The Cost of Remoteness: Evidence from 600,000 Indian Villages*

Sam Asher† Karan Nagpal‡ Paul Novosad§

April 18, 2017

PLEASE DO NOT CITE OR CIRCULATE WITHOUT PERMISSION

Abstract

This paper explores the relationship between distance to urban markets and rural economic development. We construct a high-resolution spatial dataset of over 600,000 villages in India, with information on monthly earnings, household assets, employment structure, public goods and distance to urban markets. We find that remoteness continues to be a major and persistent predictor of low living standards. Each additional 10 kilometers from a town is associated with a 3.2% reduction in mean earnings, a 5.2% reduction in non-farm employment, and a 1.9 percentage points decrease in the proportion of literate residents. Comparing these magnitudes with 1991, we find that the slope of the gradients has not diminished, despite 20 years of economic reforms and public infrastructure programs. We then test between different theories that could explain these patterns: lower state capacity (i.e. reduced provision of public goods) and higher costs of participation in urban markets. We find that even after accounting for measures of state capacity, distance to urban markets is negatively correlated with rural living standards. The cost of physical distance from urban markets is both high and persistent.

JEL Codes: O18, R12, H41

*First version August 2016. This version October 2016. We are thankful for useful discussions with James Fenske, Doug Gollin and Beata Javorcik. We are indebted to Taewan Roh and Kathryn Nicholson for exemplary research assistance.

†The World Bank

‡University of Oxford, karan.nagpal@economics.ox.ac.uk

§Dartmouth College

I Introduction

Policy-makers in all countries worry about spatial inequality. A situation in which some parts of the country grow faster than others, or regional differences in productivity are substantial, can have adverse consequences for national efficiency as well as the country's political contract. This challenge is particularly severe in developing countries, where average household consumption in richer regions can be almost 75% higher than in poorer regions of the same country; the corresponding differential for developed countries is less than 25% (The World Bank, 2009). A well-documented dimension in which this inequality gets manifested is the urban-rural wage gap, which in some countries can be as high as 700% after controlling for worker characteristics (Moretti, 2014). Does this inequality also contribute to variation in economic outcomes within rural areas? To investigate this, we assemble socioeconomic microdata on every individual in rural India (825 million observations), which we match to geocoded village data on a range of public good for 600,000 Indian villages. We find that rural wages and living standards are sharply decreasing in distance from urban areas. State capacity also decreases with remoteness, but is unable to account for the full extent of underdevelopment associated with distance from urban areas.

The challenge in such a research design is the absence of good quality and high-resolution spatial data from developing countries. We resolve this issue by assembling Indian data from 4 major administrative datasets. We obtain individual- and household-level data on earnings, assets, housing quality and occupational choices for all rural residents from the Socioeconomic and Caste Census (2012). We match this to village-level demographic data from the Population Census (1991, 2001 and 2011). The Population Census also provides us information on the quality of rural public goods and other measures of state capacity, such as connecting infrastructure. We also obtain village-level data on employment in different economic sectors from the Economic Census (1990, 1998 and 2005). Finally, we obtain geo-

coordinates for all towns and villages in the country and use them to generate straight-line distances for each village to its neighboring towns.

We find that each additional 10 kilometers of distance to the nearest town is associated with a 0.3% reduction in mean monthly earnings in the village. It is also associated with a 4% reduction in non-farm employment, and a 4% increase in the village illiteracy rate. More strikingly, we find evidence to suggest the persistence of this negative relationship between urban distance and rural living standards. Indeed, the magnitudes have increased over time: while we don't observe mean earnings for the early 1990s, we find that an additional distance of 10 kilometers to the nearest town was associated with a 3% reduction of non-farm employment in 1990 and a 2.3 percentage points increase in the proportion of village adults who were illiterate. We then test between alternate theories to explain these patterns. We find that measures of both connecting infrastructure (roads, railways) as well as public goods (electricity, primary schools, medical centers, public sources of irrigation) are strictly decreasing in urban distance. However, even after controlling for these measures, the negative relationship between urban distance and rural living standards is statistically significant. This suggests that the cost of participating in urban markets remains high, and has not been mitigated by providing additional public goods and services.

This paper contributes to multiple strands of research in economics. First, we contribute to the literature investigating differences in productivity and prices across space. Such differences can be large across urban areas of developed economies such as the United States (**?**). We show that in developing countries, there are large differences in productivity even across rural areas, a significant part of which can be explained by our measures of remoteness and urban proximity. Remoteness also contributes to higher prices of tradeable goods and weaker pass-through from global price changes (**?**), which can contribute to lower productivity (and therefore earnings) in these places.

Our paper also contributes to the literature exploring the relationship between state capac-

ity and physical remoteness, which finds that state power in Africa is decreasing in distance from the capital (Bates, 1983; Herbst, 2014; **?**; **?**). We find that state capacity in India is not just decreasing in distance from the capital, but to any urban center. The curse of remoteness is twofold: less state, and also less market due to the large distance to neighboring urban areas.

Finally, there is a growing literature estimating the causal impact of transportation costs and market access on rural economic outcomes in low-income countries (Donaldson, forthcoming; Donaldson and Hornbeck, 2016; Faber, 2014; **?**; **?**). These papers have largely showed that investments in highways and railroads increase rural living standards. This paper adds to that literature in two ways. First, we document that both economic activity and state capacity are decreasing in distance from urban markets. Second, even after controlling for transportation infrastructure, we find that remote villages are significantly poorer than their more urban-proximate counterparts, suggesting that such investments are not sufficient to close to gap between them.

The rest of the paper proceeds as follows: Section II describes our dataset and main variables of interest. Section III explains the empirical strategy. Section IV presents results and discussion. Section V concludes.

II Data

In order to study the relationship between remoteness and the rural economy, we combine data from 4 main sources: the Socioeconomic census (2012), the Population Census (1991, 2001 and 2011), the Economic Census (1990, 1998 and 2005), and geo-coordinates for all towns and villages in India. Below, we describe each source in greater detail.

II.A Socioeconomic census

The primary outcomes presented in this paper come from individual- and household-level microdata from a national socioeconomic census. Beginning in 1992, the Government of

India has conducted multiple household censuses in order to determine eligibility for various government programs (Alkire and Seth, 2012). In 1992, 1997 and 2002, these were referred to as Below Poverty Line (BPL) censuses. Households that were automatically considered above the poverty line were not included in this dataset.

