theta}\right)+\left(\frac{Z}{N}\right)\frac{(1-\phi\frac{Z_a}{Z})}{1-\phi}\right] \tag{21}$$

Further, we can rewrite the agricultural output per worker as follows:

$$\frac{Y_a}{L_a} = (A_a)^{1-\sigma}\left(\frac{X}{Y_a}\right)^{\frac{\alpha}{1-\alpha}}\left(\frac{Z_a}{Z}\right)^{1-\sigma}\left(\frac{L_a}{N}\right)^{\sigma-1}\left(\frac{Z}{N}\right)^{1-\sigma} \tag{22}$$

From (9), we know that $\frac{X}{Y_a} = \alpha P$. Using the above equations, once we know the share of factor allocations in the agriculture sector i.e., $\frac{Z_a}{Z}$ and $\frac{L_a}{N}$, we can easily compute the prices (P, w_a, r_a), the share of intermediate input X/Y_a, and values of other output and consumption variables (Y_a, Y_m, c_a, c_m).

6. Quantitative Analysis

6.1 Calibration

Starting with the production parameters, we compute the share of intermediate inputs in agricultural output i.e., α, as the average value of input intensity. We use the IHDS 2011-12 (India Human Development Survey) dataset. It is a household survey data for 42,152 households. We calculate the input intensity of each household which cultivates crops during the year. It is calculated as the ratio of the value of agricultural inputs[36] used by the household and its crop income. There are 15,755 households with non-zero input intensity. This results in $\alpha = 0.4074$. Following Restuccia et al. (2008), we use $\sigma = 0.7$. For the manufacturing sector, we use the factor shares for Indian manufacturing sector as documented in Duranton et al. (2015). This gives $\beta = 0.8194$.[37]

Next, we compute the preference parameters. Observe from equation (17) that the utility

[36] Following inputs are considered: seeds, fertilizers, pesticides, irrigation water, and new farm equipment. If we include hired equipment/animal in the list, then $\alpha = 0.5365$

[37] In Duranton et al. (2015), share of capital (total assets) is 0.41 which means that share of labor is $1 - 0.41 = 0.59$. Whereas, share of land & building is 0.13. We normalize the sum of shares of labor and land (building) to 1 so that capital is a part of the productivity.

weight of agricultural consumption, i.e. a, is the share of food expenditure in household income when the income becomes very large. We use the household monthly consumption data for year 2015 from CMIE (Center for Monitoring Indian Economy) dataset[38] to compute the share of food expenditure in income (averaged over months) for the top 0.1% households. It gives $a = 0.0515$.

Coming to the barrier parameters. The CMIE dataset also reports the monthly income of each household member from wages. We use the wages of individual members who work as "Agricultural Laborers" and "Industrial Workers" as agricultural and manufacturing wages, respectively. Their ratio gives us the value of θ as 0.564. In order to find ϕ, we use the values of factor shares used in agriculture sector in year 2011-12 and plug them in equation (19) along with the value of θ. As per the Agriculture Census of 2010-11, the percent of total land area under cultivation was 53.68%. After subtracting the area under forests (26.33%), deserts (6.92%) and water bodies (2.35%) (other than rivers and canals)[39], the share of remaining land area under cultivation then becomes 83.35%. That is, $\frac{Z_a}{Z} = 0.8335$. According to Census 2011, the share of total labor force in agriculture sector (cultivators & agricultural laborers) was 70.68%. Using these values, we get $\phi = 0.59$. It indicates that the land market is more restrictive compared to the labor market in India.

Lastly, we calibrate $\overline{a}$. Using equation (21) and agricultural good market clearing equation, we get the following expression:

$$\overline{a} = \frac{1}{1-a}\left[\frac{Y_a}{L_a}\frac{L_a}{N} - \frac{ay}{P}\right] \tag{23}$$

After plugging in the right hand side, we get $\overline{a} = 0.4538$. Finally, for simplicity, we normalize the values of A_a, A_m, Z, N to unity.

