Satellite Data for the Social Sciences: Measuring Rural Electrification with Nighttime Lights

Eugenie Dugoua,[1] Ryan Kennedy,[2*] Johannes Urpelainen[3]

[1]School of International and Public Affairs, Columbia University, 14th Floor New York, NY 10027, USA

Department of Political Science, University of Houston, 447 Philip G. Hoffman Hall, Houston, TX 77204, USA

[3]Department of Energy, Resources and Environment, Johns Hopkins University SAIS, 1619 Massachusetts Ave. NW, Office 428, Washington, DC 20036, USA

[*]Corresponding author, `rkennedy@uh.edu`

Remote sensing data has the potential to revolutionize social science. One of the most prominent examples of this is the Nighttime Lights dataset, which provides digital measures of nighttime luminosity from 1992 to 2013. This study evaluates the Nighttime Lights data against detailed rural electrification data from the 2011 Census of India. The results suggest that many nighttime luminosity measures derived from satellite data are surprisingly accurate for measuring rural electrification, even at the village level and using simple statistical tools. We also demonstrate that this accuracy can be substantially improved by using of better GIS maps, basic geoprocessing tools, and particular aggregations of nighttime luminosity. Nighttime luminosity performs worse in measuring financial inclusion or proxies of poverty, however, and detects rural electrification less accurately when the supply of power is intermittent. These results offer guidelines for when and how remote sensing data can be used when administrative data is absent or unreliable.

Summary: The promise and limitations of satellite measures of nighttime lights are demonstrated and methods of improvement are illustrated.

Introduction

Remote sensing data – information collected from satellites or high-flying aircraft – has the potential to revolutionize the social sciences. Once only the purview of state military branches, the information from these sources is increasingly being made available to the public, either through official releases from government agencies or through private sources such as Google Maps. The amount of remote sensing data available to the public is likely to increase dramatically in the next decade, as the cost of satellite technology decreases (Ma et al., 2015).

One of the most prominent examples of remote sensing data in the social sciences is the Nighttime Lights dataset provided by NOAA's National Geophysical Data Center (NOAA, n.d.). The satellites responsible for this data were originally tasked by the U.S. Defense Meteorological Satellite Program (DMSP) to estimate cloud cover by using the level of light from the Earth's surface. Only later was it realized that, by putting together a composite of cloud-free images, one could estimate a digital number (DN) of nighttime lights around the world. Economists and political scientists, in turn, realized that nighttime lights could be used to estimate electricity use and economic activity, without some of the issues of missingness or unreliability that plague official data in developing countries (Baskaran, Min, and Uppal, 2015; Chen and Nordhaus, 2011; Min et al., 2013).

To illustrate the remarkable variety in the use of nighttime lights in the social sciences, Table 2 offers a list of several recent studies that have used the nighttime lights data, the concept they attempt to proxy with the data, and their method of processing the data. Scholars have usually used the nighttime lights to measure economic output (Doll, Muller, and Morley, 2006; Chen and Nordhaus, 2011; Addison and Stewart, 2015; Henderson, Storeygard, and Weil, 2012), the level of electrification (Min et al., 2013; Min and Gaba, 2014; Baskaran, Min, and Uppal, 2015), the population of an area (Addison and Stewart, 2015), and urban extent (Small, Pozzi, and Elvidge, 2005). Changes in nighttime lights has also been used to estimate the impact of conflict (Li et al., 2015) and natural disasters (Cole et al., 2017). The methods they have used to construct these measures have varied widely. Scholars have used the maximum digital number DN, the DN sum within an area, the number of non-zero DN pixels, the average DN within an area, and others. They have also used a variety of GIS data to construct their estimates, including estimating the DN at a particular point (i.e., center of an area or a digital recording of the brightest area as observed from the ground (Min et al., 2013)), using a shapefile that provides the outline of the area of interest, or using a combination of both.

[Table 1 about here.]

Yet, the use of remote sensing data for analysis has sometimes outpaced validation. While remote sensing data is most useful where data is sparse, such as is often the case in sub-national and rural areas, the lack of validation raises questions about the quality of measurement in studies using nighttime lights as a proxy for other variables. Existing studies have used village surveys in Vietnam (Min and Gaba, 2014) as well as Senegal and Mali (Min et al., 2013), but these validation exercises have been relatively small in scale (1,331 villages total across studies), selected largely through convenience sampling (Min et al., 2013), and their results mixed with regard to household electrification. A recent machine learning approach (Jean et al., 2016) shows that poverty measures can be improved by using a combination of daytime satellite imagery and the nightlights data, but they use the nighttime lights data primarily to identify features from their more detailed daytime satellite imagery.

