Rural Roads and Local Economic Development
Online Appendix

Sam Asher and Paul Novosad

October 2, 2019

For Online Publication - Appendix

A Additional figures and tables

Table A1: Correlates of NDVI and EVI proxies for agricultural production

Panel A. NDVI/EVI on village proxies of agricultural productivity

	NDVI				EVI			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Crop suitability (log)	0.017			0.017	0.017			0.017
	(0.002)			(0.002)	(0.002)			(0.002)
Irrigation (share)		0.014		0.009		0.038		0.032
		(0.002)		(0.002)		(0.003)		(0.003)
Consumption (log)			0.028	0.026			0.043	0.036
			(0.002)	(0.002)			(0.003)	(0.003)
N	137336	137336	137336	137336	137336	137336	137336	137336
R2	0.49	0.49	0.49	0.49	0.51	0.51	0.51	0.51

Panel B. NDVI/EVI on district agricultural output

	NDVI				EVI			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agricultural output	0.047	0.027	0.309	0.210	0.342	0.342	0.210	0.172
	(0.016)	(0.015)	(0.038)	(0.037)	(0.021)	(0.021)	(0.042)	(0.038)
Fixed effects	State	State-Year	District	District, Year	State	State-Year	District	District, Year
N	2124	2124	2124	2124	2124	2124	2124	2124
R2	0.40	0.56	0.74	0.78	0.41	0.50	0.85	0.89

Notes: For validation purposes, our favored log-differenced NDVI and EVI agricultural production proxies are regressed on other likely correlates of yields. Panel A presents village level estimates of these proxies regressed on log crop suitability, share of village land irrigated, and log predicted consumption per capita, all with district fixed effects. Panel B presents district-level regressions of these proxies on the value of agricultural output (log) for the years 2000-2006. See Data Appendix for details. The sample has been restricted to states from the primary specification, where states follow PMGSY population guidelines. Heteroskedasticity robust standard errors are reported below point estimates.

Table A2: Summary statistics, by paved road at baseline

	No Road	Paved Road	Total
Primary school	0.692	0.864	0.783
	(0.462)	(0.342)	(0.412)
Medical center	0.183	0.434	0.316
	(0.387)	(0.496)	(0.465)
Electrified	0.249	0.549	0.405
	(0.432)	(0.498)	(0.491)
Crop land irrigated share	0.344	0.456	0.404
	(0.360)	(0.382)	(0.376)
Literate share	0.431	0.499	0.466
	(0.186)	(0.153)	(0.173)
Scheduled caste share	0.157	0.185	0.171
	(0.213)	(0.193)	(0.203)
Distance from nearest town	28.3	20.0	23.9
(in km)	(29.4)	(20.7)	(25.5)
Population	1513.2	1930.5	1730.8
	(30628.4)	(36167.6)	(33631.6)
Number of villages	282864	308263	591127

Notes: This table presents means and standard deviations of baseline variables and outcomes for all villages in India. 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 A3: Sectoral distribution of non-agricultural manual laborers

	Share of non-agricultural manual laborers in sector
Construction	0.60
Transport	0.07
Retail	0.05
Domestic work	0.05
Building materials	0.04
Other	0.17

Notes: This table shows the share of non-agricultural manual laborers in the five largest industries. The sample is the full rural population in the 68th round of the National Sample Survey (2011-12).

Table A4: Impact of new road on distribution of landholdings

	Landless	0-2 Acres	2-4 Acres	4+ Acres
New road	-0.009	-0.012	-0.007	0.028
	(0.036)	(0.033)	(0.016)	(0.024)
Control group mean	0.434	0.287	0.120	0.159
N	11394	11394	11394	11394
R2	0.39	0.41	0.23	0.47

Notes: This table presents regression discontinuity estimates from the main estimating equation 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. 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 and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A5: Impact of new road on agricultural labor share by land, age, and gender

Panel A. Impact by household landholding

	Landless	0-2 Acres	2-4 Acres	4+ Acres
New road	-0.117	-0.100	-0.075	-0.063
	(0.047)	(0.052)	(0.054)	(0.053)
Control group mean	0.352	0.514	0.590	0.653
N	11101	10698	10380	9945
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	-0.085	-0.093	-0.085	-0.094	-0.020	-0.044
	(0.045)	(0.045)	(0.045)	(0.044)	(0.056)	(0.061)
Control group mean	0.430	0.578	0.450	0.612	0.268	0.330
N	11421	11379	11410	11369	10781	10184
R2	0.28	0.29	0.28	0.29	0.21	0.24

Notes: This table presents regression discontinuity estimates from the main estimating equation 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 treatment 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. 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 and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A6: Consumption prediction first stage