6.2 Quantitative Results

Here, we examine the impact of reallocating agricultural land to manufacturing sector. Consider the impact of reallocating land at the country level using the results obtained in section 4.3. The negative selection case is the most practical given the non-availability of some areas to a Planner in reality. Also, using districts as administrative units is the most local and hence the most feasible reallocation exercise compared to the ones at State or India level. We assume a single model to consider the aggregate impact therefore we abstract from the heterogeneity across districts in terms of economic parameters. Instead of literally reallocating land, we rather consider an equivalent increase in agricultural productivity using equation (22) such that agricultural output per worker ($\frac{Y_a}{L_a}$) remains unchanged. Then, under this increased agricultural productivity,

[38] The dataset starts from year 2014. It covers same 1,58,624 households each month.

[39] Sources: Forests (Forest Survey Report, 2011 published by Forest Survey of India under Ministry of Environment & Forests), Deserts (http://ppqs.gov.in/), Water Bodies (http://jalshakti-dowr.gov.in/)

we compute the new equilibrium of the economy by solving a set of non-linear equations.

Table 3: Percentage changes in variables going from old to new equilibrium

Variable	A_a	$\frac{Z_a}{Z}$	$\frac{L_a}{N}$	P	w_a	r_a	GDP	Real Income
% change	97.72	−6.88	−11.51	−11.73	0.92	−4.09	7.07	9.47

Note: Real Income calculated using the price-index described in Appendix D. GDP is calculated as follows: $GDP = PY_a - X + Y_m$.

Table 3 reports the percent changes in variables when agricultural productivity is exogenously increased by 97.72%[40] Now, agriculture sector uses lesser share of each input factor – land as well as labor – along with lesser intermediate inputs. Correspondingly, manufacturing sector uses a higher share of both the factors with a greater absorption of labor than land. As a result, more land is available per worker in the agriculture sector which drives the land price down. Since the share of labor in output is almost twice in manufacturing sector than the agriculture sector i.e., $\beta > \sigma(1-\alpha)$, therefore labor becomes more productive by moving into the manufacturing sector resulting in a wage increase. We can also see this mathematically using equations (10) and (11),

$$\frac{Z_a}{L_a} \propto w_a^{(1-\alpha)}\, r_a^{\frac{(1-\sigma)(1-\alpha)-1}{\sigma}}$$

Because the rise in wages (0.92%) is not enough to cater to the increase in land per labor change (5.23%), therefore the land price r_a needed to go down. Further, with increased productivity, production of agricultural output increases whereas its price comes down. Real income rises due to: (i) increase in nominal income on account of higher labor wages, and (ii) decrease in the economy-wide price index (see Appendix D). Finally, more manufacturing output is produced with more input factors such that there is an increase of 28.93%. Though the agricultural GDP $(PY_a - X)$ decreases in the new equilibrium but the overall GDP increases by 7.07%.

In the next section, we report the equilibrium changes in counterfactual economies which have different pairs of (θ, ϕ) values than the calibrated values used in the current section.

6.3 Sensitivity to Barriers

In this section, we compare the responses of calibrated economy in the above section with those of a counterfactual one with different factor barriers. Table 4 reports the percentage changes in counterfactual economies shocked with an identical exogenous productivity increase in agriculture sector. Each economy is characterized by a different pair of barrier parameters (θ, ϕ). The

[40] Equivalent to using 39.32% lesser agricultural land.

Table 4: Percentage changes in variables – using different (θ, ϕ)

Variable	A_a	$\frac{Z_a}{Z}$	$\frac{L_a}{N}$	P	w_a	r_a	GDP	Real Income
1) $\theta = 0.564, \phi = 0.51$	97.72	-7.67	-11.34	-11.67	0.73	-3.27	7.21	9.71
2) $\theta = 0.564, \phi = 0.8$	97.72	-4.46	-11.88	-11.92	1.47	-6.40	6.27	8.30
3) $\theta = 0.543, \phi = 0.59$	97.72	-6.66	-11.67	-11.76	1.0	-4.43	7.08	9.60
4) $\theta = 0.8, \phi = 0.59$	97.72	-9.09	-8.86	-11.40	-0.04	0.2	6.17	7.48

Note: Real Income calculated using the price-index described in Appendix D. GDP is calculated as follows: $GDP = PY_a - X + Y_m$. Initial GDP is same in each case. The choice of lower bounds for (θ, ϕ) driven by the requirement that (i) initial GDP equals that of the calibrated economy, and (ii) $\overline{a}$ calculated using equation (23) satisfies $\overline{a} < c_a$.

value of factor shares and prices in the initial equilibrium of each economy is such that the GDP is same as in the above section.