This study examines the validity of nighttime lights as a measure for rural electrification with village-level data from India. We construct several different measures of luminosity, varying both the GIS data we use and the aggregation method, and test them against detailed information on the number of electrified households in six hundred thousand villages from the

2011 Census of India. India provides a unique testing ground, both because of the detail and accuracy of administrative data and the variety of regional conditions, allowing us to test the effect of regional development variation on nighttime lights accuracy.

The results suggest that nighttime lights is a surprisingly accurate measure of village household electrification, and that relatively simple linear models function quite well. However, our results also show that a large amount of variance in accuracy depends on the aggregation technique used, the underlying GIS data, regional development, and the concept being proxied by nighttime lights. The results show that remote sensing data are a promising resource when administrative records are absent or unreliable, yet they also underscore the limitations of such data for analyzing economic and social phenomena and offer practical guidelines for good measurement practice.

Data and Methods

Our ground truth variables come from the 2011 Census of India – the latest census conducted in the country (Government of India, 2011). The new census offers detailed information about electricity access for every village in India. Besides being the lowest level for which household electrification data is available, the village is an appropriate unit of analysis because it is the primary administrative unit in national rural electrification schemes. We relate the (logarithmized) number of electrified households to the nighttime lights of the village area. See SI Section S1 for data and methods.

We construct nighttime lights proxies in several ways. First, we utilized the India Lights Project's API to download as much village-level data as possible from their system using all the 2001 census codes (Min et al., 2016). The API data is organized around the month of the observation. In some years there are two such observations, but in others there are three.

We took the mean of all their available measures (maximum, minimum, mean, and median of recorded DN) across the months to produce yearly data.

Second, we downloaded night light data from NOAA for the year 2011 (DMSP OLS V4). For the core of our analysis, we use the ‘stable lights’ dataset: this includes locations with persistent lighting only. Ephemeral events, such as fires were previously discarded and background noise was identified and replaced with values of zero. We replicate our tests with the raw lights data in section S14 and find that, although this data has far fewer observations with a DN of 0, it is less accurate than its filtered counterpart. We calculate a village-level value of night lights using several different GIS file types:

- A shapefile of 2011 villages produced by ML InfoMap, an ISO 9001-2015 certified company for GIS with specific expertise in digital maps of India. The upside to this GIS file is that we have the actual shape of the village with which to calculate zonal statistics. The downside is that there are six states or union territories (out of 36) for which ML InfoMap did not provide a shapefile: Andaman and Nicobar, Arunachal, Lakshadweep, Meghalaya, Mizoram, and Nagaland (Table S1).

- A pointfile of 2011 villages produced by ML InfoMap. We calculated the village centroids from the previously discussed map and combined it with centroid point data of the states missing shapefiles. This increased the number of cases, but estimates from a point calculate statistics at about 1km around the center point of a village (size of a pixel in the nightlights data). Following others in the literature (Min et al., 2013), we also calculated the bilinear interpolation values of the point data, which takes into account neighboring pixel values.

- As an attempt to cut the balance between the point file and the shapefile, we also produced datasets were DN values were calculated within a 2-km, 3-km and 5-km circular buffer around the village centroid.

For each of these GIS files, we calculated several commonly used values: the mean, sum, and maximum of the DN. Because we found the shapefile-derived measures to have significant benefit, we focus on the sum of DN within the village boundaries. We analyze the data using a variety of tools, including both simple linear models and more complex non-linear models.

Results

Our analysis proceeds in three steps. First, we look at the raw correlation coefficients (Pearson's *r*) between the nighttime lights data and our ground truth data. Second, we use multivariate regression analysis to explore how much the nighttime lights data contributes to correctly modeling our ground truth outcomes. Finally, we check for nonlinearity in the relationship using a variety of methods. SI Section S3 offers summary statistics and SI Section S4 shows maps illustrating variation in our data across India.

Figure 1 shows the correlation between nighttime lights and the number of electrified households in a village. Although the degree of correlation varies across different measures of nighttime lights, the best measures perform well. Specifically, the correlation between the logarithmized sum of DN from shape file data and the logarithm of the total number of electrified households (variables 'log ShSum' and 'elec nbr log') is 0.63. Taking the same variables without the logarithmization gives an almost identical correlation (0.62).

