Rural Roads and Structural Transformation*

Sam Asher† Paul Novosad‡

This version: May 2017
First version: March 2013

Abstract

More than one billion people worldwide live in rural areas without access to the paved road network. We show that poor transportation infrastructure is a major constraint on the sectoral allocation of labor. Exploiting program rules from a national rural road construction program and a comprehensive dataset covering every individual in rural India, fuzzy regression discontinuity estimates show that paved roads reduce the share of workers in agriculture by ten percentage points. This sectoral reallocation is concentrated among workers with the highest returns to sectoral reallocation. Rather than facilitating firm growth, rural roads enable workers to access external labor markets.

JEL Codes: O12/O18/J43.

*We are thankful for useful discussions with Abhijit Banerjee, Lorenzo Casaburi, Melissa Dell, Ed Glaeser, Doug Gollin, Ricardo Hausmann, Rick Hornbeck, Clement Imbert, Lakshmi Iyer, Radhika Jain, Asim Khwaja, Michael Kremer, Sendhil Mullainathan, Rohini Pande, Simon Quinn, Gautam Rao, Andrei Shleifer, Seth Stephens-Davidowitz, Andre Veiga, Tony Venables and David Yanagizawa-Drott. We are indebted to Taewan Roh and Kathryn Nicholson for exemplary research assistance. This paper previously circulated as "Market Access and Structural Transformation: Evidence from Rural Roads in India." This project received financial support from the Center for International Development and the Warburg Fund (Harvard University), as well as the IGC and GLM-LIC program. All errors are our own.

†World Bank, sasher@worldbank.org

‡Dartmouth College, paul.novosad@dartmouth.edu

I Introduction

Labor productivity in agriculture is significantly lower than in other sectors of the economy (Caselli, 2005; Gollin et al., 2014; Restuccia et al., 2008). This is particularly true in developing countries, whose economies are also characterized by high population shares living in rural areas and working in agriculture (McMillan et al., 2014). Researchers going back to Lewis (1954) have suggested that labor market imperfections prevent the reallocation of labor from agriculture to higher productivity activities.[1] This paper focuses on one particular friction: the poor state of transportation infrastructure in low-income countries (Atkin et al., 2015). One billion people, or thirty-one percent of the world's rural population, live in settlements more than 2 km from a paved road. Ninety-eight percent of people lacking such access to outside markets and government services live in developing countries (World Bank, 2015). The resulting high transportation costs may prevent the flows of goods and labor necessary for structural transformation and the efficient allocation of labor.

Due to the scarcity of high spatial resolution data and the endogeneity of road placement, the economic impacts of rural roads have proven difficult for researchers to assess. The high costs and potentially large benefits of infrastructure investments mean that road construction is likely to be correlated with both economic and political characteristics of locations.[2] We overcome this challenge by taking advantage of a large-scale natural experiment in an Indian national rural road construction program, which by 2015 had built over 100,000 roads to over 185,000 villages at a cost of nearly $40 billion. The implementation guidelines produce exogenous variation in paved road access by generating discontinuities in

[1] A large literature has suggested that barriers to the reallocation of labor could result from search costs (Harris and Todaro, 1970), credit constraints (Banerjee and Newman, 1993), informational frictions (Banerjee and Newman, 1998; Bryan et al., 2014), the cost of human capital acquisition (Caselli and Coleman, 2001), and insurance networks that discourage movement out of rural areas (Munshi and Rosenzweig, 2016).

[2] Brueckner (2014) uses international oil price movements to show that investment in infrastructure responds strongly to economic growth. Burgess et al. (2015) show that the ethnic homelands of Kenya presidents receive greater road investments, although this effect disappears during periods of democracy. Blimpo et al. (2013) show in the cross section that politically marginalized areas across West Africa have lower levels of road infrastructure.

the probability of treatment at two village population thresholds (500 and 1000). We exploit these population thresholds to estimate the economic impact of rural roads using a fuzzy regression discontinuity design.

To utilize village-level variation in new roads, we construct a high spatial resolution dataset that combines household and firm microdata with village aggregates describing amenities, infrastructure and demographic information. We assemble the microdata from the 2012 Socioeconomic and Caste Census (SECC), which contains economic data for every individual and household in rural India, as well as an economic census covering all nonfarm firms. With these microdata, we are able to test hypotheses that would be impossible with aggregate data or household surveys, joining a growing body of economic research that utilizes comprehensive administrative data to investigate otherwise elusive research questions (Einav and Levin, 2014).

We find that rural roads lead to large movements of workers out of agriculture: a new road causes a 10 percentage point decrease in the share of workers in agriculture and an equivalent increase in wage labor.[3] These impacts are most pronounced among the groups with the lowest costs and highest potential gains from participation in labor markets: households with small landholdings and working age men, as well as in villages with low agricultural productivity. Roads also lead to increases in observed durable assets, but these results are not statistically significant.

We argue that our results are best explained by the increased access of rural households to labor markets beyond the village. We present a model of occupational choice across three sectors: village agriculture, village non-agriculture and external labor markets. We show that the evidence is consistent with a theory where new roads allow workers to arbitrage wage gaps between rural agricultural and outside labor markets. We consider two alternatives to this story: (i) increased agricultural productivity reduces demand for labor, and (ii)

[3]Given a control group mean of 48%, this is equivalent to a 21 percent decline.

growth in within-village nonfarm activities induces movement out of agriculture. We find no evidence for increases in agricultural investments or major land consolidation, suggesting that investments in agriculture have not reduced demand for agricultural labor. Nor is there evidence to support an large increase in nonfarm economic activity in treated villages. We estimate a statistically insignificant 1 percentage point increase in the share of workers employed in nonfarm firms within the village, far too small to account for the 10 percentage point movement of village workers out of agriculture.

This paper contributes to multiple strands of research in economics. First, we add to a large literature seeking to understand the determinants of structural transformation in the process of development. It is well established that across the developing world, labor productivity outside agriculture is much higher than within agriculture (Gollin et al., 2014; McMillan et al., 2014). We provide evidence that poor transportation infrastructure is an important barrier to the reallocation of labor out of agriculture and into wage labor markets. This should not be surprising: for sectoral arbitrage to occur, there must be both an agricultural productivity gap (Gollin et al., 2014) and sufficiently low costs to reallocating labor that it is worthwhile to do so. This paper lends support to the argument that transportation costs pose a major barrier to spatial and sectoral reallocation of labor (Bryan et al., 2014; Bryan and Morten, 2015).

Our work also adds to a related literature examining the constraints on labor market participation in developing countries. Workers in low income countries are far more likely either to be self-employed or to work in informal firms, which have lower growth and productivity than formal sector firms (La Porta and Shleifer, 2014). The majority of self-employment and informality is in the agricultural sector. We show that transport infrastructure provision can play a large role in increasing labor force participation.

Second, we add to a growing body of research that seeks to estimate the causal effects of transport infrastructure in low- and middle-income countries. Transportation infrastructure has been shown to raise the value of agricultural land (Donaldson and Hornbeck, 2016), increase agricultural trade and income (Donaldson, n.d.), reduce the risk of famine (Burgess and Donaldson, 2012), increase migration (Morten and Oliveira, 2016) and accelerate urban decentralization (Baum-Snow et al., 2017). Results on growth have proven somewhat mixed: there is evidence that reducing transportation costs can increase (Ghani et al., 2015; Storeygard, 2016), decrease (Faber, 2014) or leave unchanged (Banerjee et al., 2012) growth rates in local economic activity. These papers have largely focused on highways and railroads. We add to this literature by providing some of the first causal estimates of the impact of smaller scale roads to rural areas, as well as providing detailed estimates of the response of households and firms to the construction of transport infrastructure. Our findings complement existing studies by documenting the impact of roads on the allocation of labor across sectors.

Third, we contribute causal estimates to the literature that examines the economic impacts of rural roads specifically.[4] Such feeder roads differ in multiple ways from inter-regional transport infrastructure such as railroads and major highways. As they do not affect transport costs between cities, they are unlikely to have the same impacts that the literature has found on firm location choices, productivity and income. They also lower transport costs to rural areas that often have few firms and lack complementary infrastructure such as electricity. We extend this literature in several ways. This paper is the first large-scale study on rural roads that combines household microdata with exogenous variation in road placement; in this regard we join recent work that has estimated the impacts of major infrastructural investments

[4]Most closely related are papers that estimate the impact of rural road programs in Bangladesh (Khandker et al., 2009; Khandker and Koolwal, 2011; Ali, 2011), Ethiopia (Dercon et al., 2009), Indonesia (Gibson and Olivia, 2010), Papua New Guinea (Gibson and Rozelle, 2003) and Vietnam (Mu and van de Walle, 2011). Concurrent research on the PMGSY demonstrates that districts that built more roads experienced improved economic outcomes (Aggarwal, 2015), and that PMGSY roads lead to gains in educational enrollment and test scores (Mukherjee, 2011; Adukia et al., 2017). Other papers also suggest that the lack of rural transport infrastructure may be a significant contributor to rural underdevelopment. Wantchekon and Stanig (2015) provide evidence that transport costs are a strong predictor of poverty across sub-Saharan Africa. Fafchamps and Shilpi (2005) offer cross-sectional evidence that villages closer to cities are more economically diversified, with residents more likely to work for wages.

such as dams (Duflo and Pande, 2007) and electrification (Dinkelman, 2011; Lipscomb et al., 2013).[5] While most research has focused on agricultural outcomes, we demonstrate the large impacts that new roads can have on reallocation of labor away from agricultural activity. Finally, much of this literature has generated estimates from very small numbers of roads; our large sample both argues for a higher degree of external validity and allows us to investigate how individual and household characteristics mediate the effects of rural connectivity.

The rest of the paper proceeds as follows: Section II provides a model of how rural roads may affect local economic activity and the sectoral allocation of labor. Section III provides a description of the rural road program. Sections IV and V describe the data construction and empirical strategies. Section VI presents results and discussion. Section VII concludes.

II A Model of Labor Allocation Across Sectors

We present a simple model to highlight the role that rural transportation infrastructure may play in the process of structural transformation, which we define as the movement of labor from the farm to the nonfarm sector. We consider the allocation of labor in a representative village in the absence of and in the presence of a road. We model the absence of a road as total autarky (sufficiently high transportation costs to prevent the exchange of goods or labor) and the presence of a road as perfect integration (prices and wages equalize with outside markets, with zero transportation costs).

The economy is comprised of three sectors: rural agriculture, rural non-agriculture and an external sector, which we take to also be non-agricultural. We model production as follows: output in sector i and with connectivity c is $Y_i^c = A_i^c \ln(l_i)$, where i takes the values a for agriculture and n for non-agriculture, and c takes the values 0 if the village lacks a paved

[5] An older literature suggested that rural transport infrastructure was highly correlated with positive development outcomes (Binswanger et al., 1993; Fan and Hazell, 2001; Zhang and Fan, 2004), estimating high returns to such investments. More recent work has generally demonstrated that rural roads are associated with large economic benefits by looking at their impact on agricultural land values (Jacoby, 2000; Shrestha, 2015), estimated willingness to pay for agricultural households (Jacoby and Minten, 2009), complementarities with agricultural productivity gains (Gollin and Rogerson, 2014), search and competition among agricultural traders (Casaburi et al., 2013), and agricultural productivity and crop choice(Sotelo, 2016).

road and 1 if it has a paved road. Note that production takes only labor as an input and that productivity depends on the presence of a road. The A_i^c term subsumes the availability and prices of inputs and outputs, as well as productive technologies. A_a^0 is the numeraire: all other productivity parameters are relative to the productivity of the agricultural sector in autarky. As the village is small relative to the external labor market (comprised of all villages and towns on the road network), it is assumed that the village is a price-taker with respect to the external wage (w_e) and goods prices. Labor supply is fixed and normalized to 1, and there is no labor-leisure trade-off.

Without a road, workers are not able to access the external sector and thus are allocated between only the rural agricultural and non-agricultural sectors. With homogeneous labor and frictionless labor markets, wages equalize across sectors and the labor market equilibrium determines the share of workers in the agricultural sector: $l_a^{0*} = \frac{1}{1+A_n^0}$.