In the first 2 cases, we fix θ at its calibrated value and change only ϕ. In case 1, compared to the case in previous section, there is a greater movement of land from agriculture to manufacturing due to lower land barriers. Such that the land prices do not go down as much because there is relatively lesser land per worker. Also, wages do not increase as much since the labor movement is a bit lower. On the net, both GDP and real income increase by a higher percentage. On the other hand, in case 2, the movement of land gets more difficult with a higher value of ϕ such that the land prices decrease even more (compared to the calibrated case) as the land per worker becomes excessive. At the same time, there is a greater labor movement to manufacturing such that the wage-increase is higher due to greater shift of labor now. But, both GDP and real income decrease with respect to the calibrated case.

Next, we fix the value of ϕ and change only θ. When θ is reduced in case 3, the labor moves more freely while the land shift decreases a bit. This results in a higher wage rise and a higher land-price decrease. Such that there is a slight improvement in both GDP and real income relative to the calibrated case. Lastly, when labor is made more restricted to agriculture, a higher land movement to manufacturing compensates for the decrease in labor shift. Such that, now there is lesser land per worker in agriculture resulting in an increase in land price. Whereas, the wages decrease as more labor is stuck in the unproductive agriculture sector. As a result, increase in both GDP and real income get reduced.

To summarise, we note the following points through the above counterfactual exercise. First, the positive effect of land reallocation to manufacturing sector (through agricultural productivity shock) gets dampened if either barriers are higher. Whereas, the effect amplifies if the barriers are lower. Second, the lower movement of the factor facing higher barriers is compensated by a higher movement of the other factor. This relative movement affects the factor prices.

7. Conclusion

This paper answers the following question: how much land can be freed up for non-agricultural usage if agricultural land is allocated to its best use? For a primarily agrarian country like India, this question holds acute relevance for policymakers. We find that there is indeed potential for releasing agricultural land subject to meeting the actual production of all crops in India[41]. Our baseline results present a puzzle which prompted us to ask a methodological question: why are there so many districts performing better than a planner in growing/allocating agricultural crops? We find that the literature using GAEZ dataset had so far neglected an important issue of selection. Given the heterogeneous land quality within a large GAEZ field and the fact that not all land is used in agriculture, the effective yield could very well be different from the average yield if the entire field were used. In the results section, we allow access to heterogeneous productivity parcels to the Planner and confirm our selection hypothesis.

Lastly, we assess the quantitative impact of agricultural productivity increase, equivalent to reallocating land to manufacturing sector, using a 2-sector economic model. We find significant improvement in terms of both GDP and real income increase in the economy. However, such positive effects get diminished by the barriers faced by either factor of production – land or labor – in moving from one sector to another. Therefore, we also examine the economic importance of doing the reallocation exercise for which we find answer in the preceding sections, in the presence of different barrier levels.

[41] This is especially true for Kharif season crops which are both grown and allocated sub-optimally compared to the Rabi season crops.

References

Adamopoulos, Tasso and Diego Restuccia, "Geography and agricultural productivity: Cross-country evidence from micro plot-level data," Technical Report, National Bureau of Economic Research 2018.

Bolhuis, Marijn A, Swapnika R Rachapalli, and Diego Restuccia, "Misallocation in Indian agriculture," Technical Report, National Bureau of Economic Research 2021.

Caselli, Francesco, "Accounting for cross-country income differences," *Handbook of economic growth*, 2005, *1*, 679–741.

Costinot, Arnaud and Dave Donaldson, "How large are the gains from economic integration? Theory and evidence from US agriculture, 1880-1997," Technical Report, National Bureau of Economic Research 2016.

_ , _ , and Cory Smith, "Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world," *Journal of Political Economy*, 2016, *124* (1), 205–248.

Duranton, Gilles, Syed Ejaz Ghani, Arti Grover Goswami, William Kerr, and William Robert Kerr, "The misallocation of land and other factors of production in India," *World Bank Policy Research Working Paper*, 2015, (7221).

Frolking, Steve, Jagadeesh Babu Yeluripati, and Ellen Douglas, "New district-level maps of rice cropping in India: a foundation for scientific input into policy assessment," *Field Crops Research*, 2006, *98* (2-3), 164–177.

Gollin, Douglas, Stephen Parente, and Richard Rogerson, "The role of agriculture in development," *American economic review*, 2002, *92* (2), 160–164.