[Figure 1 about here.]

On the other hand, measures from the India Lights Project show lower correlations with household electrification. This weak association might exist because only the mean and the maximum measures are available, and not the sum of the DN over a polygon. In SI Section

S5, we compare the correlations of sum, logarithmized sum, mean, minimum, maximum, and median values across variables constructed using the shape, 2k, 3k, and 5k methodologies. The logarithmized sum consistently yields the highest correlations, indicating the importance of good geospatial information about the extent and shape of the villages.

We also compare the correlations for the maximum and the mean from the India Lights Project against the maximum and the mean of variables constructed using the shape, 2k, 3k, and 5k methodologies. The India Lights Project measures obtain the lowest correlations. Surprisingly, the DN of one pixel at the centroid of the village has a higher correlation with household electrification than any of the measures obtained from the India Lights Project. In general, though, summing over a buffer area (e.g., 2 km) substantially improves the correlation over using point estimation.

We investigate heterogeneity across Indian states in Figure 2. As the scatter plot on the left and the map on the right show, the village-level correlation between nighttime lights and the number of electrified households varies across states. In states with high levels of rural electrification and adequate electricity supply, such as Punjab in the north and Tamil Nadu in the south, these correlations are high. In states with low levels of rural electrification and intermittent supply, such as Bihar and Uttar Pradesh, the correlations are lower. Thus, a certain level of electricity access and power sector development are necessary conditions for accurate prediction with nighttime lights. For detailed analysis by state, see SI Section S6.

[Figure 2 about here.]

Figure 3 displays a hexabin plot of nighttime lights against the number of electrified households. The colors of the hexabins indicate the number of observations within that bin. As the plot shows, there is considerable variation in the number of electrified households in villages with no luminosity at all, along the y-axis. As nighttime lights increases along the x-axis, however, the variation in the numbers of electrified households from the 2011 Census of India decreases. The strong positive correlation between nighttime lights and the number of electrified households is also clear.

[Figure 3 about here.]

Figure 4 confirms this result. Without nighttime lights, there is considerable dispersion in the number of electrified households, but the dispersion decreases as the night lights grow brighter. Additional analysis in SI Section S9 demonstrates, however, that this relationship does not hold equally for different aggregation methods of the DN data.

[Figure 4 about here.]

In Table 1, we investigate the relationship between household electrification and nighttime lights using linear regression. As the first four models show, there is a strong and robust association between the two measures. The coefficient decreases as we add fixed effects for smaller administrative units, however, suggesting that nighttime lights is less suited for capturing variation in rural electrification within small geographic areas. In fact, even the inclusion of state fixed effects reduces the coefficient from 0.701 to 0.548, showing that cross-state differences explain much of the variation in household electrification. Models 5-8 show that predictive accuracy can be improved somewhat with the inclusion of a separate indicator for no luminosity at all. SI Section S8 shows that nonlinear regressions improve predictive accuracy only slightly.

The reason why nighttime lights performs poorly at predicting rural electrification at low levels is related to intermittent electricity supply. SI Section S10 examines the relationship between nighttime lights and rural electrification as a function of hours of electricity supply, and we find that the correlation between night lights and electrification is smaller for villages with fewer hours per day of electricity. In SI Section S12, we replicate these results using geocoded household survey data from 714 Indian villages (Aklin et al., 2016b,a), and note that the number of street lights in the village does not predict nighttime lights. This result might stem from the low number of street lights in a typical Indian village and their erratic use. In a separate analysis (SI Section S13), we also show that luminosity spillovers from cities bias predictions for nearby villages, but the bias is quite small.

Figure 5 explores the suitability of nighttime lights for other socio-economic variables: the percentage of households with a TV, percentage of households without assets (a proxy for extreme poverty), and the percentage of households with a bank account (a proxy for financial inclusion). The dependent variable is the logarithmized nighttime lights. The regressions also control for the number of electrified households, and in some models for the distance to the closest city. The variable most closely related to nighttime lights is TV ownership, which is unsurprising because televisions require electricity which is always used in the first place for lighting. For example, a 10 percentage point increase in TV ownership increases the DN sum by at most 25 percent, an effect comparable to that of increasing the village population by 75%. Overall, however, the relationship between these variables and nighttime lights is weak after controlling for household electrification. In the household survey data analysis from 714 Indian villages (Aklin et al., 2016b,a) (SI Section S12), controlling for the number of electrified households also significantly weakens the association between average monthly expenditure and night lights. From this analysis, nighttime lights appears more suitable for measuring rural electrification, while its utility in measuring more complex socio-economic outcomes is mixed.