With construction of a road, village labor can be allocated to either of the village sectors, or to the external labor market. All prices converge to those in the external market. Because the wage in both village sectors is pinned down by the external wage, the share of labor in agriculture is determined solely by its own productivity and the level of the outside wage: $l_a^{1*} = \frac{A_a^1}{w^e}$. Put differently, the labor allocation to the in-village sectors will be such that the marginal product of labor (and thus wage) equalizes with the external market wage.

Having characterized agriculture's share of labor in both states of the world, we can combine these solutions to generate the conditions under which new roads will lead to structural transformation:

$$l_{ag}^0 > l_{ag}^1 \iff \frac{w^e}{(1+A_n^0)} > A_a^1.$$

Road treatment leads to two independent effects on the share of labor in agriculture, as can be seen in the above expression. The first is that the revenue productivity of agriculture changes: because agricultural productivity without a road is normalized to 1, the A_a^1 on the right side of the inequality captures this change. This could be due to increased access to inputs such as fertilizer, or higher prices on the external market that are available now that transportation costs have been eliminated.[6] The second is that wages now rise to the level of the external labor market. This change is captured by the left side of the inequality.

Thus transportation investments cause labor to flow out of agriculture when the outside wage is high relative to both non-agricultural productivity without a road (which determines the agricultural labor share in autarky) and agricultural productivity with a road. If wages are higher in outside labor markets than in the village in autarky and the revenue productivity of agriculture rises with connectivity, these forces work against each other. A net flow of workers out of (into) agriculture implies that the outside wage effect is large (small) relative to the gain in agricultural productivity.

This model, simple as it is, generates additional predictions about where the effects of new roads on structural transformation should be largest. First, low agricultural productivity should lead to greater exit from agriculture upon road construction. Second, although we have not modeled worker heterogeneity, it is easy to see that workers with high external wages relative to their productivity in agriculture will be most likely to change sectors. This suggests that men, who face fewer barriers to participating in outside labor markets (higher outside wages net of opportunity costs) will be more likely to reallocate.[7] Effects may also depend on age if younger workers face higher wages in outside labor markets than in agriculture, either due to less sector-specific (and perhaps location-specific (Bazzi et al., 2016)) human capital in agriculture, or if they have lower search costs that translate into higher

[6]Sotelo (2016) estimates that paving existing roads will on average boost agricultural productivity by 15% by both increasing access to inputs and raising output prices. These changes both increase productivity directly and induce greater specialization.

[7]Attitudes against women's spending time far away from home, as well as their greater responsibilities in house work and child raising, may diminish any reallocation of female labor away from agriculture and into the labor market (Goldin, 1995).

net external wages.[8]

Likewise, if land is an important determinant of the marginal productivity of labor in agriculture, we should expect households with small landholdings to be most likely to reallocate upon receipt of a new road. This relationship between marginal labor productivity and landholding could emerge either from functioning land markets that allow more productive farmers to accumulate more land, or from market failures in land and labor that (in the limit) restrict households to farming only the land they own. Foster and Rosenzweig (2011) show that Indian farms are inefficiently small, suggesting potential frictions in these markets and high returns to reallocation from agriculture for those with small landholdings.

III Context and background

The Pradhan Mantri Gram Sadak Yojana (PMGSY) – the Prime Minister's Village Road Program – was launched in 2000 with the goal of providing all-weather access to unconnected villages across India. The focus was on the provision of new feeder roads to localities that did not have paved roads, although in practice many projects under the scheme upgraded pre-existing roads. As the objective was to connect the greatest number of locations to the external road network at the lowest possible price, link routes (terminating at a village) were to be given priority over through routes (those passing through a village to another larger road).

National guidelines determined the prioritization of road construction under the PMGSY. Most importantly for this paper, new roads were targeted to large villages, as defined by population in the 2001 Population Census. Originally, the stated goal was to provide all villages with populations greater than 1000 with connectivity by 2003 and all villages with population greater than 500 with connectivity by 2007, at which point villages over 250 were to be targeted.[9] The thresholds were lower in desert and tribal areas, as well as hilly

[8] Existing evidence lends some credence to this prediction: studying South Korea's rapid industrialization, Kim and Topel (1995) find that non-agricultural firms almost exclusively hired new entrants to the labor force; in other words, South Korea experienced rapid structural transformation at the aggregate level with little sectoral reallocation at the individual level.

[9] The unit of targeting in the PMGSY is the habitation, defined as a cluster of population whose location

states and districts affected by left-wing extremism. These rules were to be applied on a state-by-state basis, meaning that states that had achieved connectivity of all larger villages could proceed to smaller localities. However, program guidelines also laid out other rules that states could use to determine allocation. Smaller villages could be connected if they lay in the least-cost path of connecting a prioritized village. Groups of villages within 500m of each other could combine their populations. Members of Parliament and state legislative assemblies were also allowed to make suggestions that would be taken into consideration when approving construction projects. Finally, measures of local economic importance such as the presence of a weekly market could also influence allocation.

Although funded and overseen by the federal Ministry of Rural Development, responsibility for program implementation was delegated to state governments. District Rural Road Plans were drafted for every district in India, delineating a "core network" of roads that would be required to connect every village to the paved road network at the lowest possible cost. Funding came from a combination of taxes on diesel fuel (0.75 INR per liter), central government support and loans from the Asian Development Bank and World Bank. By 2015, over 400,000 km of roads had been constructed, benefiting 185,000 villages – 107,000 previously lacking an all-weather road – at a cost of more than $37 billion.[10]

IV Data

In order to estimate the economic impacts of new road provision, it was necessary to construct a village-level dataset that combines administrative data from the PMGSY program with multiple external datasets. In this section we describe the data sources and collection process.

does not change over time. Revenue villages, which are used by the Economic and Population Censuses, are comprised of one or more habitations. See National Rural Roads Development Agency (2005) for more details. In this paper, we aggregate all data to the level of the revenue village.

[10]Source: PMGSY administrative data.

IV.A Administrative Data on Road Construction

Our data on road construction come from the administrative software designed for the management of the program.[11] The data include road sanctioning and completion dates, cost and time overruns, contractor names, and quality monitoring reports.

PMGSY data are posted online at either the habitation or the road level; the data for this paper were all scraped in January 2015. There is a many-to-many correspondence between habitations and roads: roads serve multiple habitations, and habitations may be connected to multiple roads. A census village typically comprises between one and three habitations; approximately 200,000 villages, one third of the total, consist of only a single habitation. For the purposes of this paper, all variables are aggregated to the level of the census village, the geographic unit at which we measure outcomes. We consider a village to be treated by the road program if at least one habitation in the village received a completed road by the year before outcome data were collected.

We matched the administrative road data to economic, population and poverty census data at the village level. In order to generate a village correspondence across multiple datasets, we conducted a fuzzy matching of location names, along with manual cleaning and quality verification.[12] We successfully match over 85% of habitations listed in the PMGSY to their corresponding population census villages.

IV.B Socioeconomic censuses

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, 2013). In 1992, 1997 and 2002, these were referred

[11] All data are publicly available at http://omms.nic.in.

[12] For fuzzy matching, we used a combination of the reclink program in Stata, and a custom fuzzy matching script based on the Levenshtein algorithm but modified for the languages used in India. The fuzzy matching algorithm can be downloaded from the corresponding author's web site.

to as Below Poverty Line (BPL) censuses. We obtained the anonymized microdata to the 2002 BPL Census from the Ministry of Rural Development. This dataset contains individual demographic variables such as age, gender, and caste group, as well as various measures of household economic activity and assets.

The fourth such census, the Socioeconomic and Caste Census (SECC), was launched in 2011 but primarily conducted in 2012.[13] To increase the likelihood of collecting data on all individuals and households, it was based on the National Population Register (NPR) from the 2011 Population Census. The Government of India made the SECC publicly available on the internet in a mix of PDF and Excel formats. See Figure A1 for a de-identified sample page for a single household. We scraped over two million files, parsed the files into text data, and translated these from twelve different Indian languages into English. At the individual level, these data contain variables describing age, gender, occupation, caste group, disability and marital status. Data on occupations are written free-form in the SECC; after translation, we cleaned and matched these descriptions to the 2004 National Classification of Occupations. At the household level, this dataset contains variables describing housing, landholdings, agricultural assets, household assets and sources of income.

We geocoded and matched 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.

IV.C Economic and population censuses

The Indian Ministry of Statistics and Programme Implementation (MoSPI) conducted the 4th, 5th and 6th Economic Censuses respectively in 1998, 2005 and 2013. The Economic Census is a complete enumeration of all economic establishments except those engaged in

[13] 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.

crop production and plantation; there is no minimum firm size, and both formal and informal establishments are included. 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 2001 and 2011. We aggregate the microdata to the village level to obtain a measure of employment in non-farm firms.

We also use data on demographics and village-level public goods (roads, electricity, schools, etc.) from the Population Censuses of 1991, 2001 and 2011.

IV.D Sample and summary statistics

Table 1 shows village-level summary statistics for the sample of villages matched across datasets. The first column shows results for villages without a paved access road in 2001, the second column for villages with a paved road, and the third column for the pooled sample. Over 30% of villages without paved roads in 2000 received a PMGSY road by 2012.[14] Across a wide range of variables, villages without roads have lower levels of other amenities. They are further from towns, have higher illiteracy rates and are half as likely to be electrified at baseline. These differences lend further evidence to our assertion of endogenous placement of transport infrastructure, and thus the need for careful empirics to identify the causal effect.

Figure 1 provides a visual representation of the major datasets used in this project, along with year-by-year counts of the number of villages receiving PMGSY roads for the years of this study (1998 - 2014). It demonstrates that PMGSY construction is negligible before our baseline data in 2001, then slowly ramps up to a peak of over 11,000 roads constructed annually in 2008 before slowing down slightly.

[14]Approximately 20% of villages that were recorded as having a paved road in the 2001 Population Census also received PMGSY roads by 2012. This appears to have been due both to measurement error in the Population Census variables and to upgrades that were performed on existing roads.

V Empirical Strategy

For several reasons, the impacts of infrastructural investments have often proved challenging for economists to assess. First, the high cost and large potential returns of such investments mean that few policymakers are willing to allow random targeting. Political favoritism, economic potential and pro-poor targeting would lead infrastructure to be correlated with other government programs and economic growth, biasing OLS estimates in an unknown direction. Second, data are rarely available at the level of road placement, particularly in the case of rural roads. Third, the impacts of infrastructure are likely to depend on local and regional economic factors, necessitating a large sample of roads and regions to have sufficient power for tests of heterogeneity. In this section we describe the empirical strategy for the causal estimation of the impact of new rural roads.

Identification comes from the guidelines by which villages are prioritized to receive new roads. State implementing officials were instructed to target villages in the following order: (i) villages with population greater than 1000; (ii) villages with populations greater than 500; and (iii) villages with populations greater than 250. Even if selection into road treatment is partly determined by political or economic factors, these factors are not likely to change discontinuously at these population thresholds. If these rules were followed to any degree by state officials, the likelihood of treatment will discontinuously increase at these population thresholds, making it possible to estimate the effect of the program using a fuzzy regression discontinuity design.

Under the assumption of continuity at the treatment threshold, the fuzzy RD estimator (Imbens and Lemieux, 2008) estimates the local average treatment effect (LATE) of receiving a new road, for a village with population equal to the threshold:

$$\tau = \frac{\lim_{pop \to T^+} \mathbb{E}[Y_v | pop_v = T] - \lim_{pop \to T^-} \mathbb{E}[Y_v | pop_v = T]}{\lim_{pop \to T^+} \mathbb{E}[newroad_v | pop_v = T] - \lim_{pop \to T^-} \mathbb{E}[newroad_v | pop_v = T]}, \tag{1}$$

where Y_v is the outcome of interest, pop_v is the baseline village population, T is the population threshold, and $newroad_v$ is an indicator variable for whether village v received a new road in the sample period. The treatment effect can be interpreted as the discontinuous change in the outcome variable at the population threshold (the numerator) divided by the discontinuous change in the probability of treatment (the denominator).[15] The LATE estimated by our empirical design is specific to the complier set, namely those villages whose treatment status would be zero with population below the threshold and one with population above.