However, the fourth such census, the Socioeconomic and Caste Census (SECC), departed from the previous methodologies by collecting data on all households, even if they demonstrated characteristics that would exclude them from eligibility under various government schemes targeted at the poor.[1]

The Government of India has made the SECC publicly available on the internet in PDF and Excel form. In order to construct a useful microdataset, we scraped over two million files, parsed the embedded into text data, and translated these from twelve different India languages into English. At the individual level, these data contains variables describing age, gender, occupation, caste group, disability and marital status. At the household level, this dataset contains variables describing housing, landholdings, agricultural assets, household assets and sources of income. We are able to match these data to our other datasets at the village level. This dataset is unique in describing the economic conditions of every person and household in rural India, at a spatial resolution unavailable from comparable sample surveys.

II.B Economic and population censuses

The Indian Ministry of Statistics and Programme Implementation (MoSPI) conducted the 3rd, 4th and 5th Economic Censuses in 1990, 1998 and 2005. The Economic Census is a complete enumeration of all economic establishments except those engaged in crop production and plantation; there is no minimum firm size, and both formal and informal establishments are included.

[1]It is often referred to as the 2011 SECC, as the initial plan was for the survey to be conducted between June and December 2011. However, various delays meant that the majority of the surveying was conducted in 2012, with urban surveys continuing to undergo verification at the time of writing. We therefore use 2012 as the relevant year for the SECC.

The Economic Census records information on the town or village of each establishment, whether ownership is public or private, the number and demographic characteristics of employees, the sources of electricity and finance, and the caste group of the owner. The main product of the firm is also coded using the 4-digit National Industrial Classification (NIC), which corresponds roughly to a 4-digit International Standard Industrial Classification (ISIC) code. More detailed information on income or capital is not included. The main strengths of the data are its comprehensiveness and rich detail on spatial location and industrial classification of firms. We obtained location directories for the Economic Censuses, and then used a series of fuzzy matching algorithms to match villages and towns by name to the population censuses of 1991 and 2001.[2] We were able to match approximately 93% of villages between 1998 and 2005.

We also make extensive use of data from the Population Censuses of 1991, 2001 and 2011. In addition to basic demographic data, the Population Census contains variables describing local public infrastructure (roads, electricity, schools and hospitals) and household assets, all aggregated to the village level.

II.C Other data

In addition to the socioeconomic, population and economic censuses, we use cross-sectional data from the 68th (2011-12) Round of the National Sample Survey (Employment/Unemployment), which contains far fewer villages and individuals than our socioeconomic census data, but includes data on earnings, place of work and time use across primary and secondary occupations.

We use village and town latitude and longitude from ML Infomap to generate measures of straight line distances from villages to cities and highways as a proxy for market access. Highway GIS data come from both OpenStreetMap and the National Highways Authority

[2]The Economic Census of 1998 was conducted with the house listing for the 1991 population census, while the 2005 Economic Census used codes from the 2001 population census.

of India.

II.D Measures of rural living standards

To the best of our knowledge, there is no publicly available data on incomes at the village level in India, or indeed any other large developing country. We attempt to overcome this limitation by generating imputed mean monthly earnings for each village using data from the SECC and the National Sample Survey. For the highest earning member of each household, the SECC reports whether the individual earns less than Rs 5000 (USD 75), between Rs 5000 - 10,000, or more than Rs 10,000 (USD 150). From the 68th Round (2011-12) of the National Sample Survey, we know the mean monthly earnings for individuals earning in these wage ranges. Conditional on earning less than Rs 5000, the mean monthly earnings are Rs 3076; for an individual earning between Rs 5000 - 10,000, the mean monthly earnings are Rs 6,373; and for individuals earning more than Rs 10,000 per month, the average monthly earning is Rs 22,353. We use these numbers - along with the share of individuals in a village earning in each of those wage brackets - to calculate the mean monthly earnings for each village. While it is not a precise measure of income, we think this is a good proxy for rural wages.

Moreover, we use other measures of rural living standards, such as the share of households whose highest earning member has monthly income greater than Rs 5000 and Rs 10,000 (reported by the 2012 SECC), the percentage of the village workforce engaged in agriculture or other farm activities (as reported by the 2011 Population Census), the percentage of village population that is not literate (again from the 2011 Population Census), and finally the number of non-farm jobs in the village (as reported by the 2005 Economic Census). The fact that these other measures display similar patterns as our imputed monthly earnings measures gives us more confidence in our results.

II.E Measures of rural public goods

We care about public goods because they may explain some or all of the variation in rural living standards in India. We distinguish between two types of public goods - socio-economic infrastructure that contributes to the well-being of the rural population, and connecting infrastructure, that enables the rural economy to transport people and goods to neighboring and distant urban centers. In the first category, we include measures of access to education (primary and secondary schools), health (medical centers), electricity, and irrigation. In the second category, we include the village's access to paved roads, regular bus services, railway services, and distance to trunk infrastructure such as highways.

II.F Measures of urban distance

In our preferred specification, we define distance to the nearest town as the straight line distance to the nearest town with a population of more than 10,000 people in 2011. We show that our results are robust to the population threshold applied to the nearest town. While we can use actual road distances as opposed to straight line distances, they add to computational costs without enhancing our understanding in any meaningful way.

II.G Summary statistics

Table 1 shows village-level summary statistics for the entire sample of Indian villages. We divide the sample into 2 halves based on distance to the nearest town. Column 1 contains average values for our full sample of villages. Column 2 contains average values for villages whose distance to the nearest town is less than the corresponding distance for the median village, while Column 3 reports average values for villages whose distance to the nearest town is more than the distance for the median village.

Villages in our sample have a mean population of 1470 individuals in 2011, while the median is 855, suggesting a longer right tail. 99% villages in the full sample had fewer than 9118 individuals in 2011. On average, the nearest town (with population greater than 10,000) is

15.8 kilometers away. The average imputed monthly earnings for the highest earning individual in a household in the village are Rs 5040 (USD 76). In about a quarter of the households, the highest earning member earns more than Rs 5000 (USD 75) per month, in only 7% of the households the highest earning member earns more than Rs 10,000 (USD 150) per month. About half the households have a solid roof. There were about 73 jobs in the non-farm sector in the average village in 2005, most of them in locally traded services. More than 71% of the workforce is engaged in agriculture or other farm activities in 2011. 42% of the population is not literate, but today almost 85% of the villages have a primary school. Only 19% have a secondary school. 22% have a medical center. 58% villages are electrified according to the 2011 census - and these receive on average 7.4 hours of electricity across summer and winter months. 80% of the villages have a paved road, but only 46% a regular bus service, and a mere 2% have a railway station. On average, distance to the nearest highway is roughly 9 kilometers, less than the distance to the nearest town.

Comparing columns 2 and 3, we observe that villages closer to towns have, on average, larger populations, more non-farm jobs, and a lower participation rate in agriculture. Mean monthly earnings for such villages are Rs 420 (USD 6) higher than monthly earnings for villages in the more remote half, which represents about 8.7% of the mean monthly earnings in more remote villages. Households in villages closer to towns are more likely to have a solid roof, and receive slightly more hours of electricity. Such villages also have a much higher share of land irrigated by public sources, and a lower percentage of not literate people in the population . The only public goods where there are no major differences based on distance to the nearest town are primary schools - which are fairly universal - and medical centers, which are universally rare.