[Figure 5 about here.]

Using Nighttime Lights in the Social Sciences

Using data from the 2011 Census of India, we have shown that total nighttime lights over village area is a reliable measure of the progress of household electrification. The relationship is especially robust in Indian states with adequate electricity supply, whereas the remote sensing measures are less reliable in states with constrained power supplies. The measures are also not very reliable for detecting non-electrified villages, and nighttime lights appears to be less reliable for measuring other outcomes, such as extreme asset poverty or financial inclusion. The predictive power of nighttime lights also decreases as village comparisons are restricted to comparisons within smaller geographic areas, such as inside state or district boundaries.

These results offer to researchers and policymakers guidelines for the proper use of remote sensing data. In the case of nighttime lights, these measures offer reliable village-level estimates within India under a wide range of conditions, but they are much more noisy as measures of local household living standards. This result shows that using nighttime lights to measure different socioeconomic outcomes, from GDP per capita to urban growth, requires careful validation of the proxy in advance. While nighttime lights are correlated with different measures of per capita income in India, these correlations are much weaker than the simple correlation with rural electrification. To the extent rural electrification is driven by factors other than household wealth, nighttime lights might not be a good measure for economic outcomes. In the case of India, for example, the government's heavy investment in electrifying poor rural communities makes nighttime lights a problematic proxy for economic development in general.

By increasing the scale of our validation in a large, heterogeneous country like India, we have been able to expand dramatically on smaller-scale validation efforts (Min and Gaba, 2014; Min et al., 2013), and demonstrated that that the validity of nighttime lights varies widely across the Indian states. Based on this validation exercise, we propose the following rule of thumb:

nighttime lights is an adequate proxy for measuring rural electrification and local electricity consumption, but it should be used as a proxy for other social and economic outcomes with caution. For example, our findings support applications of nighttime lights to measure progress in household electrification (Min, 2015; Kroth, Larcinese, and Wehner, 2016) – a key issue in human development – but raise questions about the detection of local economic outcomes (Hodler and Raschky, 2014).

	(1) Number of electrified households	(2) Number of electrified households	(3) Number of electrified households	(4) Number of electrified households	(5) Number of electrified households	(6) Number of electrified households	(7) Number of electrified households	(8) Number of electrified households
Night lights	0.701***	0.548***	0.485***	0.458***	0.877***	0.685***	0.680***	0.676***
	(0.065)	(0.037)	(0.033)	(0.027)	(0.064)	(0.039)	(0.029)	(0.026)
Night lights absence					0.727**	0.532**	0.751***	0.850***
					(0.291)	(0.222)	(0.134)	(0.095)
Fixed effects: state	No	Yes	No	No	No	Yes	No	No
Fixed effects: district	No	No	Yes	No	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes	No	No	No	Yes
R^2	0.381	0.216	0.159	0.122	0.387	0.220	0.169	0.135
Number of observations	516769	516769	516769	516769	516769	516769	516769	516769

Values are regression coefficients with standard errors in parentheses
Dependent variable: log(number of households with electricity)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Linear regression of the log(number of electrified households) on night lights DN, with standard errors clustered by state. The night lights measure is the log of the 2011 sum of DN within a village shape. Night lights absence is a dummy variable that is 1 when there is no DN recorded within the village in 2011.

References and Notes

Addison, Douglas M., and Benjamin Stewart. 2015. “Nighttime Lights Revisited: The Use of Nighttime Lights Data as a Proxy for Economic Variables.” World Bank, Policy Research Working Paper 7496.

Aklin, Michaël, Chao-yo Cheng, Karthik Ganesan, Abhishek Jain, Johannes Urpelainen, and Council on Energy, Environment and Water. 2016a. “Access to Clean Cooking Energy and Electricity: Survey of States in India (ACCESS).” Harvard Dataverse, V1. http://dx.doi.org/10.7910/DVN/0NV9LF.