Our estimation follows the recommendations of Imbens and Lemieux (2008), Imbens and Kalyanaraman (2012) and Gelman and Imbens (2014). Our preferred specification uses local linear regression to control for the running variable (village population) on either side of the threshold. We restrict our sample to those villages whose population is within a certain bandwidth around the threshold, formally $pop_v \in [T - h; T + h]$, where h is the value of the bandwidth around threshold T. Using the methodology of Imbens and Kalyanaraman (2012), we calculate an optimal bandwidth of 84 and use a triangular kernel that places the most weight on observations close to the cutoff, as in Dell (2015).[16]

We begin by estimating the following reduced form fuzzy RDD specification:

$$Y_{v,j} = \beta_0 + \beta_1 1\{pop_{v,j} \geq T\} + \beta_2(pop_{v,j} - T) + \beta_3(pop_{v,j} - T) * 1\{pop_{v,j} \geq T\} + \zeta X_{v,j} + \eta_j + \epsilon_{v,j}, \tag{2}$$

where $Y_{v,j}$ is the outcome of interest in village v and group j, T is the population threshold, $pop_{v,j}$ is baseline village population, $X_{v,j}$ is a vector of village controls measured at baseline, and η_j is a group fixed effect, discussed below. Village controls and fixed effects are not

[15]Our design is a "fuzzy" regression discontinuity design (RDD) because the change in the probability of treatment at the threshold is less than one. Other program rules also guided road prioritization and compliance was imperfect.

[16]Results are robust to alternate bandwidths, as described below. Following the methodology of (Calonico et al., 2014), the optimal bandwidth is 78, which (unsurprisingly given the small difference) does not appreciably change the results. Results using alternative weighting functions and thresholds are available from the authors upon request.

necessary for identification but improve the efficiency of the estimation. Having subtracted the threshold T from the population controls, the change in outcome $Y_{v,j}$ for a village at the population threshold T is captured by β_1.

We make the following choices when estimating this model. In the first stage regression, in which we estimate the change in the probability of treatment, $Y_{v,j}$ is a dummy variable that takes the value one if the village has received a new road before 2012, the year of our primary outcome data.[17] For regressions in which we estimate the reduced form effect of road prioritization (i.e. being to the right of the population threshold) on economic outcomes, we discuss the definition of outcome variables as we present the results in Section VI. The vector of village controls, $X_{v,j}$, contains several village characteristics as measured in the 2001 Population Census (indicators for village amenities – primary school, medical center and electrification – the log of total agricultural land area, the share of agricultural land that is irrigated, distance in km from the closest census town, share of workers in agriculture, the illiteracy rate and the share of inhabitants that belong to a scheduled caste) and 2002 BPL Census (the share of households owning agricultural land, the share of households in subsistence agriculture, share of households earning over 4 USD cash per month, the share of households without a migrant and the share of households with only male labor). For η_j, we use district-cutoff fixed effects.[18] As the objective of this paper is to estimate the impact of receiving a paved road for the first time, we restrict our sample to villages that did not have a paved road at the start of the program.[19] Our final sample of 11,474 villages is comprised of villages that we

[17]This is the year that most data was collected for the SECC. When estimating outcomes measured in a different year, such as in the Population Census, we use the appropriate year of measurement for that particular set of regressions.

[18]Results are robust to alternative specifications using state or district fixed effects, and are available from the authors upon request.

[19]While unconnected villages were to be prioritized over those that already had some paved road, many already connected villages still received roads under the program. This is partly because road upgrades were also allowed under the rules and partly because program rules were not entirely followed. We define our sample of unconnected villages to be those that were recorded as lacking a paved road in either the 2001 Population Census (whose village amenities were recorded in 2000) or for all habitations in a village in the PMGSY administrative data.

match across our primary datasets (Population Censuses, BPL Census and Socioeconomic and Caste Census), which do not have a paved road at baseline and that fall within the optimal bandwidth of the population cutoffs for the states that followed these rules (see below).[20]

We understand the reduced form effect of road prioritization to be treatment effect of a new road times the discontinuous change in the probability of road treatment at the population threshold. To estimate the treatment effect directly, we use the following fuzzy RDD specification in which we instrument for treatment ($newroad_{v,j}$) with our road prioritization indicator variable $1\{pop_{v,j} \geq T\}$.

$$Y_{v,j} = \gamma_0 + \gamma_1 newroad_{v,j} + \gamma_2 pop_{v,j} + \gamma_3 pop_{v,j} * 1\{pop_{v,j} \geq T\} + \zeta X_{v,j} + \eta_j + \upsilon_{v,j}. \quad (3)$$

We estimate this equation using two stage least squares, where the first stage comes from Equation 2, with $newroad_{v,j}$ as the dependent variable.

The road program used multiple population thresholds to determine road prioritization: 1000, 500 and 250. Very few villages around the 250 population threshold received roads by 2012, so we limit our sample to villages with populations close to 500 and 1000. Further, only certain states followed the population threshold prioritization rules as given by the national guidelines of the PMGSY. We worked closely with the National Rural Roads Development Agency to identify the state-specific thresholds that were followed and define our sample accordingly. Our sample is comprised of villages from the following states, with the population thresholds used in parentheses: Chhattisgarh (500, 1000), Gujarat (500), Madhya Pradesh (500, 1000), Maharashtra (500), Orissa (500), and Rajasthan (500).[21] To maximize power, we pool our samples, using the same optimal bandwidth (84) for villages close to the 500

[20]We consider the Socioeconomic and Caste Census to be matched to our dataset when we were able to match it on names to the Population Census but also when the population reported in the SECC was within 20% of the population census data.

[21]These states are concentrated in north India. Southern states generally have far superior infrastructure and thus had few unconnected villages to prioritize. Other states such as Bihar had many unconnected villages but did not comply with program guidelines.

and 1000 thresholds.

The fuzzy regression discontinuity approach identifies the treatment effect of new roads under the assumption that crossing the population threshold discontinuously affects the probability of receiving a road, and nothing else of significance. We follow Imbens and Lemieux (2008) in testing for discontinuities in baseline covariates and in the density of the running variable at the population thresholds. Other threats to identification that rely on outcome variables are discussed below, in Section VI.C.

We first show that there are no discontinuities in baseline village characteristics. Table 2 presents the mean values for various village baseline characteristics, including the set of controls that we use in all regressions. Unsurprisingly, there are differences between the villages above and below the population threshold, as many village characteristics are correlated with village size. Reassuringly, however, we find no significant differences once we control for the covariates used in the fuzzy RDD specification. Figure 2 shows how our control variables vary at the cutoff, plotting the residuals from the set of controls (excluding the variable in question, running variable controls and the road prioritization indicator) and fixed effects used in our main specification against normalized village population. The black lines show a linear fit, estimated separately on either side of the cutoff, and the grey lines show the 95% confidence interval. The graphs show that baseline village characteristics are continuous at the thresholds.

We also investigate the possibility of manipulation of the running variable. We find evidence of considerable manipulation of habitation population in the official program data.[22] To resolve this issue, we instead use village population from the 2001 Population Census. Figure 3 displays two representations of the distribution of village populations in our sample,

[22]Figure A2 shows the distribution of habitation population as reported to the PMGSY, with implementation cutoffs indicated with vertical lines. There are noticeable discontinuities in density at the thresholds, suggesting that selection into treatment is not as good as random around these population cutoffs. For example, villages that are politically connected or more strategic may be able to report their population as just above 1000, even if it is not in reality. If this is occurring, the RDD approach cannot distinguish the effect of a new road from the effect of political influence.

using data from the Population Census. In the left panel, there are no noticeable discontinuities at the program prioritization thresholds. We test this formally by testing for a discontinuity in the running variable (village population) around the threshold for the pooled sample, following McCrary (2008). We estimate a discontinuity of -0.01 with a standard error of .05, failing to reject the null hypothesis of no discontinuity in the running variable.

We next examine the first stage, showing that there is a large and highly significant jump in the probability of a road road by 2012 at the population cutoff. Table 3 presents first stage estimates of the change in probability of treatment across different bandwidths h. The estimates are highly stable. Across bandwidths, there is a 21-22 percentage point increase in the probability of treatment around the cutoff. Figure 4 shows these results graphically for the optimal threshold as a scatterplot of population bin means. This graph confirms the results from Table 3: at the pooled priority threshold, there is a significant increase in the probability of treatment of approximately 21 percentage points.

VI Results

In this section, we describe and discuss the main results (Section VI.A) and robustness checks (Section VI.C). We first show that rural road construction leads to a reallocation of labor out of agriculture and into manual labor, a result that is larger for households and individuals with high potential returns to labor market participation. We then consider the mechanisms that could explain these results, finding that that the evidence best supports increased access to labor markets outside of the village.

VI.A Main results

We begin by estimating the effect of a new road on occupational choice. As approximately 92% of workers in our sample villages report their occupation to be either agricultural or manual labor, we focus our investigation on these categories. Outcomes $Y_{v,j}$ are defined as the share of workers who report working either in an agricultural occupation (cultivation, farmer, agricultural labor, or any other occupation mentioning agriculture) or in manual labor.

We find that a new road is associated with a significant occupational reallocation out of agricultural activities and into manual labor. Table 4 presents regression discontinuity estimates of the impact of a new road on the share of workers reporting occupations in agriculture (Panel A) and manual labor (Panel B). The result is stable across six different bandwidths ranging from 60 to 110.[23] New roads cause a large and stable reduction in the share of workers in agriculture (point estimates range from a reduction of 8.3 to 10.1 percentage points, depending on bandwidth) and a corresponding increase in the share of workers in manual labor (5.4 to 8.2 percentage point increase). Figure 5 presents the optimal bandwidth reduced form estimate graphically, demonstrating the significant drop in the share of workers in agriculture to the right of the population cutoff.

We restrict our analysis from this point forward to using the optimal bandwidth of 84. We next compare results using our primary measure of structural transformation (share of workers in agricultural occupations) to an alternate measure (share of households reporting cultivation as their primary income source). Table 5 presents regression discontinuity estimates from Equation 3 of the effect of a new road on these two measures. The first two columns present the impact on our primary measure, the share of workers (aged 21-60) in agriculture and manual labor. We find a 10.1 percentage point reduction in workers in agriculture (representing a 21% decrease from the control group mean) and an 8.0 percentage point increase in workers in (non-agricultural) manual labor. In contrast, the results from the third and fourth columns show that household income source does not change significantly, suggesting that many of the workers exiting agriculture are not the primary earners in the household. This is consistent with descriptive evidence from South Korean, where Kim and Topel (1995) find that non-agricultural firms tended to hire new entrants into the labor force rather than former farmers. We further examine the characteristics of the workers who

[23] This range contains both the optimal IK (84) and CCT (78) bandwidths.

change occupations later in this section.

Recent evidence has demonstrated that a reduction in transportation costs can lead to significant increases in out-migration from rural areas (Bryan et al., 2014; Morten and Oliveira, 2016). Although we are not able to measure migration choices directly, we examine three proxies that should be highly correlated with permanent migration. First we test for impacts on village population growth (Table A1). We find no evidence for significant impacts on total population growth, either in levels or annualized growth rate between the 2001 and 2011 Censuses, and can reject a change in the growth rate of 0.4%. The limitation of population growth as an outcome is that any impacts on net migration could be offset by changes to fertility and mortality. We hypothesize that migration, fertility and mortality are likely to affect different segments of the population; thus, even if they are offsetting each other in terms of total population, we should see changes in the demographics of the village. We test this prediction by estimating the impact of new roads on the age distribution (in ten year bins) and the share of each age bin that is male. In Table A2, we find no evidence either of changes to the age distribution or gender ratios in any part of that distribution. Taken together, these three pieces of evidence that new roads do not lead to major changes in out-migration.[24] The absence of demographic effects allows us to rule out large-scale migration and interpret the observed sectoral reallocation of labor as the result of changes in occupational choice rather than compositional effects due to selective migration.