III Empirical Strategy

We estimate Ordinary Least Squares (OLS) regressions in whcih we regress measures of rural living standards on the village's distance to the nearest town. We want to exclude variation arising from differences in village population: larger villages may on average receive more government attention, and therefore more public services than smaller villages. Further, there is considerable variation in governance quality and economic policies across India's 29 states - we want to exclude the variation in rural outcomes arising from these differences. Below the states in India's administrative hierarchy lie the districts. There are often substantial local differences in administration, historical conditions, or natural characteristics (eg soil fertility, rainfall etc) across India's 640 districts.[3] Hence, our preferred specification controls for village population, as well as district fixed effects. This means we estimate the relationship between rural economic outcomes and measures of remoteness for villages that are of similar population size and administered under similar political arrangements. We estimate the following equation:

$$Y_{v,t} = \beta_0 + \beta_1 distance_{v,j} + \beta_2 pop_{v,t} + \eta_v + \epsilon_{v,t}, \tag{1}$$

where $Y_{v,t}$ is the outcome of interest for village v at time t, $distance_{v,j}$ is the straight line distance in kilometers from village v to the nearest town j with population greater than 10,000, $pop_{v,t}$ is village population at time t (or at a time closest to t - for example, for SECC outcomes from 2012, the village population is from the Population Census 2011), and η_v is the district fixed effect.

The change in outcome $Y_{v,t}$ for a village for an additional kilometer of distance from the nearest town is captured by β_1, which is reported in the first row of Table 2.

To understand whether this remoteness effect is driven by size of the nearest town, we

[3] There were 640 districts in 2011-12, which is the last period for which we have economic and demographic data. There are reportedly 686 districts in India as of 2016.

run a separate regression in which we include distances to neighboring towns above different population thresholds:

$$\mathrm{Y}_{v,t} = \beta_0 + \beta_1 distance_{v,j,10,000} + \beta_2 distance_{v,j,50,000} + \beta_3 distance_{v,j,100,000} + \beta_4 distance_{v,j,500,000} + \beta_5 pop_{v,t} + \eta_v + \epsilon_{v,t}, \tag{2}$$

where $Y_{v,t}$ is the outcome of interest for village v at time t, $distance_{v,j,k,000}$ is the straight line distance in kilometers from village v to the nearest town j with population greater than k,000, $pop_{v,t}$ is village population at time t (or at a time closest to t - for example, for SECC outcomes from 2012, the village population is from the Population Census 2011), and η_v is the district fixed effect.

Finally, in order to test whether measures of state capacity or public goods explain all of the negative relationship between village mean monthly earnings and distance to nearest town, we regress mean monthly earnings on distance measures, while controlling for infrastructure (access to paved roads, railways, distance to highways) and public goods (primary school, medical center, electricity provision):

$$Y_{v,t} = \beta_0 + \beta_1 distance_{v,j} + \beta_2 pop_{v,t} + \zeta X_{v,t} + \eta_v + \epsilon_{v,t}, \tag{3}$$

where $Y_{v,t}$ is the outcome of interest for village v at time t, $distance_{v,j}$ is the straight line distance in kilometers from village v to the nearest town j with population greater than 10,000, $X_{v,t}$ is a vector of village controls (measures of connecting infrastructure and socio-economic infrastructure), $pop_{v,t}$ is village population at time t (or at a time closest to t - for example, for SECC outcomes from 2012, the village population is from the Population Census 2011), and η_v is the district fixed effect.

In all regressions, we cluster standard errors by the nearest town, as defined above, to the village. The statistical significance of results doesn't change when we cluster by district.

IV Results

Table 2 presents the results from running the regression specification in Equation 1. Column 1 contains results without any controls and shows that each additional 10 kilometers distance from the nearest town is associated with a reduction of Rs 226 in mean monthly earnings in the village. This is equivalent to 4.4% of the average monthly earnings in our full sample. In Column 2, we control for village population size in 2011, since as explained in the previous section, we expect larger villages to be better off in several economic dimensions than smaller villages. Controlling for population reduces the magnitude but doesn't change the direction or significance of the estimate. In Columns 3 and 4, we include state and district fixed effects, to exclude any variation in rural earnings arising from differences in quality of governance or local geographic factors. The inclusion of fixed effects marginally reduces the magnitude of the correlation between urban distance and earnings, but it remains negative and statistically significant. Column 4 presents results from our preferred specification, which controls for village population size and district fixed effects. While we cannot interpret the coefficient as a causal estimate, it represents a large negative correlation between remoteness and income: 10 kilometers of distance from the nearest town is associated with a reduction in mean monthly earnings of Rs 159, or 3.2% of the average monthly earnings in our full sample.

Urban proximity seems to matter: villages closer to towns have higher incomes. Table 3 further explores whether this proximity premium arises from small towns or large towns. We have 3,718 towns in our sample, which we divide into 4 population intervals. Column 1 repeats the preferred specification presented in Column 4 of Table 2. Column 2 contains results from restricting our sample to towns with more than 50,000 people, Column 3 to towns with more than 100,000 people, and Column 4 to towns with more than 500,000 people. We observe that as we increase the threshold for town size, the coefficient on the distance variable decreases.[4] Therefore, the villages benefit even from being close to small towns, and

[4]This may be because as the threshold increases, the average distance from a village to a town that

the urban proximity does not seem to be driven by large towns and cities.

These results can be easily visualized in Figure 1. The first graph represents the binscatter version of Column 1 in Table 2. In subsequent graphs, we control for village population and include state and then district fixed effects, and also vary the size threshold for the nearest town. We observe that while the gradient of the relationship between earnings and distance flattens, the negative relationship remains intact.

Evidence for the negative relationship between rural living standards and distance from urban areas is not limited to monthly earnings imputed from SECC and NSS data. We observe this negative relationship across a range of indicators from multiple sources, as presented in Table 4. We measure the share variables on a scale of 0-100, therefore we observe that each additional 10 kilometers from a town is associated with a 0.7% decrease in the share of households with monthly income greater that Rs 10,000 (approximately USD 150), and a 3.2% decrease in the share of households with a solid roof, as reported by the 2012 SECC. Further, it is associated with 4 fewer non-farm jobs, as reported by the 2005 Economic Census. Finally, as reported by the 2011 Population Census, it is associated with a 1.9% increase in the illiteracy percentage in the village, and a 3.8% increase in the percentage of the village workforce engaged in agriculture.[5]

We observe these patterns in Figure 2: the share of households with incomes above Rs 5000 ad Rs 10,000 monotonically decreases with distance to nearest town, share of households with solid roof as well as non-farm employment follow the same patterns, whereas the percentage of population engaged in agriculture and not literate follows the reverse patterns and strictly increases with distance to nearest town.