Jain, Rajni, Prabhat Kishore, and Dhirendra Kumar Singh, "Irrigation in India: Status, challenges and options," 2019.

Mohanty, Nirmal, Runa Sarkar, and Ajay Pandey, "India Infrastructure Report 2009: Land–A Critical Resource for Infrastructure'3iNetwork," 2009.

Munshi, Kaivan and Mark Rosenzweig, "Networks and misallocation: Insurance, migration, and the rural-urban wage gap," *American Economic Review*, 2016, *106* (1), 46–98.

Ranganathan, V, "Challenges of Land Acquisition," 2010.

Restuccia, Diego, Dennis Tao Yang, and Xiaodong Zhu, "Agriculture and aggregate productivity: A quantitative cross-country analysis," *Journal of monetary economics*, 2008, *55* (2), 234–250.

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

Appendix

A. Data Appendix

A.1 Irrigation Data

GAEZ dataset also provides the geographical distribution of field-level actual irrigation scenario in India in year 2011. It reports in binaries – "1" if more than 50% area of a GAEZ field is under irrigation and "0" otherwise. Figure 6 shows the irrigated GAEZ fields on a map. These comprise only 13.5% of the total geographical area of India. In the baseline scenario, we assume that Planner uses irrigation water supply and intermediate level of complementary inputs in those fields where irrigation facilities are available. And, he uses rain-fed water conditions and intermediate level of complementary inputs in the remaining fields. Because more than 40 percent of net sown area in India was irrigated in 2011-12, our baseline results serve as a lower bound for the extent of misallocation of land in agriculture (Jain et al., 2019).

A.2 Data Cleaning

We merge the district boundaries with both the potential yields and actual yields data for 2011. Each observation is then at the crop-district level. We clean the merged data following these steps:

1. The potential yield is reported as negative or zero in a field in which there is no parcel of land suitable to grow a particular crop. If all the fields in a district are unsuitable for growing a particular crop, then we drop all such crop-district observations.

2. If the (actual) area of production for a crop in a district is greater than the maximum available GAEZ area for producing that crop in that district, then we drop this crop-district observation.

3. If the volume of production for a crop in a district is greater than the maximum volume based on GAEZ yields for producing that crop in that district, then we drop this crop-district observation.

4. If the total area of production for *all* crops in a district is greater than the maximum available GAEZ area for that district, then we drop this district.

The cleaned data is used for solving the planner's problem. Because some districts as well as crops get dropped after the above steps, we observe many districts labelled as "No Data" in the maps.

Table 5: Crop Yields (in tonnes per hectare)

Crop	All-India Yield		Average District Yield	
	Kharif	Rabi	Kharif	Rabi
Banana	10.42	10.59	10.82	10.58
Barley	1.39	2.21	1.69	1.74
Coconut	2.10	2.10	2.24	2.24
Cotton	0.14	0.096	0.10	0.13
Gram	0.71	0.87	0.52	0.88
Groundnut	0.78	1.16	0.81	1.05
Maize	1.91	3.40	1.79	2.48
Onion	1.56	2.90	1.82	2.11
Pearl millet	0.98	1.28	1.14	1.45
P Beans	4.63	0.94	4.62	1.24
Pigeonpea	0.68	1.01	0.79	0.89
Potato	2.68	5.52	2.54	3.21
Rapeseed	0.36	1.06	0.36	0.76
Sorghum	0.95	0.72	0.96	1.12
Sugarcane	6.57	8.06	5.47	5.56
Sunflower	0.52	0.56	0.99	0.96
Sweet potato	2.74	1.67	2.56	2.09
Tobacco	1.45	1.85	1.51	1.79
Wheat	1.30	2.66	1.16	1.96
Wetland Rice	2.06	1.80	1.85	1.69
Soybean	1.01	NA	0.99	NA