	Coefficient	(SE)	p-value
Owns land	9657	(1239)	0.000
Two-wheeled vehicle	34253	(2874)	0.000
Four-wheeled vehicle	85686	(14868)	0.000
Landline phone	24639	(8154)	0.003
Mobile phone	23997	(995)	0.000
Both landline and mobile	31479	(6895)	0.000
HH income 5000 - 10000 INR	10076	(1878)	0.000
HH income 10000+ INR	38933	(4779)	0.000
Refrigerator	29477	(2868)	0.000
Number of rooms in home	3429	(599)	0.000
Grass wall	12808	(3551)	0.000
Mud wall	13372	(3269)	0.000
Plastic wall	19748	(6754)	0.003
Wood wall	9217	(3745)	0.014
Brick wall	23030	(3451)	0.000
GI wall	14184	(4505)	0.002
Stone wall	17065	(4492)	0.000
Concrete wall	22316	(3515)	0.000
Grass roof	-2920	(1770)	0.099
Tile roof	-6508	(1772)	0.000
Slate roof	2316	(3018)	0.443
Plastic roof	6474	(8259)	0.433
GI roof	-3359	(1889)	0.075
Brick roof	-9605	(2387)	0.000
Stone roof	11637	(5121)	0.023
Concrete roof	1432	(2519)	0.570
Owns home	-1334	(5550)	0.810
Kisan credit card	12441	(4584)	0.007
Constant	24538	(6572)	0.000
N = 25279			
R2 = 0.359			

Notes: This table presents estimates from the regression of total household consumption on all economic well-being measures that are used to predict consumption. The sample is all rural households in the IHDS-II, with observations weighted according to sampling weights. No other controls are used.

Table A7: Impact of new road on all predictors of consumption

	Coefficient	(SE)	p-value	N	R2
Owns land	0.006	(0.036)	0.87	11432	0.39
Two-wheeled vehicle	-0.003	(0.021)	0.89	11432	0.35
Four-wheeled vehicle	0.001	(0.007)	0.85	11432	0.22
Landline phone	-0.003	(0.004)	0.41	11432	0.08
Mobile phone	0.045	(0.041)	0.26	11432	0.47
Both landline and mobile	-0.009	(0.005)	0.09	11432	0.06
HH income from 5000 - 10000	-0.007	(0.024)	0.76	11432	0.19
HH income over 10000	0.006	(0.015)	0.68	11432	0.20
Refrigerator	0.005	(0.013)	0.70	11432	0.26
Mean number of rooms in home	0.063	(0.086)	0.46	11432	0.36
Grass wall	0.040	(0.028)	0.16	11432	0.25
Mud wall	-0.054	(0.052)	0.30	11432	0.40
Plastic wall	-0.002	(0.005)	0.63	11432	0.07
Wood wall	0.000	(0.012)	0.98	11432	0.12
Brick wall	0.004	(0.035)	0.91	11432	0.41
GI wall	0.001	(0.004)	0.76	11432	0.05
Stone wall	0.003	(0.030)	0.93	11432	0.14
Concrete wall	-0.005	(0.011)	0.69	11432	0.09
Grass roof	-0.003	(0.041)	0.95	11432	0.43
Tile roof	0.013	(0.045)	0.78	11432	0.60
Slate roof	0.016	(0.024)	0.52	11432	0.28
Plastic roof	-0.024	(0.010)	0.02	11432	0.18
GI roof	0.001	(0.021)	0.97	11432	0.51
Brick roof	-0.001	(0.008)	0.93	11432	0.28
Stone roof	0.015	(0.025)	0.56	11432	0.50
Concrete roof	-0.004	(0.018)	0.81	11432	0.43
Owns home	0.007	(0.008)	0.36	11432	0.11
Kisan credit card	-0.007	(0.017)	0.65	11432	0.35

Notes: This table presents regression discontinuity estimates from the main estimating equation of the effect of new road construction on village shares of all dummy variables used in the consumption prediction exercise (except for number of rooms, which is the village mean). The specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates for all estimates except for consumption and poverty, which report bootstrapped standard errors as described in the data appendix.

Table A8: Impact of new road on log predicted consumption, by education and occupation

Panel A. Consumption by education level

	No education	Primary or below	Middle school+
New road	-0.017	0.013	0.007
	(0.039)	(0.042)	(0.045)
Control group mean	9.39	9.54	9.75
N	11306	11340	11272
R2	0.27	0.31	0.33

Panel B. Consumption by occupation

	Agriculture	Non-ag manual labor	Other
New road	-0.055	-0.002	0.030
	(0.081)	(0.086)	(0.040)
Control group mean	9.40	9.62	9.59
N	8534	8583	11350
R2	0.26	0.40	0.39

Notes: This table presents regression discontinuity estimates from the main estimating equation of the effect of a new road on log predicted consumption. In Panel A, which divides households by education, Columns 1, 2, and 3 show results for households where the primary earner is illiterate, has primary education or below, and has middle school education or above, respectively. Panel B divides households by the occupation of the primary earner: agriculture, non-agricultural manual labor, and other. 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 and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Bootstrapped standard errors are reported below point estimates; see Data Appendix for details.