What is striking is that these patterns have not improved over the last 20 years. While we don't have village-level income data for the early 1990s, we do observe other village-level

meets the size criterion increases, and the influence of the town correspondingly decreases.

[5]This matters because labour productivity in agriculture is generally lower than in non-farm employment (Caselli, 2005; Gollin et al., 2014; Restuccia et al., 2008)

proxies such as non-farm jobs, percentage of workforce engaged in agriculture and percentage of village residents who are not literate. Table 5 reports these results. Each additional 10 kilometers from a town is associated with 1.17 fewer non-farm jobs (as reported by the 1990 Economic Census), 2.45% higher participation in agriculture, and a 2.15% higher illteracy rate (as reported by the 1991 Population Census). The results suggest that the gradient of economic outcomes against urban distance has not changed significantly in the last 20-21 years.

IV.A State Capacity

One explanation for the negative relationship between urban distance and rural economic conditions could be differences in the capacity of the state to provide public goods in more remote areas. This may be due to higher cost of provision in these areas, or a resource-constrained state focussing investments in more productive villages where the marginal returns are higher. It may also result from citizens in more remote villages having a more constrained voice: fewer voters live in more remote locations, and hence politicians may have weaker incentives to provide public goods. On the other hand, a state that cares about spatial inequality may respond to this situation by increasing the relative provision of public services to more remote villages.

We distinguish between two broad types of public goods: socio-economic infrastructure, such as schools, medical centers, electricity and public irrigation, and connecting infrastructure, such as paved roads, highways, railways and bus services. Table 6 and Table 7 provide estimates to suggest that both types of public goods are negatively correlated with urban distance, as reported by the 2011 Population Census.

We observe that each additional 10 kilometers of distance from the nearest town is associated with a 3.29 percentage points reduction in the probability that the village receives any electricity, a 0.45 hours reduction in the average hours of electricity received by the village, a 3% reduction in the share of village land which is irrigated, and a 0.26 points reduction

in the probability that the village has a secondary school. Probability of village having a primary school and a medical center are not affected by urban distance - this follows from our observation in Table 1 that nearly all villages have a primary school, whereas very few villages have a medical center, as reported by the 2011 Population Census.

Figure 3 illustrates these patterns through binscatter plots. We observe that primary and secondary schools and medical centers don't vary significantly with urban distance, the variation is number of daily hours of electricity is very small, while a clear negative pattern is discernible village electrification and irrigation.

Moving to connecting infrastructure, we observe that villages farther away from towns suffer the additional handicap of receiving more limited connecting infrastructure or transport services that could have helped to mitigate the negative consequences of remoteness. Each additional 10 kilometers of urban distance is associated with a 2% reduction in the probability that the village has a paved road, a 0.9% reduction in the probability of receiving bus services, and a 0.4% reduction in probability of receiving railway services. Moreover, villages that are more distant from towns are also more distant from trunk infrastructure such as highways: each additional 10 kilometers distance from the nearest town increases the village's average distance to a highway by 2.63 kilometers. Figure 4 illustrates these patterns.

Estimates in Table 8 suggest that the negative relationship between urban distance and public goods was also present in the early 1990s.

IV.B State capacity versus cost of transportation

Does the negative relationship between distance to urban areas and public good provision, as presented in the previous sub-section, explain away the negative relationship between distance to urban areas and rural living standards that we saw in Table **??**? If it does not, then there may be other sources of this distance advantage, such as the cost of transportation between rural and neighboring urban areas.

We answer this question by running the regression specification in Equation **??**, where we regress mean monthly earnings on distance to nearest town, while controlling not just for village population and district fixed effects, but also for the level of provision of socio-economic and connecting infrastructure, as set up in Equation **??**. Table 9 presents results from this regression.

We observe that while controlling for connecting infrastructure, such as the presence of paved roads (Column 2), or railway services, bus services and distance to trunk highways (column 4) reduces the magnitude of the negative urban distance coefficient, it does not completely take away the effect. Controlling for measures of socio-economic infrastructure (schools, medical centers, electrification, irrigation), as in Column 3, yield a similar result. Finally, in Column 5, we observe that even when we control for a wide range of public goods, the negative coefficient on urban distance remains statistically significant. Controlling for socio-economic and connecting infrastructure, each additional 10 kilometers of distance to the nearest town reduce mean monthly earnings by about Rs 112 (approximately USD 2), which is 2.2% of the average in our full sample. Therefore, while a restricted provision of public goods explains a portion of the negative relationship between urban distance and rural living standards, it does not explain the full amount, from which we can surmise that cost of transportation to urban markets remains a substantial barrier for rural economic prosperity.

V Conclusion

In this paper, we have documented extensive evidence for the negative effects of remoteness, as defined by distance to the nearest town, on a village's living standards. We have shown that average village earnings, share of households earning above a certain threshold, and non-farm employment all decrease monotonically with distance to nearest town, while the share of villagers engaged in agriculture and the percentage of population not literate increases monotonically with distance. Strikingly, these patterns are similar to the patterns

in 1990-91, thus highlighting the persistence of this remoteness disadvantage.

We then showed that in addition to living standards, measures of state capacity, such as socio-economic and connecting infrastructure, are also negatively correlated with distance to urban centers. While this partially explains the lower living standards in more remote villages, it doesn't explain the full variation. From this, we surmise that costs of transporting goods and people remain an unsurmounted barrier, and contribute to a village's general poverty. This casts the existing causal evidence regarding the effect of infrastructure projects on household outcomes in low-income countries in new light. While we cannot ascribe causal interpretations to our estimates, this emphasizes the need for more rigorous work on understanding how rural and urban development interact with each other.