Table 6: State-level Inefficiency (in %) using districts as administrative units

State	Baseline	Positive Selection	Negative Selection
Andaman & Nicobar Island	44.97	80.34	64.57
Andhra Pradesh	38	50.69	46.28
Arunachal Pradesh	59.44	74.45	70.69
Assam	32.07	58.57	56.13
Bihar	49.37	76.95	76.6
Chandigarh	−18.79	77.44	41.47
Chhattisgarh	57.22	63.38	58.61
Dadara & Nagara Haveli	−24.03	−16.17	−20.5
Goa	11.25	23.97	1.19
Gujarat	21.89	35.12	37.53
Haryana	13.82	16.98	16.61
Himachal Pradesh	68.74	80.73	76.46
Karnataka	43.48	58.39	57.25
Kerala	−29.5	−11.03	−13.2
Madhya Pradesh	26.05	31.90	16
Maharashtra	37.75	49.46	36.89
Manipur	17.78	35.71	16.19
Meghalaya	19.05	64.74	42.42
Mizoram	56.26	86.52	85.85
Nagaland	−26.48	−2.42	−2.31
Odisha	31.38	52.20	48.84
Puducherry	−34.21	−18.83	−20.2
Punjab	−8.53	−2.80	1.06
Rajasthan	30.66	41.82	50.7
Sikkim	26.38	39.25	37.82
Tamil Nadu	−9.63	11.51	9.82
Tripura	17.96	26.13	17.79
Uttar Pradesh	45.32	54.94	54.66
Uttarakhand	47.08	61.85	59.47
West Bengal	13.68	25.53	24.04

Note: Data is for Kharif season. States of Jammu & Kashmir and Jharkhand are not included as they had very little actual area (147 & 422 hectares respectively) left after data cleaning.

Figure 6: Irrigated (more than 50%) GAEZ fields in India

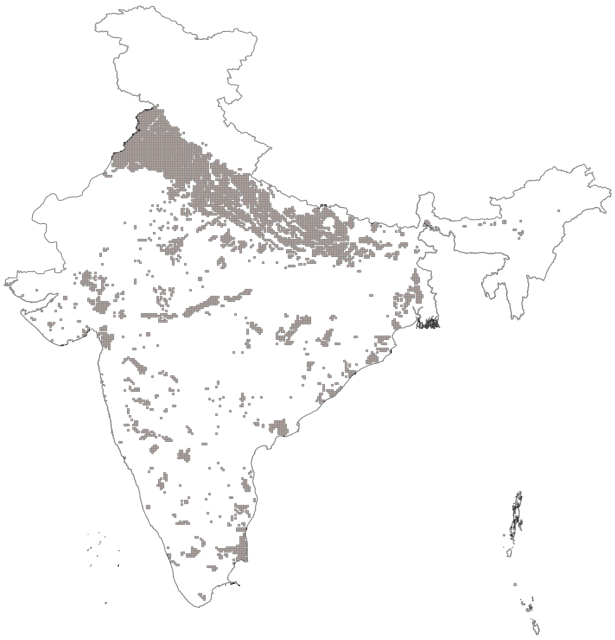
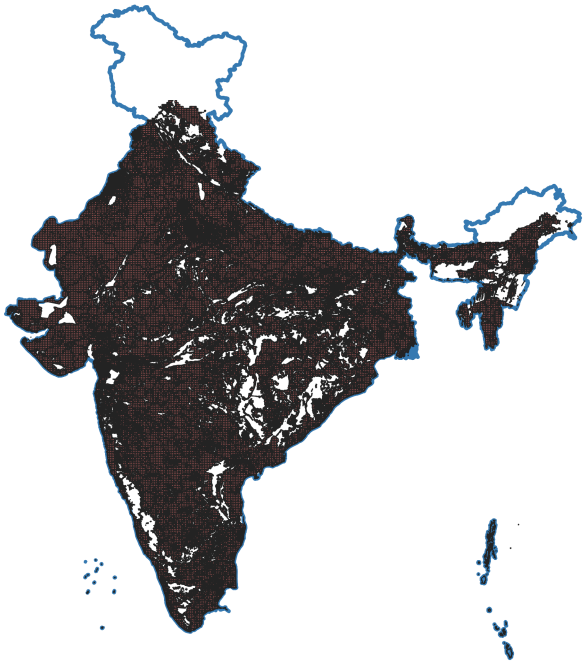