The model suggests that those who exit agriculture in favor of non-farm labor market opportunities will be those for whom the losses of agricultural income are smallest and the labor market gains are largest. By using individual-level census data, we can examine the distribution of treatment effects across subgroups with different factor endowments. As the dominant sector of the rural Indian economy is agriculture, land endowments may play a major role

[24] We suspect that this difference with Morten and Oliveira (2016) is due to the difference in road type: the construction of a paved rural road is unlikely to significantly change the one-time cost of permanent migration relative to the lifetime benefits, in contrast to the major changes induced by highway construction.

in determining which workers respond most to a rural road. We first examine the impact of road construction on the landholding distribution in Table A3. We find that a new road does not significantly change the share of households that are landless, own less than 2 acres, or have between 2 and 4 acres of agricultural land. However, we do find a 3.4 percentage point increase in the share of households with over four acres of land (significant at the 10 percent level). We are hesitant to over-interpret one marginally significant result out of four tests, but it is possible that there is some land consolidation following the construction of a road. Regardless, we do not find major changes in the landholding distribution and thus treat ex post observed landholdings as a baseline variable upon which to conduct heterogeneity analysis. Panel A of Table 6 presents our main specification, estimating the effect separately by size of landholdings. We find that movement out of agriculture is strongest for workers in households without land, and that this effect is monotonically decreasing in landholding size.[25] The decrease in agriculture for those with no land (12.2 percentage points) is much larger as a percentage of the control group mean: our estimates suggest that 35% of workers with no land exit agriculture, compared to just 10% in households with more than four acres of land.[26] These results are consistent with recent work finding that the inheritance of land in India can significantly reduce rates of migration and participation in non-agricultural occupations (Fernando, 2016) and suggest that the lack of transport infrastructure may be one cause of the inefficiently small size of many farms in rural India (Foster and Rosenzweig, 2011). These effects also suggest that new roads may be a progressive investment in that those with the least agricultural wealth (as proxied by landholding) show the largest labor market effects.

We next examine the heterogeneity of the treatment effect as a function of age and gender (Table 6, Panel B). There are no differential results by age: the point estimate for workers

[25]We cannot however statistically reject equality between any of these estimates.

[26]It is important to note that productivity in agriculture will only depend on landholdings if there are market failures such that it is more productive to work on one's own land. An extensive literature investigates common failures in agricultural land and labor markets in low income countries. See, for example, de Janvry et al. (1991).

aged 21-40 (a 9.8 percentage point decrease in the share in agriculture) is almost identical to the effect for workers aged 41-60 (a 9.5 percentage point decrease). While the differences are not significantly different, we do find that men are more likely to exit agriculture as compared to women, particularly in the younger cohort (-9.6 percentage point effect for men compared to -3.8 percentage point for women). These estimates could be the result of a male physical advantage in non-agricultural work or attitudes against women's working far away from home that may prevent reallocation of female labor away from agriculture (Goldin, 1995). However, as a percentage of the control group mean, the estimates for male and female workers are much closer.

VI.B Potential mechanisms and additional outcomes

We have thus far established the causal impact of rural roads on occupation choice. Rural roads lead to a large reallocation of labor out of agriculture and into manual labor. There are two possibilities for the destination of workers leaving agriculture upon road treatment: increasing employment in non-agricultural firms in the village, and increased participation in external labor markets. We test between these mechanisms by estimating growth in employment in nonfarm firms, using the 2013 Economic Census. We use two measures of nonfarm in-village employment from this dataset. First, in order to generate a result comparable to our main result (a 10 percentage point decline in the share of workers in agriculture), we examine the share of total workers employed by in-village nonfarm firms. Second, we estimate the impact on log employment in these firms.[27] Table 7 presents these results.[28] Panel A provides estimates for employment. In column 1 we estimate that a new road leads to a statistically insignificant 1.3 percentage point decrease in the share of village workers in village nonfarm firms. This is far too small to explain the observed exit from

[27]Because some villages have 0 employment by this measure, we add 1 to the employment before the log transformation.

[28]We define the sample in the way as in earlier tables, but additionally trim outliers to eliminate villages where the number of workers in village nonfarm firms is greater than the total number of workers resident in the village.

agriculture, meaning that the majority of workers leaving agriculture are finding work outside of the village.[29] Columns 2-4 estimate impacts on log employment in all firms, tradable sector firms and non-tradable sector firms, respectively. We estimate a total effect on employment in nonfarm firms of 22 log points (not significant at the 10 percent level), with results nearly twice as large for tradable as for non-tradable firms. While this points towards firm growth as being driven by increased productivity in tradable firms, our estimates are imprecise and we lack the data on inputs, productivity and prices to shed further light on these results. In Panel B, we generate equivalent estimates for the number of firms. We find that the number of firms grows by 24 log points (significant at the 10 percent level). Again, this result is larger for tradable firms, although the difference with non-tradable firms is not significant.[30]

One possibility outside the scope of our model is that labor saving agricultural investments may actually lower demand for labor in the agricultural sector. Much of the existing literature on rural roads focuses on agricultural outcomes, finding evidence that connectivity results in increased agricultural land values (Jacoby, 2000), increased productivity (Sotelo, 2016) and lower market prices for agricultural output (Casaburi et al., 2013). Bustos et al. (2016) find that technical change in soy production in Brazil was strongly labor saving, leading to a reduction in the agricultural share of the workforce in soy growing areas. Our data allow us to test for three such changes to agricultural production that may reduce the demand for agricultural labor: ownership of mechanized farm equipment (tractors, etc), ownership of irrigation equipment, and the exit of households from land ownership. In Table 8, we present estimates of the impact of a new road on these outcomes. Column 1 estimates the impact of a new road on the share of households owning mechanized farm equipment, column 2 on

[29] We also find that a new road increases the likelihood of a regular bus route serving the village, suggesting that the road does facilitate greater passenger traffic to external markets. Results available upon request. Asher et al. (2016) uses this outcome as an empirical application of the use of classification trees to examine heterogeneous treatment effects in moment-based models.

[30] In Table A4 we test whether roads lead to changes in the share of working age adults that are either not working or whose occupations we are unable to classify. We find no changes in either outcome, establishing that our results are not driven by labor force exit or other data issues.

the share of households owning irrigation equipment, and column 3 on the share of households owning agricultural land. We find no significant impacts on any of these measures, suggesting that capital intensification of agriculture is not taking place, and thus a reduction of labor demand in agriculture is unlikely.[31]

The model predicts that labor reallocation will be strongest in villages with low agricultural productivity. Table 9 presents estimates of the impact of a new road on the share of workers in agriculture (columns 1 and 2) and on log employment in nonfarm village firms (columns 3 and 4), with the sample split by median agricultural productivity. We use as our measure of agricultural productivity the caloric productivity per acre, as estimated by the Food and Agriculture Organization's GAEZ project.[32] We find that that the point estimate on the impact of road construction on the share of workers in agriculture is over twice as large for the below-median agricultural productivity sample, although the difference is not statistically significant. We find similar results for growth of employment in village firms in the nonfarm sector: the low agricultural productivity sample experiences 45 log point employment growth (significant at the 5 percent level), compared to an insignificant 6 log points in the high productivity sample.

Finally, we use asset data to examine whether new road construction causes improvements in economic outcomes. Table 10 presents estimates on asset ownership using our main regression discontinuity specification. We create a village-level asset index using the four household assets listed in the 2012 Socioeconomic and Caste Census: a house of solid material (having both solid walls and roof), a refrigerator, a motorized vehicle, and a phone. All four components of the index are the share of households in the village having that asset, and have been

[31] As discussed earlier, we do find one piece of evidence for changes in household landholdings: a 3.4 percentage point increase (significant at the 10 percent level) in the share of households with more than 4 acres of land. More detailed data on agricultural production would be required for further investigation of this possibility.

[32] See Costinot et al. (2016) for a detailed description of this dataset. As cereals are the dominant crops in India and this project focuses on the least connected parts of rural India, we use the measure of rain-fed cereal productivity per acre under low agricultural inputs.

standardized to have a mean of zero and standard deviation of one. Column 1 presents results using this measure, while columns 2-5 show the results for the share of households owning each of the assets. Point estimates are positive for all assets but are not statistically significant. The p-value for the asset index is 0.21. It remains possible that occupations and potentially even earnings change quickly but that durable assets take more time to accumulate.[33]

VI.C Robustness

In this section we explore the possibility that factors other than the road treatment may be driving our results. Reassuringly, we find no evidence supporting such concerns.

As a placebo exercise, we run our first stage and reduced form estimation on the share of workers in agriculture for the set of villages not in our main sample, where there is not a discontinuous increase in road treatment at the population threshold. If other determinants of sectoral allocation varied discontinuously at the treatment threshold, we might incorrectly attribute their effects to rural road treatment. This sample is defined as villages for which we have data and close to the population cutoffs (500, 1000) in major states that did not follow the rules at all (Andhra Pradesh, Assam, Bihar) and villages close to the 1000 cutoff in states that only followed the 500 population cutoff: Gujarat, Maharashtra, Orissa and Rajasthan. Table A5 presents the estimates of these regressions. There is no evidence of either a first stage or reduced form effect on agricultural labor share for the placebo sample, indicating that our results are not due to other discontinuous differences in villages whose effect we spuriously attribute to new roads.

A different threat to our identification could come from any other policy that used the

[33]These measures of economic outcomes may underestimate the welfare effects of increased nonfarm labor market participation if, for example, wages serve as insurance against agricultural risk, as demonstrated by Kochar (1999). It is also possible that roads lower the volatility of agricultural income, as demonstrated Allen and Atkin (2016) with respect to Indian highway construction. Another way that we may fail to accurately estimate the impact on welfare is due to our inability to measure consumption, which many researchers have argued is a better proxy for welfare than income. See Meyer and Sullivan (2003) for a discussion of the trade-offs between these different measures. Finally, it may be that the medium-run effects that we measure (average time between road construction and outcome measurement is 3.5 years) are lower than the long-run impacts.

same thresholds as the PMGSY. In fact, one national government program did prioritize villages above 1000 population: the Total Sanitation Campaign (Spears, 2015), which incentivizes rural local governments to improve sanitation. We present two reasons why it is highly unlikely that this program is spuriously driving our results. First, there is little theoretical reason to believe that investments in sanitation could drive a large reallocation of labor away from agriculture. Second, in Table A6 we present reduced form estimates of the impact of road prioritization on four measures of sanitation. We find no evidence that being above the 1000 population threshold is associated with improved outcomes in any of these measures.

VII Conclusion

Access to the outside world via paved roads, easily taken for granted in many rich countries, is far from a reality for many of the world's rural poor. High transportation costs potentially inhibit gains from the division of labor, economies of scale and specialization. Recent work has begun to demonstrate the role of trunk infrastructure (railroads and highways) on economic activity. However, despite the emphasis of both theorists and development policymakers on the importance of transportation costs, causal estimates of the impact of rural roads have been difficult to generate.

In this paper we estimate the economic impacts of the Pradhan Mantri Gram Sadak Yojana, a large-scale program in India with the objective to provide universal access to paved "all-weather" roads in rural India. We exploit discontinuities in the probability of paved road construction at village population thresholds, finding that new paved roads lead to a large reallocation of labor out of agriculture and into (manual) labor markets. Rather than facilitating growth of the nonfarm sector in rural areas, roads appear to facilitate the access of rural labor to external employment.

Why so many workers remain in low productivity agriculture when higher wages are available in other locations and sectors is a classic question in development economics. Our findings suggest that the poor state of rural transportation infrastructure in developing countries must be taken seriously as a barrier to the efficient allocation of labor across space and sectors. This should not be surprising: arbitrage is only possible if the costs of reallocation are less than the gains. We do not resolve the entire puzzle. Migration is an obvious alternate way of accessing labor market opportunities outside of the village. It is beyond the scope of this paper to examine the many potential barriers to migration, but recent research has suggested that transportation costs are an important factor (Morten and Oliveira, 2016; Bryan et al., 2014). We find no evidence of a rise in migration, lending credence to research proposing factors other than the state of rural transport infrastructure to explain India's low rates of rural-urban migration (see, for example, Munshi and Rosenzweig (2016)).