Table A9: First stage and reduced form estimates, main and placebo samples

Panel A. Main sample first stage and reduced form effects

	First stage	Reduced form				
	Road by 2012	Transport	Occupation (ag share)	Firms	Ag production	Consumption
Road priority	0.215	0.088	-0.073	0.060	0.018	0.007
	(0.017)	(0.040)	(0.034)	(0.035)	(0.027)	(0.030)
Control group mean	0.25	0.00	0.00	-0.00	-0.00	0.00
N	11432	11432	11432	10678	11432	11432
R2	0.30	0.20	0.30	0.31	0.54	0.50

Panel B. Placebo sample first stage and reduced form effects

	First stage	Reduced form				
	Road by 2012	Transport	Occupation (ag share)	Firms	Ag production	Consumption
Road priority	-0.002	-0.002	-0.016	0.010	-0.047	-0.013
	(0.017)	(0.060)	(0.039)	(0.040)	(0.032)	(0.035)
Control group mean	0.26	0.44	-0.22	0.23	-0.26	0.33
N	9142	9138	9081	8457	9142	9142
R2	0.35	0.29	0.41	0.49	0.51	0.47

Notes: This table presents a comparison of estimates of the effect of PMGSY prioritization on a village's probability of treatment (first stage) and reduced form estimates of the effect of PMGSY prioritization on indices of the five major families of outcomes, for both the main sample (Panel A) and a placebo sample of villages close to the thresholds that were not followed (Panel B). 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 and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A10: Impact of new road on indices of major outcomes, by kernel and bandwidth

	Triangular			Rectangular		
	60	80	100	60	80	100
Transport	0.404	0.411	0.401	0.419	0.430	0.307
	(0.208)	(0.188)	(0.172)	(0.205)	(0.182)	(0.154)
	[0.05]	[0.03]	[0.02]	[0.04]	[0.02]	[0.05]
Ag occupation	-0.290	-0.337	-0.332	-0.343	-0.362	-0.260
	(0.181)	(0.162)	(0.148)	(0.176)	(0.157)	(0.133)
	[0.11]	[0.04]	[0.02]	[0.05]	[0.02]	[0.05]
Firms	0.394	0.281	0.235	0.275	0.159	0.172
	(0.177)	(0.158)	(0.144)	(0.172)	(0.153)	(0.131)
	[0.03]	[0.07]	[0.10]	[0.11]	[0.30]	[0.19]
Ag production	0.145	0.093	0.071	0.102	0.080	0.050
	(0.139)	(0.125)	(0.114)	(0.137)	(0.121)	(0.104)
	[0.30]	[0.46]	[0.54]	[0.46]	[0.51]	[0.63]
Consumption	0.112	0.063	0.035	0.098	0.030	-0.023
	(0.154)	(0.138)	(0.126)	(0.149)	(0.133)	(0.112)
	[0.47]	[0.65]	[0.78]	[0.51]	[0.82]	[0.84]
N	[8339]	[11099]	[13871]	[8339]	[11099]	[13871]

Notes: This table presents regression discontinuity estimates from the main estimating equation of the effect of a new road on indices of the major outcomes in each of the five families of outcomes: transportation, occupation, firms, agriculture and welfare. We show robustness to three different bandwidth choices (60, 80, 100) and two different kernel weighting choices (rectangular and triangular). See Data Appendix for details of index construction. The specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Coefficients are presented for each regression with standard errors in parentheses and p-values in brackets.

Table A11: Impact of new road on population growth, age distribution and gender ratios

Panel A. Population growth (2001-2011)

	Log	Level
New road	-0.024	-9.662
	(0.029)	(20.275)
Control group mean	6.43	653.06
N	11432	11432
R2	0.79	0.83

Panel B. Age group share

	11-20	21-30	31-40	41-50	51-60
New road	-0.004	-0.003	0.002	-0.002	0.002
	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)
Control group mean	0.24	0.19	0.15	0.11	0.07
N	11432	11432	11432	11432	11432
R2	0.22	0.19	0.26	0.38	0.40

Panel C. Male share by age group

	11-20	21-30	31-40	41-50	51-60
New road	-0.010	0.003	0.004	-0.006	0.017
	(0.009)	(0.008)	(0.008)	(0.010)	(0.013)
Control group mean	0.52	0.52	0.51	0.52	0.51
N	11432	11432	11432	11432	11432
R2	0.13	0.19	0.10	0.07	0.05

Notes: This table presents regression discontinuity estimates from the main estimating equation of the effect of PMGSY treatment on village demographics. Panel A presents results on 2011 village population, both in log and level. Panel B presents results on the share of the village population in ten-year age bins. Panel C presents results on the share of the population in each age bin that is male. Dependent variables in Panels B and C are generated from the SECC microdata. 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 and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A12: Impact of new road on unemployment

	Unemployed	Unclassifiable
New road	0.010	-0.009
	(0.024)	(0.010)
Control group mean	0.430	0.018
N	11432	11432
R2	0.30	0.17

Notes: This table presents regression discontinuity estimates from the main estimating equation 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 specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A13: Impact of new road on sanitation

	Open Defecation	Latrine (on premises)	Pit Latrine (with slab)	Pit Latrine (without slab)
New road	0.006	-0.003	0.019	-0.010
	(0.038)	(0.036)	(0.017)	(0.012)
Control group mean	0.891	0.105	0.019	0.011
N	1776	1776	1776	1776
R2	0.25	0.27	0.09	0.08