Figure 7: Area left for use by planner after Negative Selection



Source: Land use data from https://www.openstreetmap.org/

B. Results for Rabi season

B.1 Baseline Results

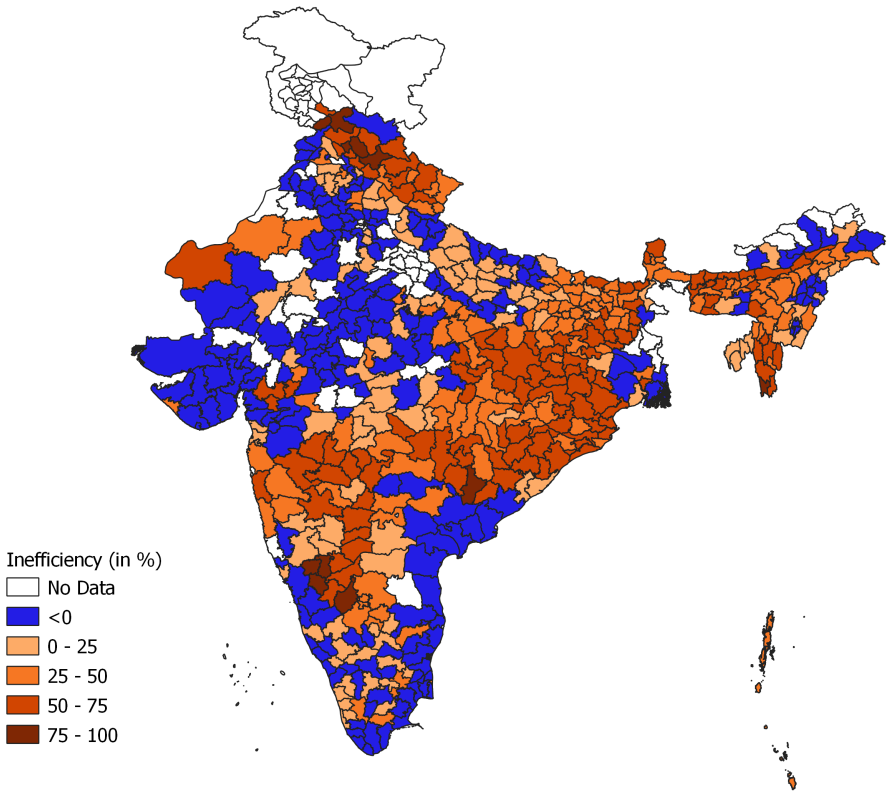


Figure 8: Inefficiency comparison across districts for Rabi season when Planner uses mixed water supply and intermediate inputs

Out of 555[42] districts, 195 i.e. 35% of them are more efficient than the Planner's solution (almost double than Kharif case). As before, it implies that these districts are using better human inputs than the ones assumed by the Planner. This also suggests that majority districts in India use better cultivation techniques for Rabi crops than Kharif crops.

Inefficiency at state level is highest in Himachal Pradesh alike Kharif season and lowest in Haryana. Here, the actual area used is around 48.82 million hectares while the optimal area is 47.5 million hectares. That is, Planner could save only 1.32 million hectares or 2.7% of agricultural land during Rabi season. However, this number increases to 15.05 million hectares (or

[42] Out of 618 districts, 587 were left after data cleaning and further 32 had no feasible solution for Planner's problem.

26.87%) of agricultural land if Planner uses irrigation water supply and intermediate complementary inputs everywhere. Still, it will be smaller than 20.72 million hectares that could be saved during Kharif season under a much lower level of inputs. Note that, 34.61 million hectares (or 51.82%) could be freed during Kharif season if Planner uses irrigation everywhere.