Foster and Rosenzweig (2007) argue that economists do not adequately understand the flows of capital and labor between rural and urban areas in developing countries. This paper adds to a growing literature on the linkages in labor markets across space, and suggests that transportation infrastructure may be an important determinant of such flows. However, a limitation of this paper is that we cannot study these flows directly. We hope that future research will shed light on the nature and causes of labor flows across sectors and between rural and urban areas. There is increasing evidence that rural workers are an important component of urban labor supply (Imbert and Papp, 2016). If so, the impacts of improving transportation linkages between rural and urban areas will also be felt by urban inhabitants, with potentially large consequences for both urban wages and firm behavior.

Many researchers have puzzled over India's low rates of urbanization and structural transformation when compared to other developing countries. This paper provides evidence that workers can participate in non-agricultural labor markets without moving to cities when there is adequate access to external labor markets. India's high population densities and superior infrastructure may help to explain why its structural transformation has approximately matched the speed of sub-Saharan Africa while urbanizing much more slowly. At the same time, our results suggest that India's low rate of structural transformation when compared to China may be due in part to its much lower rate of investment in transportation infrastructure. More research is needed to understand the policies that have enabled structural transformation away from low productivity agriculture in certain low-income countries and not in others.

References

Adukia, Anjali, Sam Asher, and Paul Novosad, "Educational Investment Responses to Economic Opportunity: Evidence from Indian Road Construction," 2017. Working paper.

Aggarwal, Shilpa, "Do Rural Roads Create Pathways out of Poverty? Evidence from India," 2015. Working Paper.

Ali, Rubaba, "Impact of Rural Road Improvement on High Yield Variety Technology Adoption: Evidence from Bangladesh," 2011. Working paper.

Alkire, Sabina and Suman Seth, "Identifying BPL Households: A Comparison of Methods," *Economic and Political Weekly*, 2013, *48* (2), 49–57.

Allen, Treb and David Atkin, "Volatility, Insurance, and the Gains from Trade," 2016. Working paper.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P Ryan, "Classification Trees for Heterogeneous Moment-Based Models," 2016. NBER Working Paper No. 22976.

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen, "Markup and Cost Dispersion across Firms: Direct Evidence from Producer Surveys in Pakistan," *The American Economic Review*, 2015, *105* (5). Working Paper.

Banerjee, Abhijit and Andrew F. Newman, "Information, the Dual Economy, and Development," *The Review of Economic Studies*, 1998, *65* (4), 631–653.

—— **, Esther Duflo, and Nancy Qian**, "On the Road: Access to Transportation Infrastructure and Economic Growth in China," 2012. NBER Working Paper No. 17897.

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

Baum-Snow, Nathaniel, Loren Brandt, J. Vernon Henderson, and Matthew A Turner, "Roads, Railroads and Decentralization of Chinese Cities," *Review of Economics and Statistics*, 2017.

Bazzi, Samuel, Arya Gaduh, Alexander Rothenberg, and Maisy Wong, "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *The American Economic Review*, 2016, *106* (9).

Binswanger, Hans P., Shahidur R. Khandker, and Mark R. Rosenzweig, "How infrastructure and financial institutions affect agricultural output and investment in India," *Journal of Development Economics*, 1993, *41* (2).

Blimpo, M. P., R. Harding, and L. Wantchekon, "Public Investment in Rural Infrastructure: Some Political Economy Considerations," *Journal of African Economies*, 2013, *22* (AERC Supplement 2).

Brueckner, Markus, "Infrastructure, Anocracy, and Economic Growth: Evidence from International Oil Price Shocks," 2014. Working paper.

Bryan, Gharad and Melanie Morten, "Economic Development and the Spatial Allocation of Labor: Evidence From Indonesia," 2015. Working paper.

—— **, Shyamal Chowdury, and Ahmed Mushfiq Mobarak**, "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh," *Econometrica*, 2014, *82* (5).

Burgess, Robin and Dave Donaldson, "Railroads and the Demise of Famine in Colonial India," 2012. Working paper.

—— **, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padro i Miquel**, "The Value of Democracy: Evidence from Road Building in Kenya," *American Economic*

Review, 2015, *105* (6).

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli, "Agricultural Productivity and Structural Transformation. Evidence from Brazil," *The American Economic Review*, 2016, *106* (6), 1320–1365.

Calonico, Sebastian, Md Cattaneo, and Rocio Titiunik, "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 2014, *82* (6), 2295–2326.

Casaburi, Lorenzo, Rachel Glennerster, and Tavneet Suri, "Rural Roads and Intermediated Trade: Regression Discontinuity Evidence from Sierra Leone," 2013. Working Paper.

Caselli, Francesco, "Accounting for Cross-Country Income Differences," in "Handbook of Economic Growth," Vol. 1 2005, chapter 9, pp. 679–741.

___ and Wilbur John II Coleman, "The U.S. Structural Transformation and Regional Convergence: A Reinterpretation," *Journal of Political Economy*, 2001, *109* (3), 584–616.

Costinot, Arnaud, Dave Donaldson, 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).

de Janvry, Alain, Marcel Fafchamps, and Elisabeth Sadoulet, "Peasant household behaviour with missing markets: Some Paradoxes Explained," *The Economic Journal*, 1991, *101* (409), 1400–1417.

Dell, Melissa, "Trafficking Networks and the Mexican Drug War," *American Economic Review*, 2015, *105* (6).

Dercon, Stefan, Daniel O. Gilligan, John Hoddinott, and Tassew Woldehanna, "The impact of agricultural extension and roads on poverty and consumption growth in fifteen Ethiopian villages," *American Journal of Agricultural Economics*, 2009, *91* (4), 1007–1021.

Dinkelman, Taryn, "The Effects of Rural Electrification on Employment: New Evidence from South Africa," *American Economic Review*, 2011, *101* (7), 3078–3108.

Donaldson, Dave, "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review (forthcoming).*

___ and Richard Hornbeck, "Railroads and American Economic Growth: A "Market Access" Approach," *Quarterly Journal of Economics*, 2016, *131* (2).

Duflo, Esther and Rohini Pande, "Dams," *Quarterly Journal of Economics*, 2007, *122* (2), 601–646.

Einav, Liran and Jonathan Levin, "Economics in the Age of Big Data," *Science*, 2014, *346* (6210), 1–5.

Faber, Benjamin, "Trade Integration, Market Size, And Industrialization: Evidence From China's National Trunk Highway System," *Review of Economic Studies*, 2014, *81* (3).

Fafchamps, Marcel and Forhad Shilpi, "Cities and Specialisation: Evidence from South Asia," *The Economic Journal*, 2005, *115* (503), 477–504.

Fan, Shenggen and Peter Hazell, "Returns to Public Investments in the Less-Favored Areas of India and China," *American Journal of Agricultural Economics*, 2001, *83* (5), 1217–1222.

Fernando, A. Nilesh, "Shackled to the Soil: The Long-Term Effects of Inherited Land on Labor Mobility and Consumption," 2016. Working paper.

Foster, Andrew D. and Mark R. Rosenzweig, "Economic Development and the Decline of Agricultural Employment," in "Handbook of Development Economics," Vol. 4 2007, chapter 47, pp. 3051–3083.

—— **and** —— , "Are Indian Farms Too Small? Mechanization, Agency Costs, and Farm Efficiency," 2011. Working paper.

Gelman, Andrew and Guido Imbens, "Why high-order polynomials should not be used in regression discontinuity designs," 2014. NBER Working Paper No. 20405.

Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr, "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing," *Economic Journal*, 2015.

Gibson, John and Scott Rozelle, "Poverty and Access to Roads in Papua New Guinea," *Economic Development and Cultural Change*, 2003, *52* (1), 159–185.

—— **and Susan Olivia**, "The effect of infrastructure access and quality on non-farm enterprises in rural Indonesia," *World Development*, 2010, *38* (5), 717–726.

Goldin, Claudia, "The U-shaped female labour force function in economic development and economic history," in T. Paul Schultz, ed., *Investment in Women's Human Capital*, Chicago and London: University of Chicago Press, 1995.

Gollin, Douglas and Richard Rogerson, "Productivity, transport costs and subsistence agriculture," *Journal of Development Economics*, 2014, *107*, 38–48.

—— **, David Lagakos, and Michael E Waugh**, "The Agricultural Productivity Gap," *Quarterly Journal of Economics*, 2014, *129* (2), 939–993.

Harris, John R. and Michael P. Todaro, "Migration, unemployment and development: a two-sector analysis," *The American Economic Review*, 1970, *60* (1), 126–142.

Imbens, Guido and K. Kalyanaraman, "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 2012, *79* (3).

Imbens, Guido W. and Thomas Lemieux, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 2008, *142* (2), 615–635.

Imbert, Clement and John Papp, "Short-term Migration, Rural Workfare Programs and Urban Labor Markets: Evidence from India," 2016. Working paper.

Jacoby, Hanan G, "Access to Markets and the Benefits of Rural Roads," *The Economic Journal*, 2000, *110* (465), 713–737.

—— **and Bart Minten**, "On measuring the benefits of lower transport costs," *Journal of Development Economics*, 2009, *89*, 28–38.

Khandker, Shaidur R. and Gayatri B Koolwal, "Estimating the Long-term Impacts of Rural Roads: A Dynamic Panel Approach," 2011. World Bank Policy Research Paper No. 5867.

—— **, Zaid Bakht, and Gayatri B. Koolwal**, "The Poverty Impact of Rural Roads: Evidence from Bangladesh," *Economic Development and Cultural Change*, 2009, *57* (4), 685–722.

Kim, Dae-Il and Robert H Topel, "Labor Markets and Economic Growth: Lessons from Korea's Industrialization, 1970-1990," in Richard B. Freeman and Lawrence F. Katz, eds., *Differences and Changes in Wage Structures*, University of Chicago Press, 1995, chapter 7, pp. 227–264.

Kochar, Anjini, "Smoothing Consumption by Smoothing Income: Hours-of-Work Responses to Idiosyncratic Agricultural Shocks in Rural India," *The Review of Economics and Statistics*, 1999, *81* (1), 50–61.

La Porta, Rafael and Andrei Shleifer, "Informality and Development," *Journal of Economic Perspectives*, 2014, *28* (3), 109–126.

Lewis, W. Arthur, "Economic Development with Unlimited Supplies of Labour," *The*

Manchester School, 1954, *22* (2), 139–191.

Lipscomb, Molly, Ahmed Mushfiq Mobarak, and Tania Bahram, "Development effects of electrification: evidence from the geologic placement of hydropower plants in Brasil," *American Economic Journal: Applied Economics*, 2013, *5* (2), 200–231.

McCrary, Justin, "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

McMillan, Margaret, Dani Rodrik, and Iñigo Verduzco-Gallo, "Globalization, Structural Change, and Productivity Growth, with an Update on Africa," *World Development*, 2014, *63*, 11–32.

Meyer, Bruce D. and James X. Sullivan, "Measuring the Well-Being of the Poor Using Income and Consumption," *The Journal of Human Resources*, 2003, *38* (2003).

Morten, Melanie and Jaqueline Oliveira, "The Effects of Roads on Trade and Migration: Evidence from a Planned Capital City," 2016. Working paper.

Mu, Ren and Dominique van de Walle, "Rural Roads and Local Market Development in Vietnam," *Journal of Development Studies*, 2011, *47* (5).

Mukherjee, Mukta, "Do Better Roads Increase School Enrollment? Evidence from a Unique Road Policy in India," 2011. Working paper.

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

National Rural Roads Development Agency, "Pradhan Mantri Gram Sadak Yojana - Operations Manual," Technical Report, Ministry of Rural Development, Government of India 2005.

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.

Shrestha, Slesh A., "The Effect of Roads on Farmland Values: Evidence from the Topography-based Road Network in Nepal," 2015. Working paper.

Sotelo, Sebastian, "Domestic Trade Frictions and Agriculture," 2016. Working paper.