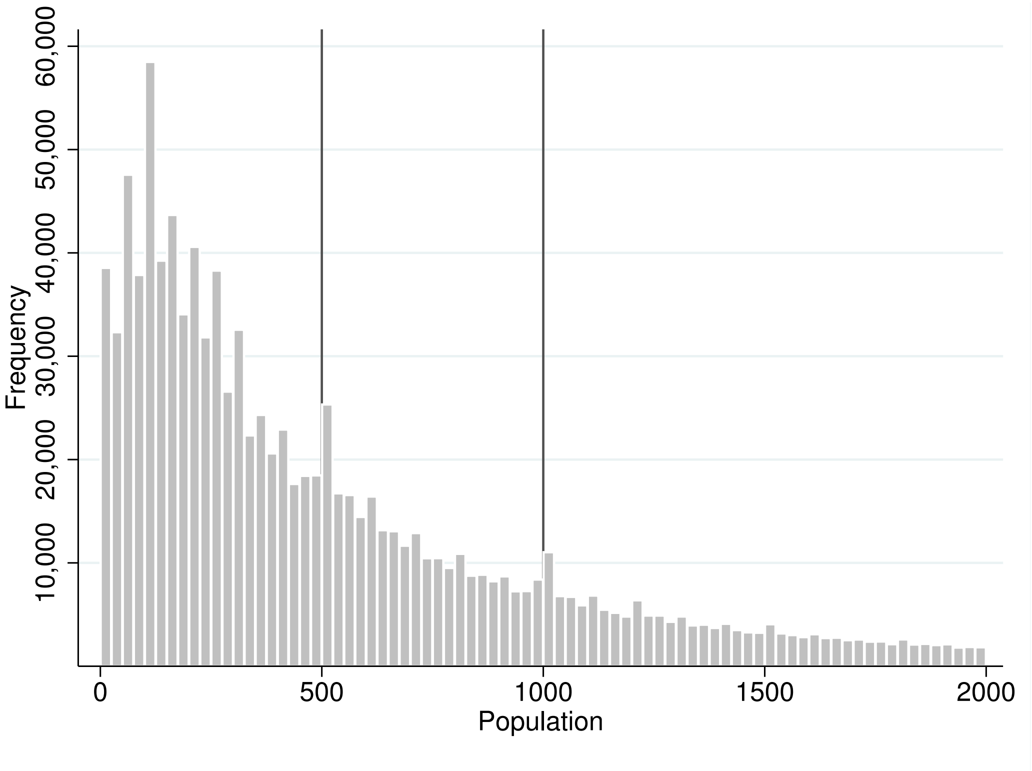
Notes: The Total Sanitation Campaign (TSC) is stated to have "aimed to transition rural households from open defecation to use of on-site pit latrines" (Spears, 2015). The program began construction of latrines in 2001. The outcomes considered here are 2011 Population Census measures of (in order) percentages 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 1,000, the threshold used by the TSC. The sample of states here come from our main PMGSY specification. The specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Table A14: Spillovers: impact of new road on nearby villages

	Transportation	Ag occupation	Firms	Ag production	Consumption	Unemployment rate
New road	-0.049	-0.001	-0.165	0.036	0.060	-0.007
	(0.135)	(0.132)	(0.141)	(0.100)	(0.114)	(0.009)
p-value	0.72	1.00	0.24	0.72	0.60	0.45
N	11403	11403	11403	11403	11403	11403
R2	0.51	0.52	0.46	0.71	0.65	0.70