Aklin, Michaël, Chao-yo Cheng, Johannes Urpelainen, Karthik Ganesan, and Abhishek Jain. 2016b. “Factors Affecting Household Satisfaction with Electricity Supply in Rural India.” *Nature Energy* 1: 16170.

Baskaran, Thushyanthan, Brian Min, and Yogesh Uppal. 2015. “Election Cycles and Electricity Provision: Evidence from a Quasi-experiment with Indian Special Elections.” *Journal of Public Economics* 126: 64–73.

Burlig, Fiona, and Louis Preonas. 2016. “Out of the Darkness and Into the Light? Development Effects of Rural Electrification in India.” Energy Institute at Haas, Working Paper 268.

Chen, Xi, and William D. Nordhaus. 2011. “Using Luminosity Data as a Proxy for Economic Statistics.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (21): 8589–8596.

Cole, Tony A, David W Wanik, Andrew L Molthan, Miguel O Román, and Robert E Griffin. 2017. “Synergistic Use of Nighttime Satellite Data, Electric Utility Infrastructure, and Ambient Population to Improve Power Outage Detections in Urban Areas." *Remote Sensing* 9 (3): 286.

Doll, Christopher N.H., Jan-Peter Muller, and Jeremy G. Morley. 2006. "Mapping Regional Economic Activity from Night-Time Light Satellite Imagery." *Ecological Economics* 57 (1): 75–92.

Filho, C.R. De Souza, J. Zullo, Jr, and C. Elvidge. 2004. "Brazil's 2001 Energy Crisis Monitored from Space." *International Journal of Remote Sensing* 25 (12): 2475–2482.

Government of India. 2011. "2011 Census Report, Houselisting and Housing Census Data Highlights." `http://www.censusindia.gov.in/2011census/hlo/hlo_highlights.html`.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Hodler, Roland, and Paul A. Raschky. 2014. "Regional Favoritism." *Quarterly Journal of Economics* 129 (2): 995–1033.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–794.

Kroth, Verena, Valentino Larcinese, and Joachim Wehner. 2016. "A Better Life for All? Democratization and Electrification in Post-Apartheid South Africa." *Journal of Politics* 78 (3): 774–791.

Li, Xi, Rui Zhang, Chengquan Huang, and Deren Li. 2015. "Detecting 2014 Northern Iraq

Insurgency using Night-Time Light Imagery." *International Journal of Remote Sensing* 36 (13): 3446–3458.

Ma, Yan, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. 2015. "Remote Sensing Big Data Computing: Challenges and Opportunities." *Future Generation Computer Systems* 51: 47–60.

Min, Brian. 2015. *Power and the Vote: Electricity and Politics in the Developing World*. New York: Cambridge University press.

Min, Brian, and Kwawu Mensan Gaba. 2014. "Tracking Electrification in Vietnam Using Nighttime Lights." *Remote Sensing* 6 (10): 9511–9529.

Min, Brian, Kwawu Mensan Gaba, Chris Elvidge, and Anand Thakker. 2016. "nightlights.io: Twenty Years of India Lights." Online Data Resource, `http://nightlights.io/`.

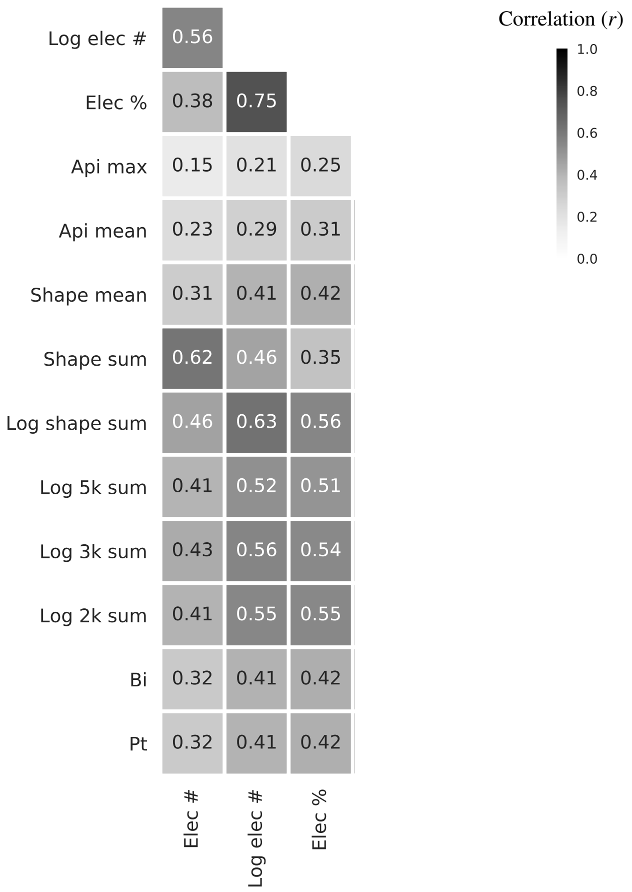
Min, Brian, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou. 2013. "Detection of Rural Electrification in Africa Using DMSP-OLS Night Lights Imagery." *International Journal of Remote Sensing* 34 (22): 8118–8141.