Table 1: Summary statistics

	Full Sample	Closer Villages	Remoter Villages
Distance to nearest town (kms)	15.79	8.296	23.55
	(10.87)	(3.235)	(10.56)
Population (2011)	1469.9	1653.3	1279.9
	(1991.0)	(2126.0)	(1821.3)
Mean monthly earnings (2012)	5040.1	5246.9	4825.8
	(2380.9)	(2416.9)	(2323.7)
Percent households with income over 5,000 (2012)	23.25	25.61	20.80
	(22.90)	(22.92)	(22.62)
Percent households with income over 10,000 (2012)	7.493	8.300	6.656
	(11.27)	(11.52)	(10.95)
Percent households with solid roof (2012)	48.40	53.43	43.19
	(34.26)	(32.13)	(35.60)
Non-farm employment (2005)	73.74	84.22	62.88
	(253.2)	(282.2)	(218.7)
Percent population illiterate (2011)	42.32	40.46	44.25
	(14.11)	(12.86)	(15.05)
Percent population engaged in agriculture (2011)	70.79	67.51	74.19
	(27.23)	(27.13)	(26.91)
Percent villages electrified (2011)	58.33	64.15	52.30
	(49.30)	(47.96)	(49.95)
Daily hours of electricity (2011)	7.417	7.692	7.133
	(7.930)	(7.543)	(8.303)
Percent villages with primary school (2011)	84.90	85.11	84.68
	(35.81)	(35.60)	(36.02)
Percent villages with secondary school (2011)	18.89	19.70	18.05
	(39.14)	(39.77)	(38.46)
Percent villages with health center (2011)	21.80	20.97	22.65
	(41.29)	(40.71)	(41.86)
Percent land irrigated (2011)	57.52	67.55	47.11
	(38.64)	(35.75)	(38.77)
Paved Road Access (2011)	79.57	82.77	76.25
	(40.32)	(37.76)	(42.55)
Percent villages with railway service (2011)	2.355	2.907	1.783
	(15.16)	(16.80)	(13.23)
Percent villages with bus service (2011)	46.43	46.51	46.34
	(49.87)	(49.88)	(49.87)
Distance to nearest highway (kms)	8.902	6.895	10.98
	(8.225)	(6.854)	(8.977)
Observations	502117	255547	246570

Notes: This table presents means and standard deviations for observed outcomes for all villages in our sample. The 2005 data is from the Economic Census 2005, 2011 data is from the Population Census 2012, and the 2012 data from the Socioeconomic Census 2012. The "closer villages" column presents values for villages whose distance to the nearest town is less than distance to nearest town for the median village. The "remoter villages" column presents values for villages whose distance to the nearest town is greater than distance to nearest town for the median village.

Table 2: Urban distance and rural earnings, 2012

	(1)	(2)	(3)	(4)
Distance to nearest town (kms)	-22.624	-20.751	-17.641	-15.909
	(2.763)***	(2.701)***	(1.111)***	(1.121)***
Population (2011)		0.059	0.076	0.073
		(0.009)***	(0.004)***	(0.003)***
Fixed effects	None	None	State	District
N	529223	515978	515978	515978
R2	.01422	.01715	.1931	.2757

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: The table presents OLS estimates from Equation X, where we regress imputed mean earnings on distance to the nearest town. Imputed mean earnings are based on assigning monthly earnings of 3,076 rupees to households whose highest earning member reports income of less than 5,000 rupees, 6,373 rupees to households reporting greater than 5,000 but less than 10,000 rupees and 22,353 rupees to households reporting greater than 10,000 rupees. These numbers are mean monthly earnings for earners in these wage ranges as reported by the 68th Round (2011-12) of the National Sample Survey. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 3: Town size and rural earnings, 2012

	Nearest Town Pop>10,000	Nearest Town Pop>50,000	Nearest Town Pop>100,000	Nearest Town Pop>500,000
Distance to nearest town	-15.909	-7.115	-4.510	-1.394
	(1.121)***	(0.708)***	(0.509)***	(0.336)***
Population (2011)	0.073	0.074	0.075	0.076
	(0.003)***	(0.003)***	(0.003)***	(0.003)***
Fixed effects	District	District	District	District
N	515978	515978	515978	515978
R2	.2757	.2736	.2727	.2715

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: The table presents OLS estimates from Equation X, where we regress imputed mean earnings on distance to the nearest town. Village size matters for mean earnings, but distance depresses earnings even when controlling for population. Imputed mean earnings are based on assigning monthly earnings of 3,076 rupees to households whose highest earning member reports income of less than 5,000 rupees, 6,373 rupees to households reporting greater than 5,000 but less than 10,000 rupees and 22,353 rupees to households reporting greater than 10,000 rupees. These numbers are mean monthly earnings for earners in these wage ranges as reported by the 68th Round (2011-12) of the National Sample Survey. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 4: Urban distance and other rural indicators, 2005-2011

	Income > 10,000	Solid Roof	Non-farm Jobs	Percent in Agriculture	Percent Illiterate
Distance to nearest town (kms)	-0.070	-0.319	-0.385	0.384	0.190
	(0.005)***	(0.021)***	(0.072)***	(0.017)***	(0.009)***
Population (2011)	0.000	0.001	0.060	-0.001	-0.000
	(0.000)***	(0.000)***	(0.002)***	(0.000)***	(0.000)***
Fixed effects	District	District	District	District	District
N	515978	515966	515978	512492	515978
R2	.2492	.6195	.4175	.2875	.4643

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: This table presents OLS estimates for a range of rural economic indicators, using our preferred specification with district fixed effects. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 5: Urban distance and rural indicators, 1990-91

	Non-farm Jobs (1990)	Percent in Agriculture (1991)	Percent Illiterate (1991)
Distance to nearest town (kms)	-0.117	0.245	0.215
	(0.066)*	(0.012)***	(0.009)***
Population (1991)	0.065	-0.002	-0.001
	(0.002)***	(0.000)***	(0.000)***
Fixed effects	District	District	District
N	508080	507824	508080
R2	.08669	.2472	.4838

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: This table presents OLS estimates for a range of indicators about rural economic activity, using our preferred specification with district fixed effects. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 6: Urban distance and public goods, 2011-12

	Electrification	Daily hours of electricity	Irrigation	Primary School	Secondary School	Medical Center
Distance to nearest town (kms)	-0.329	-0.045	-0.304	-0.021	-0.026	0.002
	(0.024)***	(0.004)***	(0.025)***	(0.019)	(0.010)**	(0.015)
Population (2011)	0.002	0.000	0.000	0.004	0.009	0.007
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Fixed effects	District	District	District	District	District	District
N	515978	515978	500108	515965	515978	515978
R2	.4727	.5979	.6023	.1885	.2311	.2734

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: This table presents OLS estimates for a range of indicators for state capacity and provision of public goods, using our preferred specification with district fixed effects. Public good availability is reported from 2011. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 7: Urban distance and connecting infrastructure, 2011

	Paved Road	Distance to Highway	Bus Service	Railway Service
Distance to nearest town (kms)	-0.197	0.263	-0.090	-0.037
	(0.021)***	(0.013)***	(0.019)***	(0.004)***
Population (2011)	0.003	-0.000	0.005	0.001
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Fixed effects	District	District	District	District
N	515972	515978	515978	515978
R2	.3249	.3372	.3455	.02683

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: This table presents OLS estimates for a range of indicators for state capacity and provision of public goods, using our preferred specification with district fixed effects. Public good availability is reported from 2011. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Table 8: Urban distance and public goods, 1991