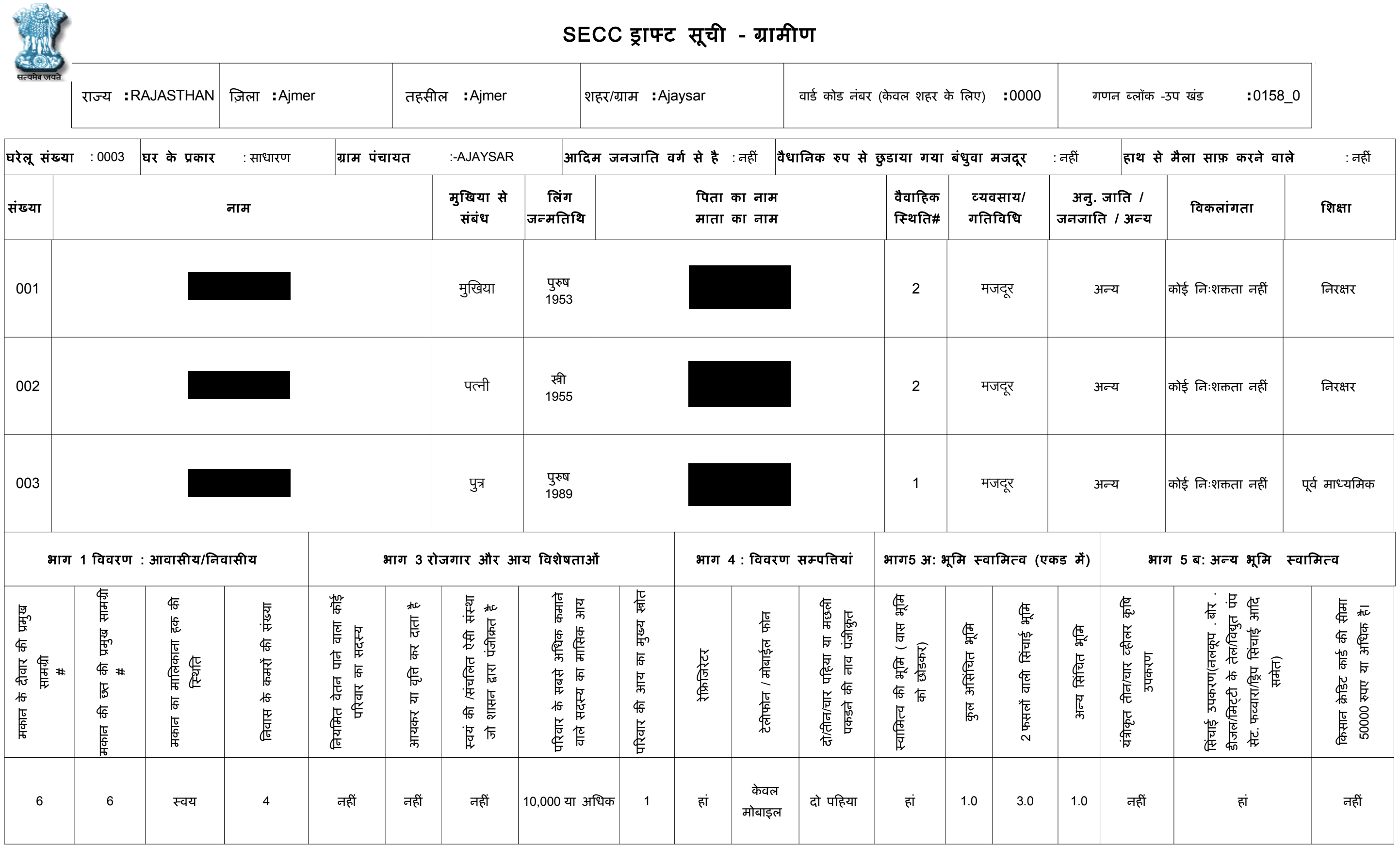
Notes: This table presents regression discontinuity estimates from the main estimating equation of the effect of a new road on outcomes in nearby villages. Dependent variables are indices of the five families of outcomes (transportation, occupation, firms, agriculture, and welfare), plus a sixth column for the unemployment rate. A catchment area for a PMGSY sample village is defined as other villages within 5 km. Outcomes are aggregated across spillover villages. Otherwise the specification is identical to the main regression specification for estimating direct effects. See Data Appendix for details of index construction. The specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects (see Section V for details). Heteroskedasticity robust standard errors are reported below point estimates.

Figure A1: Histogram of habitation populations (PMGSY OMMS)



Notes: The figure shows the histogram of the habitation populations as reported in the PMGSY Online Monitoring and Management System. The vertical lines show the program eligibility thresholds at 500 and 1,000. Due to evidence of manipulation in the PMGSY administrative data, the running variable used in the analysis is population from the 2001 Population Census.

Figure A2: 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 छत की प्रमुख सामग्री #	मकान का मालिकाना हक की स्थिति	निवास के कमरों की संख्या	नियमित वेतन पाने वाला कोई परिवार का सदस्य	आयकर या वृत्ति 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 पकड़ने 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

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.

B Data Appendix

Section IV gives an overview of the data used in this paper. This data appendix provides more detail on the data sources and construction of the main variables.

B1 Administrative Data on Road Construction

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

PMGSY data are posted online (http://omms.nic.in) 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.[1] We successfully match over 85% of habitations listed in the PMGSY to their corresponding population census villages.

B2 Socioeconomic censuses

Data on occupation, earnings and assets 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 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, which we use to construct baseline control variables.

The fourth such census, the Socioeconomic and Caste Census (SECC), was launched in 2011 but primarily conducted in 2012.[2] To increase the likelihood of collecting data on all

[1] For fuzzy matching, we used a combination of the reclink program in Stata, and a custom fuzzy matching script called masala_merge2 based on the Levenshtein algorithm but modified for the languages used in India. The fuzzy matching algorithm can be found in the included code.

[2] 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

individuals and households, it was based on the National Population Register (NPR) from the 2011 Population Census. To increase transparency, the Government of India made the SECC publicly available at http://secc.gov.in in a mix of PDF and Excel formats; currently only aggregated data is available on the website. See Figure A2 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 (NCO).[3]. Our main occupational variables (share of workers in agriculture and share of workers in non-agricultural manual labor) are based on this classification: agricultural workers are those with NCO single digit code 6 (skilled agricultural workers) or NCO 2 digit 92 (agricultural laborers), while non-agricultural manual laborers are those with NCO single digit code 9 (elementary occupations) excluding those in agriculture (code 92).

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.