NOAA. n.d. "Version 4 DMSP-OLS Nighttime Lights Time Series." `https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`. Accessed: 2017-10-18.

Small, Christopher, Francesca Pozzi, and Christopher D. Elvidge. 2005. "Spatial Analysis of Global Urban Extent from DMSP-OLS Night Lights." *Remote Sensing of Environment* 96 (3): 277–291.

Acknowledgments

A replication package with all data and code will be uploaded on a public repository upon publication (Harvard Dataverse, `http://dataverse.org/`). We thank Brian Min and Semee Yoon for excellent comments on previous drafts.



Elec #	Number of households that are electrified in the village, respectively
Log elec #	Log of the number of households that are electrified in the village, respectively
Elec %	Percentage of households that are electrified in the village
Bi	Luminosity of the pixel at the longitude and latitude of the village centroid using linear interpolation of the surrounding pixels
Pt	Luminosity of the pixel at the longitude and latitude of the village centroid
Log Nk sum	Log of the sum of the luminosity of all the pixels within a N-km circle, respectively, centered at the village centroid
Log shape Sum	Log of the sum of the luminosity of the pixels within the shape boundaries of the village
Shape sum	Sum of the luminosity of the pixels within the shape boundaries of the village
Shape mean	Mean luminosity of pixels inside the shape boundaries of the village
Api max	Max of the luminosity provided by the India Lights Project API.
Api mean	Mean of the luminosity provided by the India Lights Project API.

Figure 1: Pearson correlation coefficients (*r*) between electrification and luminosity variables.