Spears, Dean, "Effects of sanitation on early-life health: Evidence from a governance incentive in rural India," 2015. Working paper.

Storeygard, Adam, "Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa," *Review of Economic Studies*, 2016, *83* (3).

Wantchekon, Leonard and Piero Stanig, "The Curse of Good Soil? Land Fertility, Roads, and Rural Poverty in Africa," 2015. Working paper.

World Bank, "The Rural Access Index," 2015. http://www.worldbank.org/transport/transportresult.

Zhang, Xiaobo and Shenggen Fan, "How Productive is Infrastructure? A New Approach and Evidence from Rural India," *American Journal of Agricultural Economics*, 2004, *86* (2), 492–501.

Table 1: Summary statistics, by paved road at baseline

	No Road	Paved Road	Total
Primary school	0.876	0.956	0.917
	(0.329)	(0.206)	(0.275)
Medical center	0.232	0.530	0.387
	(0.422)	(0.499)	(0.487)
Electrified	0.419	0.748	0.591
	(0.493)	(0.434)	(0.492)
Distance from nearest town (in km)	24.91	20.85	22.79
	(21.95)	(17.87)	(20.04)
Land irrigated share	0.379	0.420	0.401
	(0.352)	(0.358)	(0.356)
Ln land area	5.054	5.712	5.397
	(1.055)	(1.078)	(1.117)
Literate share	0.440	0.506	0.474
	(0.155)	(0.133)	(0.148)
Scheduled caste share	0.162	0.176	0.169
	(0.179)	(0.162)	(0.170)
Share of households with land	0.723	0.673	0.697
	(0.243)	(0.254)	(0.250)
Share of HH with subsistence agriculture as primary income source	0.444	0.421	0.432
	(0.269)	(0.238)	(0.253)
Share of households earning 4 USD per month or above	0.792	0.830	0.812
	(0.268)	(0.226)	(0.248)
Population (2001)	973.6	1902.9	1457.2
	(1092.6)	(1947.1)	(1661.6)
Population (2011)	1164.1	2169.9	1687.5
	(1341.5)	(2253.9)	(1938.9)
Employment in nonfarm firms (1998)	33.81	119.7	78.49
	(83.65)	(349.9)	(262.5)
Number of nonfarm firms (1998)	17.27	54.27	36.53
	(32.72)	(94.77)	(74.36)
Employment in nonfarm firms (2013)	62.37	221.3	145.1
	(141.5)	(459.9)	(354.9)
Number of nonfarm firms (2013)	33.21	111.7	74.04
	(63.51)	(182.1)	(143.9)
PMGSY road by 2012	0.300	0.201	0.248
	(0.458)	(0.401)	(0.432)

Notes: This table presents means and standard deviations of baseline variables and outcomes. The first column presents summary statistics for villages without a paved road in the 2001 Population Census, the second column for villages with a paved road, and the third column for the pooled sample.

Table 2: Balance

Variable	Below threshold	Over threshold	Difference of means	p-value on difference	RD estimate	p-value on RD estimate
Primary school	0.950	0.961	0.011	0.003	-0.019	0.58
Medical center	0.153	0.177	0.023	0.001	-0.072	0.27
Electrified	0.411	0.445	0.034	0.000	-0.028	0.74
Distance from nearest town (km)	26.868	26.734	-0.135	0.747	-4.196	0.24
Land irrigated (share)	0.274	0.288	0.014	0.014	-0.012	0.79
Ln land area	5.115	5.228	0.114	0.000	-0.078	0.46
Literate (share)	0.453	0.460	0.007	0.010	-0.014	0.55
Scheduled caste (share)	0.141	0.144	0.003	0.314	-0.029	0.34
Land ownership (share)	0.739	0.737	-0.002	0.705	0.005	0.89
Subsistence ag (share)	0.443	0.439	-0.004	0.434	0.036	0.39
HH income >US$4 (share)	0.756	0.761	0.005	0.324	-0.012	0.80
N	6049	5425				

Notes: The table presents mean values for village characteristics, measured in the baseline period, for all variables used as controls in the main specification. The first nine variables come from the 2001 Population Census, while the final five come from the 2002 BPL Census. Columns 1 and 2 show the unconditional means for villages below and above the treatment threshold, respectively. Column 3 shows the difference of means across columns 1 and 2 and column 4 shows the p-value for the difference of means. Column 5 shows the regression discontinuity estimate, following the specification in Equation 3, of the effect of being above the treatment threshold on the baseline variable (with the outcome variable omitted from the set of controls), and column 6 is the p-value for this estimate, using heteroskedasticity robust standard errors. An optimal bandwidth of ± 84 around the population thresholds has been used to define the sample of villages (see text for details), such that the sample for the estimation are villages with a population in the range of 416-584 for the 500 threshold and 916-1084 for the 1000 threshold.

Table 3: First stage effect of road prioritization on road treatment

	±60	±70	±80	±90	±100	±110
Road priority	0.219	0.217	0.215	0.212	0.212	0.214
	(0.020)***	(0.018)***	(0.017)***	(0.016)***	(0.015)***	(0.014)***
F statistic	124.928	143.568	161.055	176.516	195.915	219.604
N	8145	9523	10899	12225	13608	15002
R2	0.30	0.30	0.30	0.29	0.29	0.29

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

Notes: This table presents first stage estimates from Equation 3 of the effect of being above the treatment threshold on a village's probability of treatment. The dependent variable is a indicator variable that takes on the value one if a village has received a PMGSY road before 2012. The first column presents results for villages with populations within 60 of the population threshold (440-560 for the low threshold and 940-1060 for the high threshold). The second through sixth columns expand the sample to include villages within 70, 80, 90, 100 and 110 of the population thresholds. The sample consists of villages that did not have a paved road at baseline (see text for details). The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 4: Impact of road on occupation, by bandwidth

Panel A. Share of workers in agriculture

	±60	±70	±80	±90	±100	±110
New road	-0.083	-0.095	-0.099	-0.101	-0.099	-0.091
	(0.0501)*	(0.0470)**	(0.0445)**	(0.0425)**	(0.0403)**	(0.0379)**
Control group mean	0.4686	0.4702	0.4704	0.4697	0.4706	0.4706
N	8099	9466	10838	12154	13531	14917
R2	0.2911	0.2827	0.2781	0.2748	0.2744	0.2758

Panel B. Share of workers in non-agricultural manual labor

	±60	±70	±80	±90	±100	±110
New road	0.0537	0.0696	0.0773	0.0815	0.0791	0.0735
	(0.0500)	(0.0469)	(0.0444)*	(0.0424)*	(0.0402)**	(0.0379)*
Control group mean	0.4541	0.4528	0.4524	0.4530	0.4525	0.4527
N	8099	9466	10838	12154	13531	14917
R2	0.2740	0.2661	0.2611	0.2580	0.2574	0.2574

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of a new road on occupational choice. Panel A presents regression discontinuity estimates for the share of workers reporting agriculture as their occupation while Panel B presents regression discontinuity estimates for the share of workers reporting manual labor as their occupation. The first column presents results for villages with populations within 60 of the population threshold. The second through sixth columns expand the sample to include villages within 70, 80, 90, 100 and 110 of the population thresholds. For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The sample consists of villages that did not have a paved road at baseline (see text for details). The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 5: Impact of new road on occupation and income source

	Occupation		Household Income Source	
	Agriculture	Manual Labor	Agriculture	Manual Labor
New road	-0.101	0.080	-0.033	-0.006
	(0.044)**	(0.044)*	(0.045)	(0.044)
Control group mean	0.476	0.449	0.419	0.506
N	11474	11474	11474	11474
R2	0.28	0.26	0.31	0.28

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on occupational choice and household source of income. Column 1 estimates the impact on the share of workers in agriculture. Column 2 estimates the effect on the share of workers in manual labor (excluding agriculture). Columns 3 and 4 provide estimates of the impact of a new road on the share of households reporting cultivation and manual labor as the primary source of income. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 6: Impact of new road on share of workers in agriculture, by household and worker characteristics

Panel A. Impact by household landholding

	Landless	0-2 Acres	2-4 Acres	4+ Acres
New road	-0.122***	-0.107**	-0.081*	-0.067
	0.038	0.042	0.043	0.043
Control group mean	0.351	0.513	0.590	0.654
N	11148	10731	10429	10000
R2	0.22	0.18	0.19	0.22

Panel B. Impact by age and gender

	All		Male		Female	
	21-40	41-60	21-40	41-60	21-40	41-60
New road	-0.098**	-0.095**	-0.096**	-0.095**	-0.038	-0.053
	0.046	0.046	0.045	0.045	0.058	0.062
Control group mean	0.430	0.578	0.450	0.611	0.269	0.330
N	11464	11423	11453	11413	10820	10226
R2	0.27	0.29	0.27	0.28	0.21	0.23

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on occupational choice. The dependent variable in each regression is the share of workers in agriculture, for that specific category. Panel A examines whether treatments effects vary by the size of the household landholding. Column 1 estimates the impact for workers in households without agricultural land, column 2 for workers in households with greater than 0 acres but but weakly less than two acres, column 3 for workers in households with more than 2 acres but weakly less than 4 acres, and column 4 for households with 4 or more acres of land. Panel B examines whether treatment effects vary by age and gender. The first two columns present results for workers aged 21-40 and 41-60. The next two present the same results for males workers only, while the final two present the same results for female workers. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 7: Impact of new road on firms

Panel A. Employment in firms in village

	Share of workers	Log employment	Log employment (tradables)	Log employment (non-tradables)
New road	0.013	0.223	0.394*	0.230*
	0.022	0.149	0.224	0.127
Control group mean	0.11	2.94	1.54	2.46
Control group mean (level)		32.1	16.5	15.6
N	10608	10608	10608	10608
R2	0.30	0.30	0.34	0.21

Panel B. Number of firms in village

	Log firms	Log firms (tradables)	Log firms (non-tradables)
New road	0.240*	0.310*	0.226**
	0.138	0.188	0.114
Control group mean	2.42	1.30	1.95
Control group mean (level)	17.3	8.9	8.4
N	10608	10608	10608
R2	0.32	0.35	0.22

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on firms. The dependent variable in each regression comes from the 2013 Economic Census. Panel A presents results on employment in nonfarm economic establishments, while Panel B presents results on the number of these establishments. In column 1, the dependent variable is the share of total workers in the village working in such establishments. Columns 2-4 present results on log employment (A) and number of firms (B) for firms in all industries, firms in tradable industries and firms in non-tradable industries. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates

Table 8: Impact of new road on agricultural inputs

	Mechanized farm equipment	Irrigation equipment	Land ownership
New road	0.001	-0.000	0.003
	(0.012)	(0.028)	(0.036)
Control group mean	0.041	0.141	0.571
N	11473	11474	11474
R2	0.26	0.43	0.39

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on measures of agricultural intensification. Column 1 estimates the impact on the share of households owning mechanized farm equipment, column 2 on the share of households owning irrigation equipment and column 3 on the share of households owning land. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 9: Impact of new road, by agricultural productivity

	Occupation in Agriculture		Log in-village nonfarm employment	
	Low Productivity	High Productivity	Low Productivity	High Productivity
New road	-0.138*	-0.059	0.450**	0.059
	0.073	0.052	0.213	0.208
Control group mean	0.516	0.432	2.851	3.042
N	5981	5493	5597	5011
R2	0.25	0.32	0.39	0.24

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the heterogeneous effects of new road construction by agricultural productivity (see text for details). The dependent variable in the first two columns is share of workers in agriculture, while the dependent variable for the next two columns is log employment in nonfarm firms located in the village. The sample for the "Low Productivity" columns includes all villages with below-median agricultural productivity, and the sample for the "High Productivity" columns includes all villages with above-median agricultural productivity. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