C. Inefficiency results under Irrigated and Intermediate inputs usage by Planner

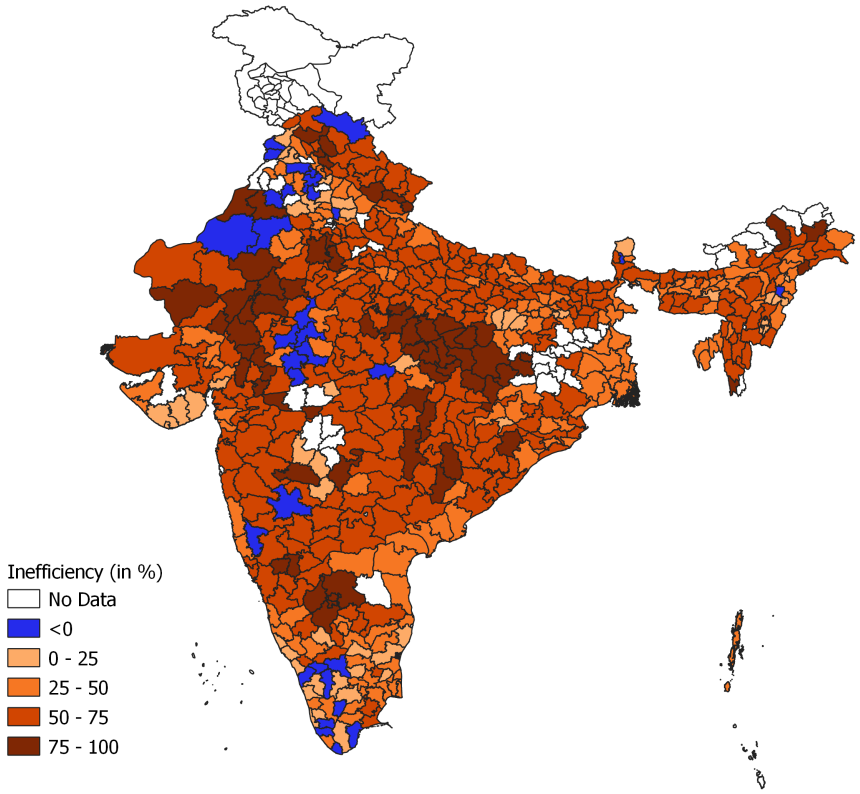


Figure 9: Inefficiency comparison across districts for Kharif season when Planner uses irrigation water supply and intermediate inputs

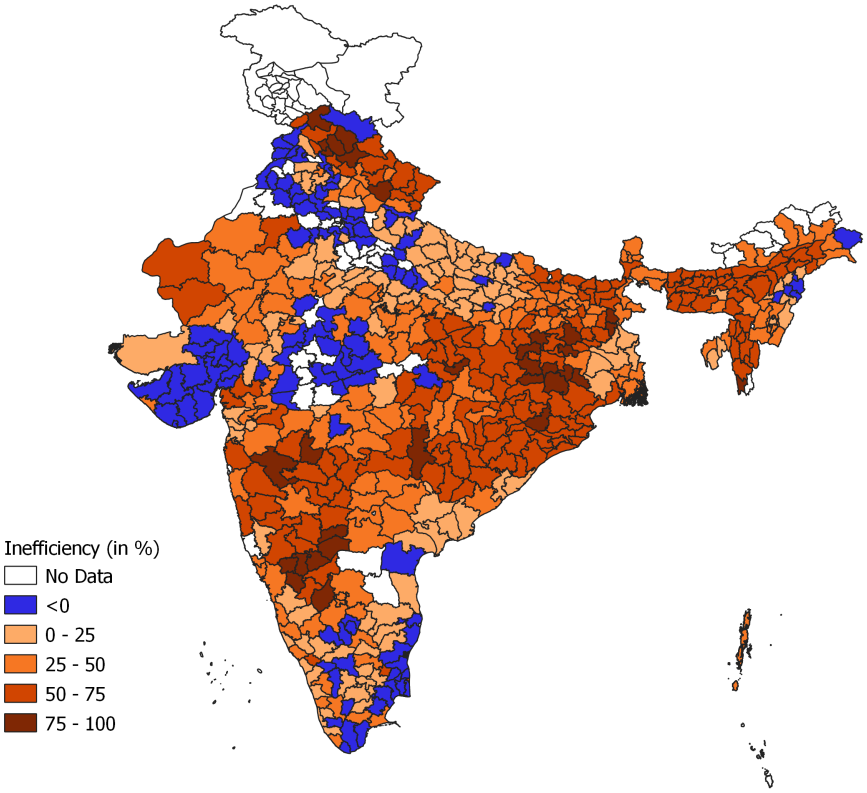


Figure 10: Inefficiency comparison across districts for Rabi season when Planner uses irrigation water supply and intermediate inputs

D. Calculating Price-Index in the economy

To compute the price index of the economy, we solve the following problem:

$$\min_{c_a, c_m} Pc_a + c_m \tag{24}$$

subject to

$$U \geq 1$$

Solving for c_a, c_m, we get

$$c_m = \frac{P(1-a)(c_a - \overline{a})}{a}, \text{ and}$$

$$c_a = exp\left(1-(1-a)log\left(\frac{P(1-a)}{a}\right)\right)+\overline{a}$$

Plugging them back in (24), we obtain the price index.