	Electrification (1991)	Irrigation (1991)	Primary School (1991)
Distance to nearest town (kms)	-0.004	-0.165	-0.004
	(0.000)***	(0.019)***	(0.019)
Population (1991)	0.000	0.000	0.010
	(0.000)***	(0.000)***	(0.000)***
Fixed effects	District	District	District
N	508077	489425	508080
R2	.242	.6238	.4535

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

Notes: This table presents OLS estimates for a range of indicators for state capacity and provision of public goods, using our preferred specification with district fixed effects. Public good availability is reported from 2011. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

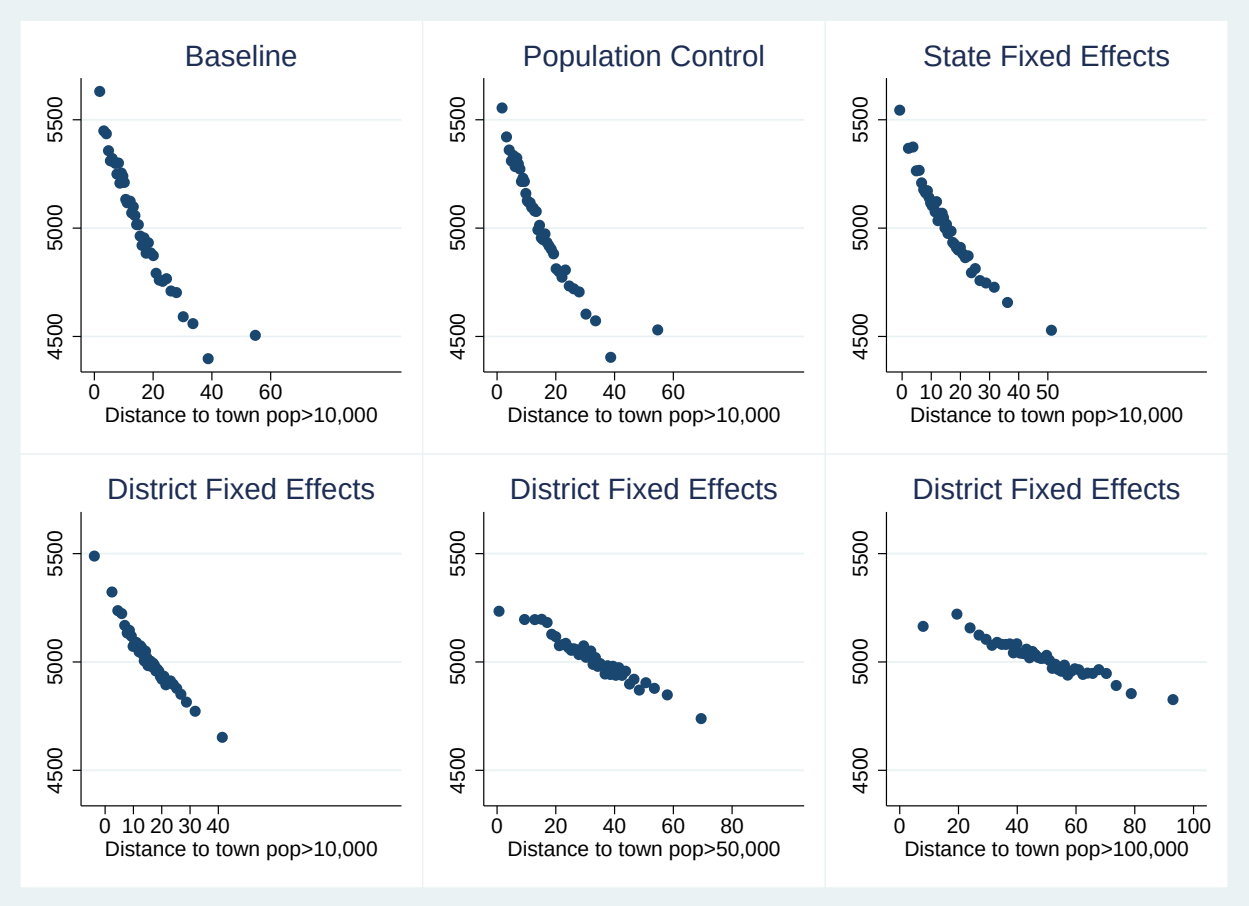
Table 9: State capacity v/s cost of transportation

	(1)	(2)	(3)	(4)	(5)
Distance to nearest town (kms)	-15.909	-15.417	-14.315	-12.214	-11.192
	(1.121)***	(1.114)***	(1.176)***	(1.118)***	(1.161)***
Population (2011)	0.073	0.067	0.031	0.052	0.022
	(0.003)***	(0.002)***	(0.002)***	(0.002)***	(0.002)***
Paved Road Access (2011)		2.437		1.760	1.300
		(0.165)***		(0.140)***	(0.144)***
Percent villages electrified (2011)			0.822		0.533
			(0.166)***		(0.169)***
Daily hours of electricity (2011)			8.430		7.613
			(1.300)***		(1.286)***
Percent villages with primary school (2011)			0.432		0.103
			(0.126)***		(0.123)
Percent villages with secondary school (2011)			2.995		2.620
			(0.094)***		(0.089)***
Percent villages with health center (2011)			0.969		0.737
			(0.081)***		(0.082)***
Percent land irrigated (2011)			2.466		2.290
			(0.235)***		(0.231)***
Distance to nearest highway (kms)				-11.530	-10.837
				(0.950)***	(0.946)***
Percent villages with bus service (2011)				2.372	1.842
				(0.116)***	(0.120)***
Percent villages with railway service (2011)				2.611	2.372
				(0.194)***	(0.191)***
Fixed effects	District	District	District	District	District
Connecting Infrastructure	None	Paved Roads	None	All	All
Public Goods	None	None	All	None	All
N	515978	515972	500096	515972	500096
R2	.2757	.2773	.2816	.2815	.2854