B3 Economic and population censuses

The Indian Ministry of Statistics and Programme Implementation (MoSPI) conducted the 6th Economic Census in 2013. The Economic Census is a complete enumeration of all economic establishments except those engaged in crop production, defense and government administration. Establishments are any location, commercial or residential, where an economic activity is carried out. There is no minimum firm size, and both formal and informal establishments are enumerated, including people working out of their houses. The dataset, as well as earlier rounds, is available for purchase from MoSPI. We obtained the location directory for the Economic Census, and then used a series of fuzzy matching algorithms to match villages and towns by name to the population census of 2011. Employment is defined as the number of workers at the firm on the work day prior to the enumerator's visit, including casual wage laborers. We aggregate the microdata to the village level to obtain a measure of employment in village nonfarm firms. We use the sum of employment in all firms reported in the 2013 Economic Census to produce an endline measure of nonfarm employment. The Economic Census also reports the sector of the firm, which we use to test for heterogeneous effects across the five largest sectors in our sample (livestock, forestry, manufacturing, retail and education), which together account for 79% of employment in in-village nonfarm firms. For all regressions using this data, we define the outcome variable as $log(employment_{i,v} + 1)$, where employment is

as the relevant year for the SECC.

[3] All available rounds of the Indian Classification of Occupations, produced by the Ministry of Statistics and Programme Implementation since 1962, are available at http://mospi.nic.in/classification/national-industrial-classification.

the sum of employment in all firms in sector i in village v. To ensure that outliers do not drive our results, we restrict our sample in regressions using outcomes from the Economic Census to villages where total employment is less than total inhabitants in the village.

The Primary Census Abstract (PCA) and Village Directory (VD) give us village-level data in the Population Censuses of 2001 and 2011. They are available for free download on the Census website (http://censusindia.gov.in/DigitalLibrary/Tables.aspx). The 2001 data provides control variables for the main regressions and is used to establish baseline balance for the regression discontinuity, while 2011 data is used to measure endline outcomes. The PCA is the source for demographic information (such as total population) and the VD for village characteristics and amenities (such as roads, electricity, schools, regular availability of transportation, etc.).

We also test for outcomes from two new measures of agricultural inputs from the 2011 Population Census Village Directory. The first is crop choice. The census records the three major crops for each village—from this we generate an indicator variable for whether the village grows any non-subsistence crops, which we define as anything other than cereals (rice, wheat, etc) and pulses (lentils, chickpeas, etc). The second is total agricultural land, which we transform into logs.

These censuses also provide the basis for linking the various other datasets. We use a key provided by the 2011 Population Census to link data from 2011 to 2001. GIS data of village boundaries in 2011, procured from ML Infomap (a digital mapping firm) and based on official census maps, is used for the aggregation of gridded remote sensing to the village level.

We have combined multiple rounds of the economic and population censuses into a single dataset, referred to as the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG), Version 1.0. Asher et al. (2019) provides details of its construction and guidance on its use. Version 1.0 of the dataset, which contains a limited set of core variables and keys to link to three rounds of the Population Census and four rounds of the Economic Census, can be found at https://doi.org/10.7910/DVN/DPESAK.

B4 Agricultural production

As no comprehensive village-level data is collected on agricultural production in India, we use two commonly-used and closely related vegetative indices (VIs) to proxy for agricultural production in baseline and endline survey periods: the normalized difference in vegetation index (NDVI) and the enhanced vegetation index (EVI), which is very similar but uses additional information from the blue part of the electromagnetic spectrum. NDVI and EVI are chlorophyll-sensitive measures of plant matter, generated at global coverage and 250 m resolution by the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Earth Observing System-Terra satellite. NDVI is built using near infrared and red bands, while EVI uses additional information from the blue band to reduce atmospheric interference and the influence of background vegetation (Son et al., 2014). NDVI and EVI have shown to be equivalently effective for crop classification tasks (Wardlow and

Egbert, 2010), and have also been shown to be equally successful at predicting wheat yields in Canada when combined with agroclimate data (Kouadio et al., 2014). Each image represents a 16-day composite where each pixel value is optimized considering cloud cover obstruction, image quality, and viewing geometry via the MODIS VI algorithm (Huete et al., 2002). Composite images were downloaded from the Columbia University IRI Data Library (https://iridl.ldeo.columbia.edu) for the years 2000-2014 for nine 16-day periods from late May through mid-October, covering the major (kharif) cropping season in India (Selvaraju, 2003).

For each composite image, pixels were spatially averaged to village polygons. After village aggregation within each 16-day composite, three proxies for agricultural production were calculated for each year's growing season: the difference between early-season VI (the mean of the first three 16-day composites) and the max VI value observed at the village level (Labus et al., 2002; Rasmussen, 1997), mean VI (Mkhabela et al., 2005), and cumulative NDVI (Rojas, 2007) (the sum of NDVI from each of the nine composites during the growing season).[4] All VI measures are then log transformed for the regressions to allow for an interpretable effect. We prefer the differenced measure because it effectively controls for non-crop vegetation (such as forest cover) by measuring the change in vegetation from the planting period (when land is fallow) to the moment of peak vegetation.