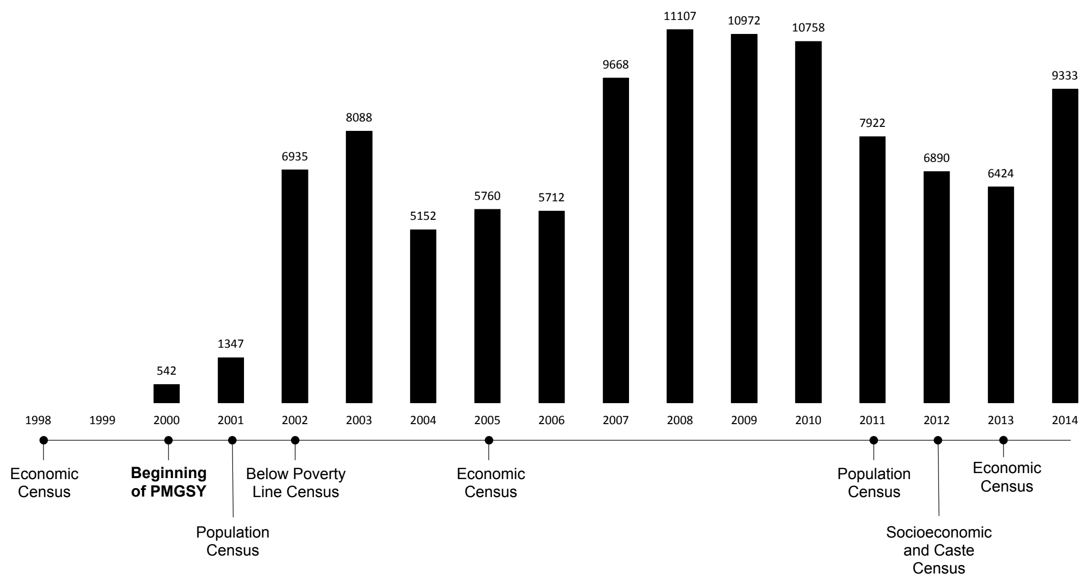
Table 10: Impact of new road on household assets

	Asset Index	Solid House	Refrigerator	Vehicle	Phone
New road	0.327	0.042	0.009	0.003	0.041
	(0.295)	(0.029)	(0.013)	(0.024)	(0.042)
Control group mean	-1.777	0.220	0.036	0.140	0.444
N	11474	11474	11474	11474	11474
R2	0.58	0.67	0.27	0.38	0.48

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

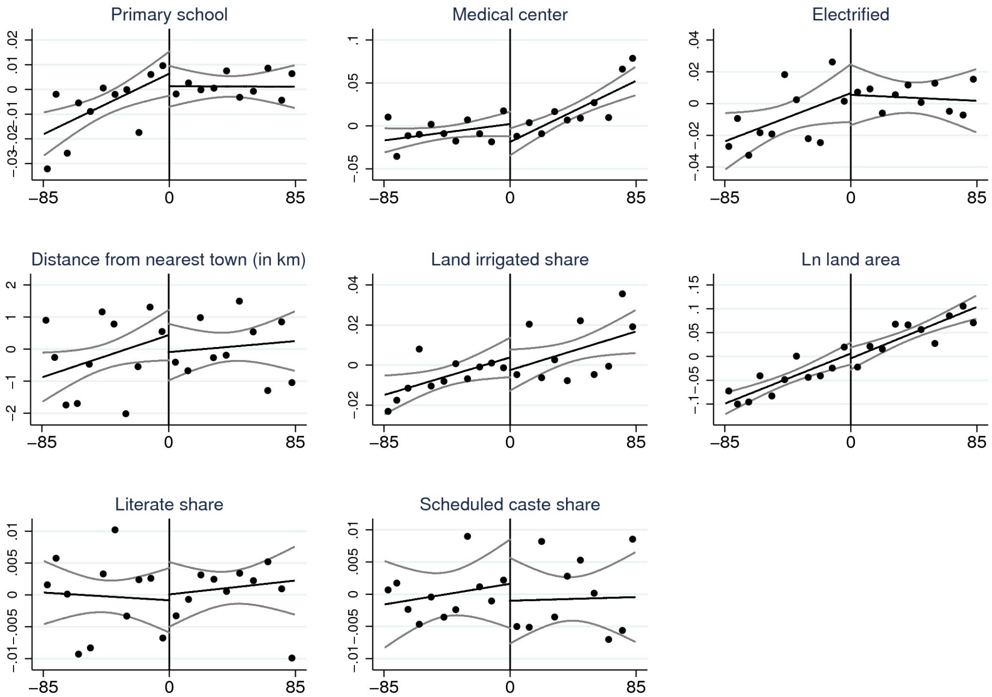
Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on household assets. Column 1 takes as a dependent variable an index of the four household assets listed in the 2012 Socioeconomic and Caste Census: a house of solid material (having both solid walls and roof), a refrigerator, a motorized vehicle, and a phone. All four components of the index are the share of households in the village having that asset, and have been standardized to have a mean of zero and standard deviation of one. Columns 2-5 present estimates for each of these outcomes individually, with the dependent variable defined as the share of households having the asset in question. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Figure 1: Timeline of Data Sources, with Count of Villages Treated



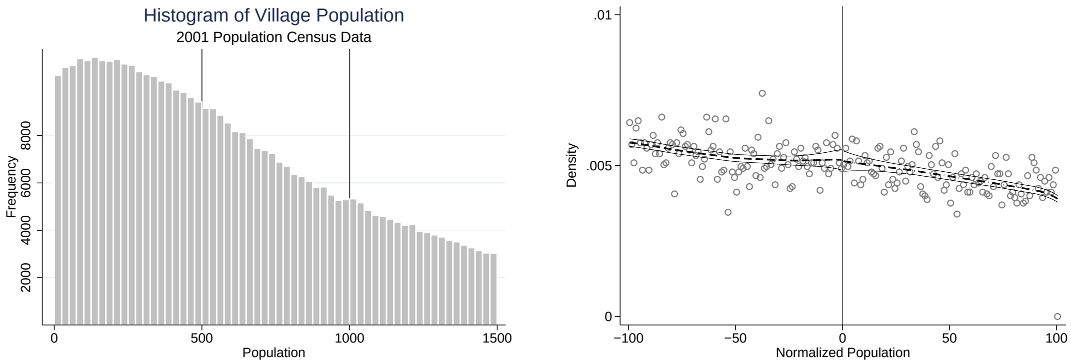
Notes: The figure shows when the population and poverty censuses of India used as primary data sources in this paper were conducted. Note that while the Socioeconomic and Caste Census (SECC) was intended to be conducted exclusively in 2011, and it is often referred to with this year, it was conducted primarily in 2012. The bar graph above represents the number of villages receiving PMGSY roads in each year in our full village-level dataset. Exact counts are also listed.

Figure 2: Balance of baseline village characteristics



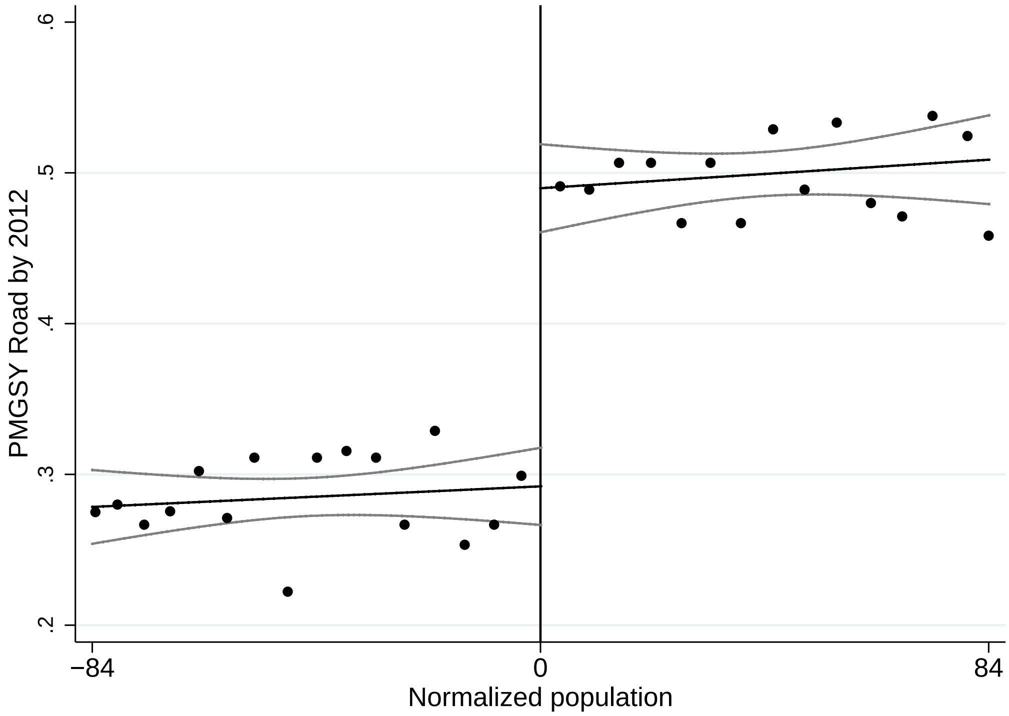
Notes: The figure plots residualized baseline village characteristics (after controlling for all variables in the main specification other than population) over normalized village population in the 2001 Population Census. Points to the right of zero are above treatment thresholds, while points to the left of zero are below treatment thresholds. Each point represents approximately forty observations. As in the main specification, a linear fit is generated separately for each side of 0, with 95% confidence intervals displayed. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the threshold (see text for details).

Figure 3: Distribution of running variable



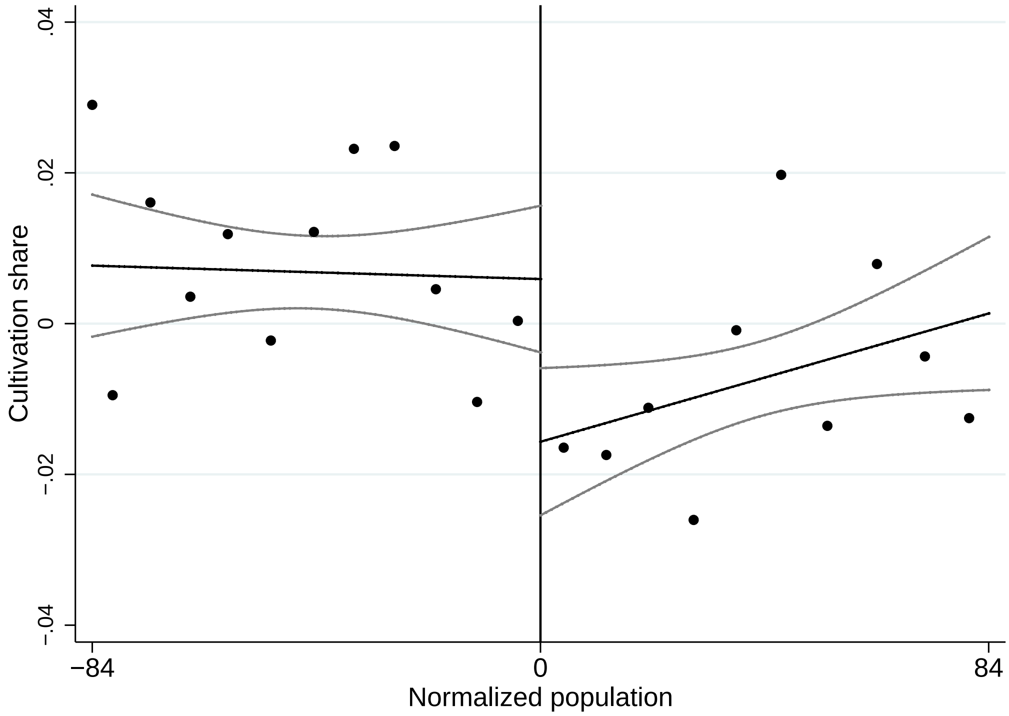
Notes: The figure shows the distribution of village population around the population thresholds. The left panel is a histogram of village population as recorded in the 2001 Population Census. The vertical lines show the program eligibility cutoffs used in this paper, at 500 and 1000. The right panel uses the normalized village population (reported population minus the threshold, either 500 or 1000). It plots a non-parametric regression to each half of the distribution following McCrary (2008), testing for a discontinuity at zero. The point estimate for the discontinuity is -0.01, with a standard error of 0.05.