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$

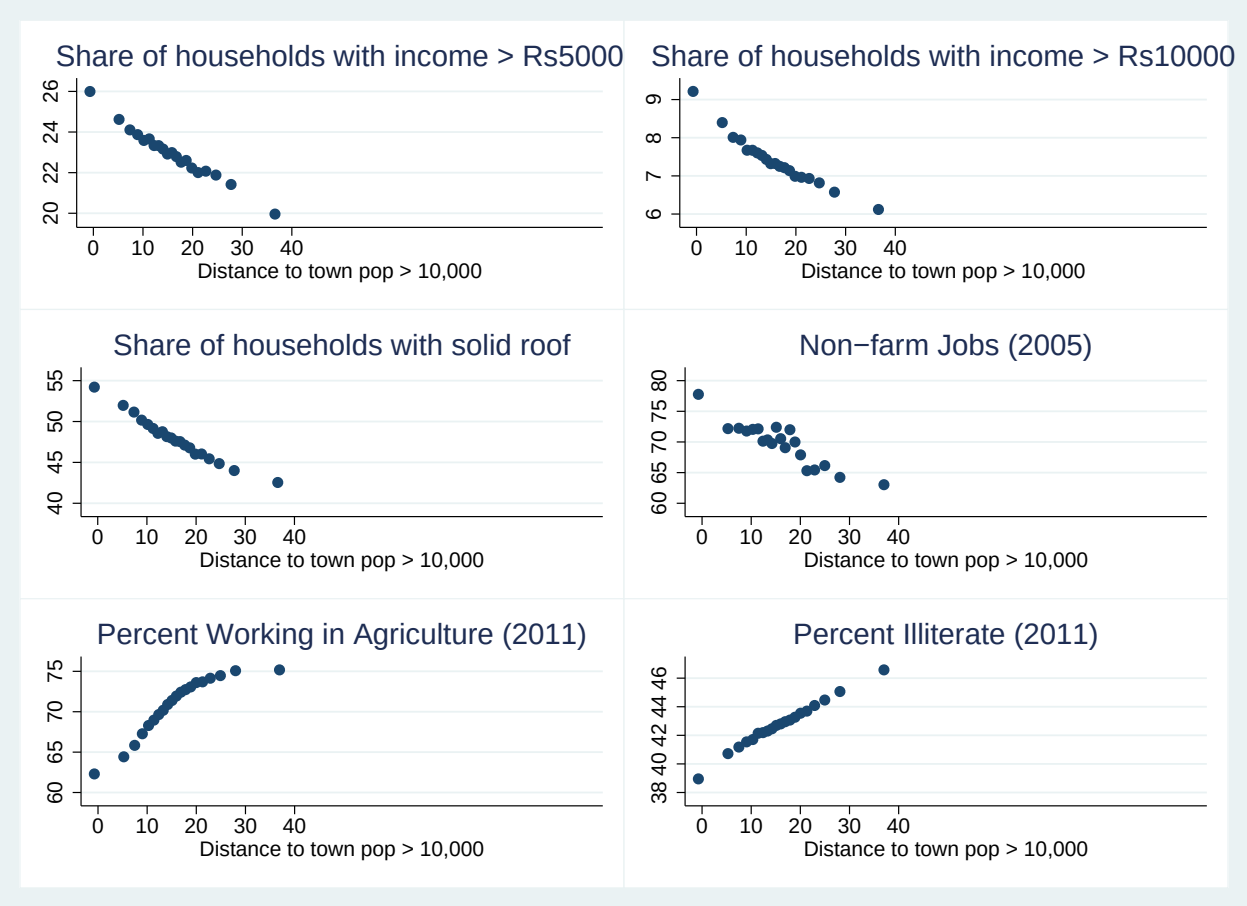
Notes: The table presents OLS estimates from Equation X, where we regress imputed mean earnings on distance to the nearest town and a range of indicators for village access to public goods. While controlling for public goods reduces the magnitudes of the OLS estimates, they still remain large and statistically significant. Imputed mean earnings are based on assigning monthly earnings of 3,076 rupees to households whose highest earning member reports income of less than 5,000 rupees, 6,373 rupees to households reporting greater than 5,000 but less than 10,000 rupees and 22,353 rupees to households reporting greater than 10,000 rupees. These numbers are mean monthly earnings for earners in these wage ranges as reported by the 68th Round (2011-12) of the National Sample Survey. Standard errors have been clustered by the nearest town for each village and are reported below point estimates.

Figure 1: Mean Earnings and Urban Distance



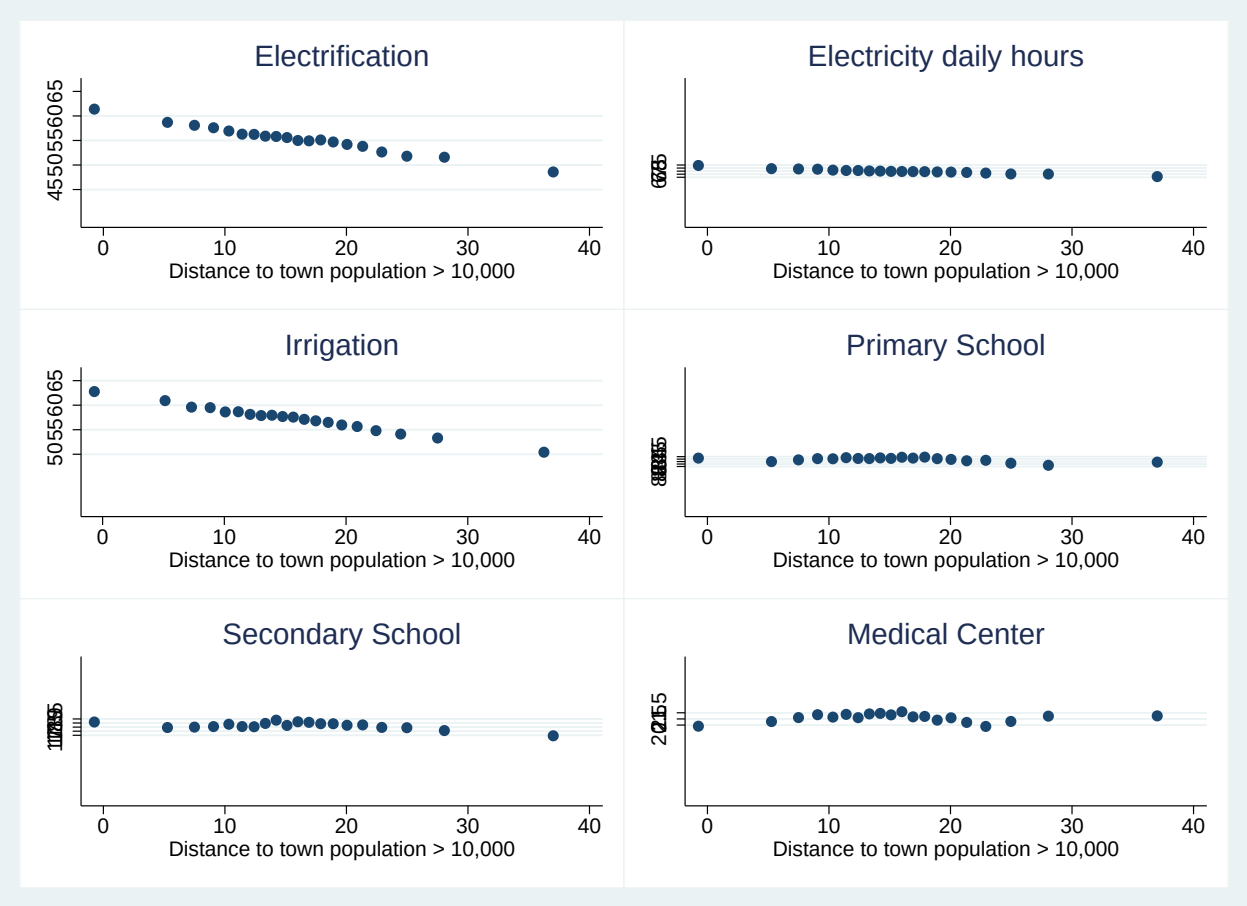
Notes: The figure shows how imputed mean village earnings change both as a function of distance to all towns, and as a function of distance to larger towns with population over 100,000. Imputed mean earnings are based on assigning monthly earnings of 3,076 rupees to households whose highest earning member reports income of less than 5,000 rupees, 6,373 rupees to households reporting greater than 5,000 but less than 10,000 rupees and 22,353 rupees to households reporting greater than 10,000 rupees. These numbers are mean monthly earnings for earners in these wage ranges as reported by the 68th Round (2011-12) of the National Sample Survey.