We use additional likely correlates of agricultural production to validate the use of these growing-season VI measures as a proxy for agricultural output at the village level (Table A1). Cross-sectional regressions with state fixed effects were run using log endline year (2011-2013 average) growing season change in NDVI (as described above) as the dependent variable. At the village level, these correlates are: cereal crop potential production measure (low input usage, log) from the FAO Global Agro-Ecological Zones (GAEZ) aggregated to the village level (raw data available at http://www.fao.org/nr/gaez/en/; share of village agricultural land area under any type of irrigation; and per capita annual predicted consumption (described below). Additionally, panel NDVI data was regressed at the district level on agricultural output from the Planning Commission's series of district domestic product data, across a consistent sample of districts (raw data available at http://planningcommission.nic.in/plans/stateplan/index.php?state=ssphdbody.htm). While these remotely sensed measures of agricultural production do not capture other determinants of agricultural earnings such as quality or price changes, their strong correlation with both agricultural productivity measures and real measures of production supports using them to estimate impacts of roads on village agricultural production.

B5 Consumption

We combine data from 2012 SECC and the concurrent IHDS-II (Indian Human Development Survey, 2011-12, available at https://ihds.umd.edu) to predict village-level consumption measures following the methodology in Elbers et al. (2003). To do this, using IHDS data, we regress total household consumption on dummy variables that are equiv-

[4] To reduce noise, we define our endline measure as the average of the measures for 2011, 2012 and 2013, and our baseline measures as the average of the measure for 2000, 2001 and 2002.

alent to all asset and earnings information contained in the SECC.[5] The results of this regression are given in Table A6. We then use the coefficients to predict household-level consumption in the SECC microdata. This is used to generate consumption per capita at the individual level, which is in turn used to produce village level statistics for mean predicted consumption per capita, per capita predicted consumption at different village percentiles, and share of the population below the poverty line.[6] For the purpose of regressions, consumption variables are winsorized at the 1st and 99th percentiles, and log transformed. As outlined in Elbers et al. (2003), in order to get correct standard errors and p-values, we perform a double bootstrap, first in the IHDS regressions to generate 1,000 different asset coefficient vectors, and then over villages in our main sample.[7]

This method is supported by a large literature on predicting consumption and proxying welfare using asset and related data. Early work showed that in the United States, up to 78% of the variation in total consumption could be predicted by a linear regression on food consumption, housing expenditures and valuation, vehicle ownership, size of the family, and age (Skinner, 1987). Hentschel et al. (2000) show that this method yields unbiased estimates of poverty and performs well except when sample sizes are very small. McKenzie (2005) evaluates the ability of this method to generate accurate measures of inequality and poverty, finding that it better predicts non-durable consumption than other methods considered; he also validates this measure by finding that predicted consumption and directed measured consumption generate highly similar conclusions on the relationship between inequality and schooling in Mexico. Both McKenzie (2005) and Young (2012) make the further point that assets have the advantage of likely capturing real, permanent income better than consumption measured at any moment in time. Predicted consumption using this method has also been widely used (most notably by the World Bank) to generate poverty estimates using census data for areas not covered by detailed (and expensive) household consumption surveys (Bedi et al., eds, 2007). While we are undoubtedly missing some of the variation in consumption not explained by these assets and income variables, our large sample sizes (median village has 152 households in the SECC) and wide range of assets covered (from housing materials to vehicles to mobile phones) give us confidence that our measure of predicted consumption is sufficiently precise to pick up major changes in consumption.

For an alternative way of aggregating information across assets, we create an index at the village level by taking the primary component of the indicator variables described above in the SECC microdata, normalized to have a mean of 0 and standard deviation of 1 within our sample.

[5] These variables are roof material (grass, tile, slate, plastic, GI metal, brick, stone, and concrete), wall material (grass, mud, plastic, wood, brick, GI sheets, stone, and concrete), number of rooms, phone ownership (landline only, mobile only, and both landline and mobile), house ownership (owned), vehicle ownership (two wheeler and four wheeler), land ownership, kisan credit card, refrigerator, and highest individual income in household (between 5,000 and 10,000 rupees and more than 10,000 rupees).

[6] We use the official rural poverty line of INR 27/day from the Tendulkar Committee Report (Government of India, 2014).

[7] To speed up the computation of the bootstrapped estimations, we modify GNU Parallel code (Tange, 2011).

The only earnings variable available at the village level comes from the SECC. It records monthly earnings of the highest earning member of the household, censored into three bins: 0 to 4,999 rupees, 5,000 to 9,999 rupees and 10,000+ rupees. As 85% of households report being in the lowest bin, we define our earnings variable to be the share of households in the top two bins (with the highest earner earning 5,000 rupees or more).

We generate another consumption proxy using lights at night, as measured by satellites. Night lights are a proxy for consumption that have the advantage of high resolution and objective measurement over a 20+ year period (Henderson et al., 2011). We match gridded data to village polygons, sum over all pixels in the village and then take the log of the value plus 1 in order to not drop observations that take the value 0. To increase precision, we define our dependent variable as the log of the mean value from 2011, 2012 and 2013 (plus 1), and include a control for log mean baseline light (plus 1) in 2000-2002.

B6 Spillovers

Spillover effects of PMGSY road construction on nearby villages are assessed using 2001 Population Census GIS data purchased from ML InfoMap. Catchment areas with radii of 5 km were constructed by measuring distances from the centroids of villages in the sample to the centroids of all other villages. Outcomes were then aggregated across all villages within these catchment areas, constructed in the same manner as for the non-spillover regressions. On average, there are 15 villages per 5 km catchment area. 55 percent of non-sample villages within a catchment appear in more than one catchment at 5km. These villages are double counted, but should not bias the estimates due to the exogeneity of road construction in our regression discontinuity sample.