Figure 4: First stage: effect of road prioritization on probability of PMGSY road by 2012



Notes: The figure plots the probability of getting a PMGSY road by 2012 over village population in the 2001 Population Census. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the population thresholds (see text for details). Populations are normalized by subtracting the cutoff.

Figure 5: Reduced form: effect of road prioritization on share of workers in agriculture



Notes: The figure plots the residualized share of households reporting cultivation as the primary source of income (after controlling for all variables in the main specification other than population) over normalized village population in the 2001 Population Census. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the population thresholds (see text for details). Populations are normalized by subtracting the cutoff.

A For Online Publication - Appendix: Additional figures and tables

Table A1: RD estimate of PMGSY road on population growth

	Pop growth (level)	Pop growth (annualized percent)
New road	-3.374	-0.000
	(14.434)	(0.002)
Control group mean	1.019	1.019
N	10972	10972
R2	0.91	0.27

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of a new road by 2010 on population growth between the 2001 and 2011 Population Census. In column 1 the outcome is total village population in 2011, while in column 2 the outcome is annualized population growth between the 2001 and 2011 Population Censuses. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 250 rupees per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A2: RD estimate of PMGSY road on age distribution and gender ratios

Panel A. Age group share

	11-20	21-30	31-40	41-50	51-60
New road	-0.003	-0.005	0.004	-0.002	0.002
	0.005	0.005	0.004	0.004	0.003
Control group mean	0.239	0.188	0.148	0.114	0.072
N	11474	11474	11474	11474	11474
R2	0.22	0.19	0.26	0.38	0.40

Panel B. Male share by age group

	11-20	21-30	31-40	41-50	51-60
New road	-0.010	0.002	0.004	-0.004	0.018
	0.009	0.008	0.008	0.010	0.013
Control group mean	0.521	0.517	0.509	0.520	0.515
N	11474	11474	11474	11474	11474
R2	0.13	0.20	0.10	0.08	0.06

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of PMGSY treatment on village demographics. Panel A presents results on the share of the village population in ten-year age bins. Panel B presents results on the share of the population in each age bin that is male. All dependent variables are generated from the SECC microdata. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A3: RD estimate of PMGSY road on distribution of landholdings

	Landless	0-2 Acres	2-4 Acres	4+ Acres
New road	-0.009	-0.018	-0.007	0.034
	(0.029)	(0.027)	(0.013)	(0.019)*
Control group mean	0.433	0.287	0.120	0.160
N	11440	11440	11440	11440
R2	0.39	0.41	0.22	0.47

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on the share of village households with landholdings in a given range. The first column reports the estimate effect on the share of households reporting no agricultural land, followed by three columns for households owning agricultural land. The sample consists of villages that did not have a paved road at baseline, with baseline population within the optimal bandwidth (84) of the threshold (see text for details). The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A4: Impact of road on unemployment

	Unemployed	Unclassifiable
New road	0.014	-0.010
	(0.024)	(0.010)
Control group mean	0.430	0.018
N	11474	11474
R2	0.30	0.17

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

Notes: This table presents regression discontinuity estimates from Equation 3 of the effect of new road construction on the occupational choice. In the first column, the dependent variable is the share of working age adults (18-60) who do not work outside of the house (household work, student, unemployed, etc), while in the second column the dependent variable is the share of working age adults whose occupation does not make clear whether or not they work. For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The sample consists of villages that did not have a paved road at baseline (see text for details). The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A5: Effect of road prioritization on treatment and share of workers in agriculture, for primary and placebo sample

Panel A. Outcome: Road treatment (first stage)

	Main Sample	Placebo Sample
Road Priority	0.212***	-0.013
	0.017	0.021
Outcome Mean	0.25	0.28
N	11474	6454
R2	0.30	0.37

Panel B. Outcome: Share of workers in agriculture (reduced form)

	Main Sample	Placebo Sample
Road Priority	-0.021**	0.001
	0.009	0.012
Outcome Mean	0.48	0.45
N	11474	6411
R2	0.30	0.45

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

Notes: This table presents a comparison of estimates from Equation 2 of the effect of PMGSY prioritization on a village's probability of treatment and reduced form estimates of the effect of PMGSY prioritization on the share of households reporting cultivation as their primary source of income for the main sample of states that adhered to the implementation cutoffs and a placebo sample of states that did not follow the cutoffs. The first column presents estimates for the sample of states who followed the cutoff rules, while the second column presents estimates for the sample that did not follow the cutoff. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the threshold (see text for details). For each regression, the outcome mean for the control group (villages with population below the threshold) is also shown. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 250 rupees per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table A6: Reduced form estimate of new road on sanitation

	Open Defecation	Latrine in Premises	Pit Latrine - with slab	Pit Latrine - without slab
Road priority	0.005	-0.004	0.007	-0.005
	(0.014)	(0.013)	(0.006)	(0.004)
N	1775	1775	1775	1775
r2	0.26	0.27	0.11	0.09

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

Notes: The Total Sanitation Campaign (TSC) is stated to have "aimed to transition rural households from open defecation to use of on-site pit latrines" (Spears, 2015). The program began construction of latrines in 2001. The outcomes considered here are 2011 measures of (in order) percentages of households who report: open defecation; the existence of a latrine within premises; an in-house pit latrine with slab or ventilated improved pit; and an in-house pit latrine without slab/open pit. The sample has been restricted to villages with population within the optimal bandwidth (84) of 1000, the cutoff used by the TSC. The sample of states here come from our main PMGSY specification. The specification includes baseline village-level controls for amenities (primary school, medical center, electrification, distance to nearest town), log total acres under cultivation, share of agricultural land irrigated, share of population literate, share of population belonging to a scheduled caste, share of households owning land, share of households with subsistence agriculture as the main source of income, and share of households earning more than 4 USD per month, as well as district-threshold fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Figure A1: Sample page from SECC

SECC ड्राफ्ट सूची - ग्रामीण

राज्य :RAJASTHAN	ज़िला :Ajmer	तहसील :Ajmer	शहर/ग्राम :Ajaysar	वार्ड कोड नंबर (केवल शहर के लिए) :0000	गणन ब्लॉक -उप खंड :0158_0

घरेलू संख्या : 0003	**घर के प्रकार** : साधारण	**ग्राम पंचायत** :-AJAYSAR	**आदिम जनजाति वर्ग से है** : नहीं	**वैधानिक रुप से छुडाया गया बंधुवा मजदूर** : नहीं	**हाथ से मैला साफ़ करने वाले** : नहीं

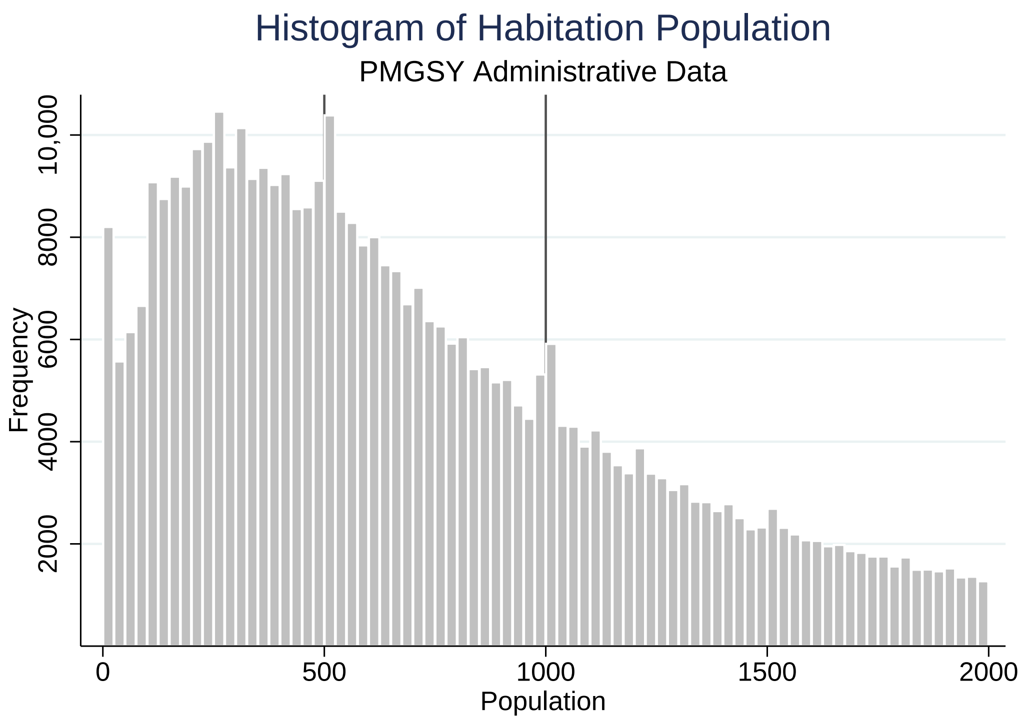
संख्या	नाम	मुखिया से संबंध	लिंग जन्मतिथि	पिता का नाम माता का नाम	वैवाहिक स्थिति#	व्यवसाय/ गतिविधि	अनु. जाति / जनजाति / अन्य	विकलांगता	शिक्षा
001	[redacted]	मुखिया	पुरुष 1953	[redacted]	2	मजदूर	अन्य	कोई निःशक्तता नहीं	निरक्षर
002	[redacted]	पत्नी	स्त्री 1955	[redacted]	2	मजदूर	अन्य	कोई निःशक्तता नहीं	निरक्षर
003	[redacted]	पुत्र	पुरुष 1989	[redacted]	1	मजदूर	अन्य	कोई निःशक्तता नहीं	पूर्व माध्यमिक

भाग 1 विवरण : आवासीय/निवासीय				भाग 3 रोजगार और आय विशेषताओं					भाग 4 : विवरण सम्पत्तियां			भाग5 अ: भूमि स्वामित्व (एकड में)				भाग 5 ब: अन्य भूमि स्वामित्व		
मकान के दीवार की प्रमुख सामग्री #	मकान की छ्त की प्रमुख सामग्री #	मकान का मालिकाना हक की स्थिति	निवास के कमरों की संख्या	नियमित वेतन पाने वाला कोई परिवार का सदस्य	आयकर या वृति कर दाता है	स्वयं की /संचलित ऐसी संस्था जो शासन द्वारा पंजीकृत है	परिवार के सबसे अधिक कमाने वाले सदस्य का मासिक आय	परिवार की आय का मुख्य स्रोत	रेफ्रिजिरेटर	टेलीफोन / मोबाईल फोन	दो/तीन/चार पहिया या मछली पकडने की नाव पंजीकृत	स्वामित्व की भूमि (वास भूमि को छोडकर)	कुल असिंचित भूमि	2 फसलों वाली सिंचाई भूमि	अन्य सिंचित भूमि	यंत्रीकृत तीन/चार व्हीलर कृषि उपकरण	सिंचाई उपकरण(नलकूप . बोर . डीजल/मिट्टी के तेल/विद्युत पंप सेट. फव्वारा/ड्रिप सिंचाई आदि समेत)	किसान क्रेडिट कार्ड की सीमा 50000 रुपए या अधिक है।
6	6	स्वय	4	नहीं	नहीं	नहीं	10,000 या अधिक	1	हां	केवल मोबाइल	दो पहिया	हां	1.0	3.0	1.0	नहीं	हां	नहीं

Ver:4.0.5a DB: 070 09 Mar 2014 06:06:02 PM Signature____________ Page 5 of 200

Notes: This is a sample page taken from a PDF file that was scraped from secc.gov.in. Individual-level variables are name, relationship with head of household, gender, date of birth, parents' names, marital status, occupation, caste category, disability and education. Household-level variables are wall material, roof material, house ownership, dwelling room count, salaried job, payment of income tax, ownership of registered enterprise, monthly income, source of income, asset ownership (refrigerator, telephone, vehicle, mechanized farm equipment, irrigation equipment, Kisan credit card), and land ownership.

Figure A2: Histogram of habitation populations (PMGSY OMMS)



Notes: The figure shows the histogram of village population as reported in the PMGSY Online Monitoring and Management System. The vertical lines show the program eligibility cutoffs at 500 and 1000.