Figure 2: Rural Economy and Urban Distance



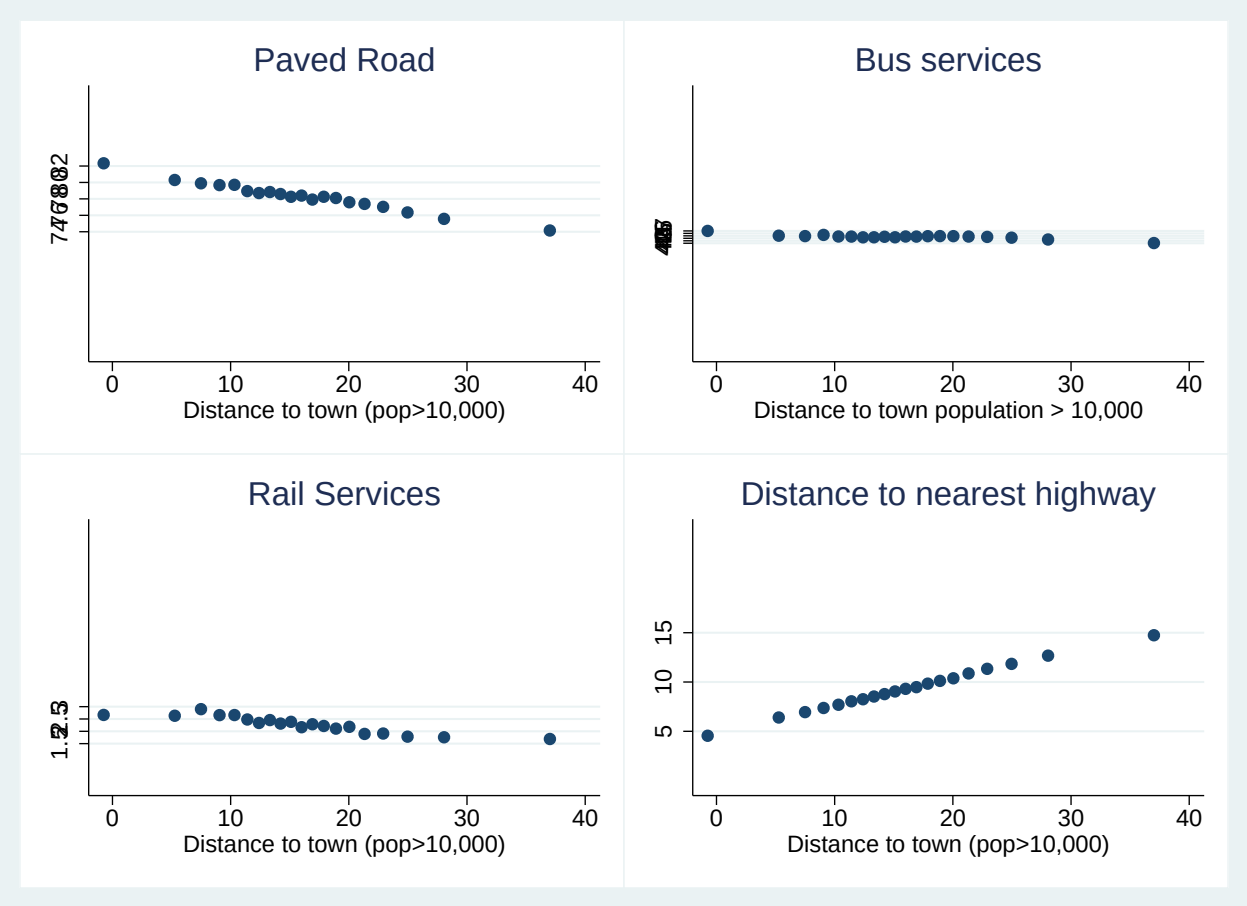
Notes: The figure shows how other indicators of economic activity and economic structure, such as the number of non-farm jobs, extent of illiteracy and participation in agriculture change as a function of distance to all towns, as well as distance to larger towns (population greater than 100,000).

Figure 3: Urban Distance and State Capacity



Notes: The figure shows how measures of state capacity, or the availability of public goods, change as a function of distance to all towns, as well as distance to larger towns (population greater than 100,000).

Figure 4: Urban Distance and Connecting Infrastructure



Notes: The figure shows how measures of state capacity, or the availability of public goods, change as a function of distance to all towns, as well as distance to larger towns (population greater than 100,000).

References

Alkire, Sabina and Suman Seth, "Identifying BPL Households: A Comparison of Methods," 2012.

Bates, Robert H, "Modernization, ethnic competition, and the rationality of politics in contemporary Africa," *State versus ethnic claims: African policy dilemmas*, 1983, *152*, 171.

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

Donaldson, Dave, "Railroads of the Raj: Estimating the impact of transportation infrastructure.," forthcoming.

___ **and Richard Hornbeck**, "Railroads and American Economic Growth: A Market Access Approach," *The Quarterly Journal of Economics*, 2016, p. qjw002.

Faber, Benjamin, "Trade integration, market size, and industrialization: Evidence from China's National Trunk Highway System," *The Review of Economic Studies*, 2014, p. rdu010.

Gollin, Douglas, David Lagakos, and Michael E Waugh, "Agricultural productivity differences across countries," *The American Economic Review*, 2014, *104* (5), 165–170.

Herbst, Jeffrey, *States and power in Africa: Comparative lessons in authority and control*, Princeton University Press, 2014.

Moretti, Enrico, "Cities and growth," *International Growth Center*, 2014.

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.

The World Bank, "World Development Report 2009: Reshaping Economic Geography," Technical Report, World Bank 2009.