B7 Family-wise indices

In order to address concerns of multiple hypothesis testing, we follow Anderson (2008) in generating five indices for our main families of outcomes: transportation, labor market, firms, agriculture and assets/consumption. Each of these is generated by demeaning its component outcomes and converting to effect sizes through dividing by control group standard deviation; demeaned values are then combined by weighting according to the inverse of the covariance matrix. The transportation index is comprised of five indicator variables for availability of motorized transit as measured in the 2011 Population Census: public buses, private buses, vans, taxis and auto-rickshaws. The labor market index is comprised of the share of workers in agriculture and the opposite of the share of workers in manual labor (so that their covariance is positive), both coming from the SECC. The firms index is comprised of log of employment plus 1 in all nonfarm firms in the 2013 Economic Census; it does not include the other firm outcomes as they are simply disaggregations of total employment by sector. The agriculture index is comprised of our favored measure of agricultural yields (differenced NDVI, described above) and each of the measures of agricultural inputs: share of households owning mechanized farm equipment, share of households owning irrigation equipment, share of households owning land, log total cultivated acres and an indicator for non-cereal/pulse (subsistence) crops among the primary three crops in the village (coming from a combination of the Population Census and SECC). Finally, the asset/consumption index is comprised of log predicted consumption per capita, the primary component asset index, log night light luminosity and the share of households with the primary earner making more than 5,000 INR per month.

References

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

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: a Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad, "The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG)," 2019. Working paper.

Bedi, Tara, Aline Coudouel, and Kenneth Simler, eds, *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, Washington, D.C.: The World Bank, 2007.

Elbers, Chris, Jean Lanjouw, and Peter Lanjouw, "Micro-level Estimation of Poverty and Inequality," *Econometrica*, 2003, *71* (1), 355–364.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "A Bright Idea for Mesuring Economic Growth," *American Economic Review*, 2011, *101* (3), 194–199.

Hentschel, Jesko, Jean Olson Lanjouw, Peter Lanjouw, and Javier Poggi, "Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador," *The World Bank Economic Review*, 2000, *14* (1), 147–165.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices," *Remote Sensing of Environment*, 2002, *83* (1-2), 195–213.

Kouadio, Louis, Nathaniel K. Newlands, Andrew Davidson, Yinsuo Zhang, and Aston Chipanshi, "Assessing the Performance of MODIS NDVI and EVI for Seasonal Crop Yield Forecasting at the Ecodistrict Scale," *Remote Sensing*, 2014, *6* (10), 10193–10214.

Labus, M. P., G. A. Nielsen, R. L. Lawrence, R. Engel, and D. S. Long, "Wheat Yield Estimates Using Multi-temporal NDVI Satellite Imagery," *International Journal of Remote Sensing*, 2002, *23* (20), 4169–4180.

McKenzie, David J., "Measuring Inequality with Asset Indicators," *Journal of Population Economics*, 2005, *18* (2), 229–260.

Mkhabela, Manasah S., Milton S. Mkhabela, and Nkosazana N. Mashinini, "Early Maize Yield Forecasting in the Four Agro-ecological Regions of Swaziland Using NDVI Data Derived From NOAA's-AVHRR," *Agricultural and Forest Meteorology*, 2005, *129* (1-2), 1–9.

Rasmussen, M. S., "Operational Yield Forecast Using AVHRR NDVI Data: Reduction of Environmental and Inter-annual Variability," *International Journal of Remote Sensing*, 1997, *18* (5), 1059–1077.

Rojas, O., "Operational Maize Yield Model Development and Validation Based on Remote Sensing and Agro-meteorological Data in Kenya," *International Journal of Remote Sensing*, sep 2007, *28* (17), 3775–3793.

Selvaraju, R., "Impact of El Niño-southern Oscillation on Indian Foodgrain Production," *International Journal of Climatology*, 2003, *23* (2), 187–206.

Skinner, Jonathan, "A Superior Measure of Consumption from the Panel Study of Income Dynamics," *Economics Letters*, 1987, *23* (2), 213–216.

Son, N. T., C. F. Chen, C. R. Chen, V. Q. Minh, and N. H. Trung, "A Comparative Analysis of Multitemporal MODIS EVI and NDVI Data for Large-Scale Rice Yield

Estimation," *Agricultural and Forest Meteorology*, 2014, *197*, 52–64.

Tange, Ole, "GNU Parallel: The Command-line Power Tool," *The USENIX Magazine*, 2011, *36* (1), 42–47.

Wardlow, Brian D. and Stephen L. Egbert, "A Comparison of MODIS 250-m EVI and NDVI Data For Crop Mapping: A Case Study for Southwest Kansas," *International Journal of Remote Sensing*, 2010, *31* (3), 805–830.

Young, Alwyn, "The African Growth Miracle," *Journal of Political Economy*, 2012, *120* (4), 696–739.