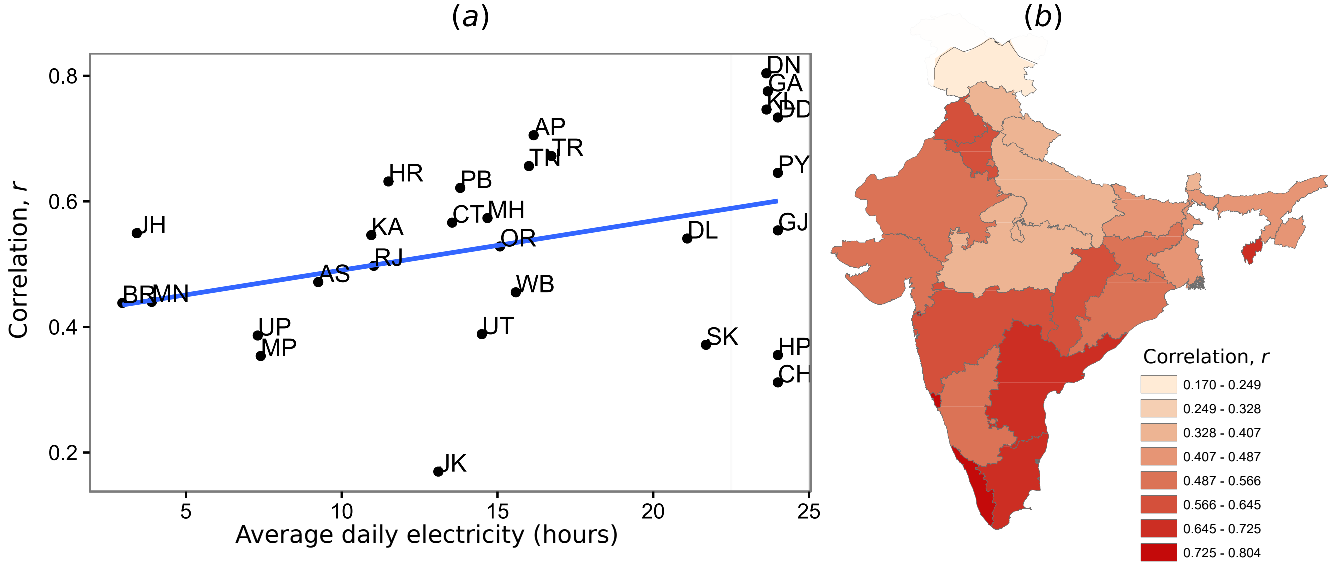


Figure 2: Pearson correlation coefficients (r) between the logarithmized number of electrified households and the logarithmized sum of the 2011 shape file night lights measure, state by state. The scatter plot (a) shows the correlation coefficients as a function of average hours of supply (Government of India, 2011), names corresponding with abbreviations can be found in Table S2; map (b) places the correlation coefficients on a map of India.

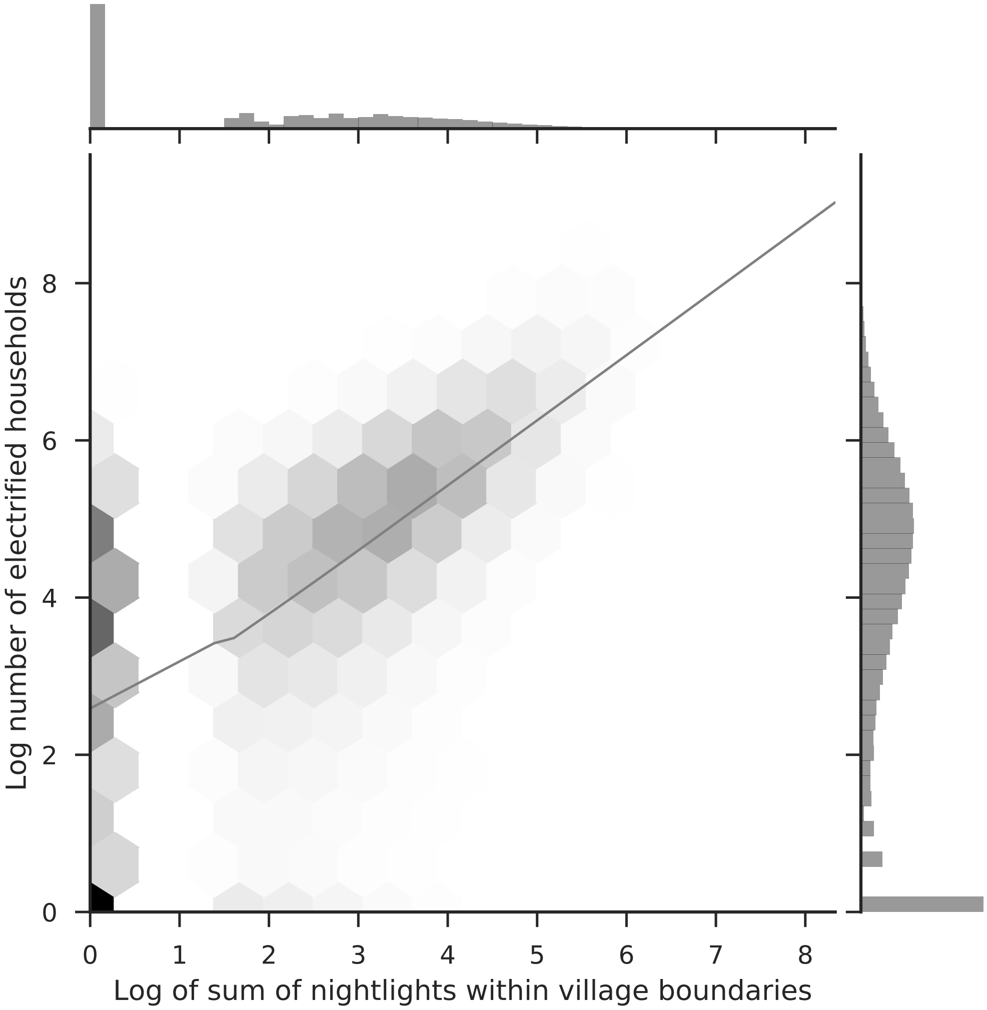


Figure 3: Hexabin plot of the log number of electrified households against the log sum of nightlights for 2011.

Note: Nightlights were summed within village boundaries as defined in the shape files. Dark colors indicate more observations in each hexabin. The two histograms illustrate the univariate distribution of the variables.

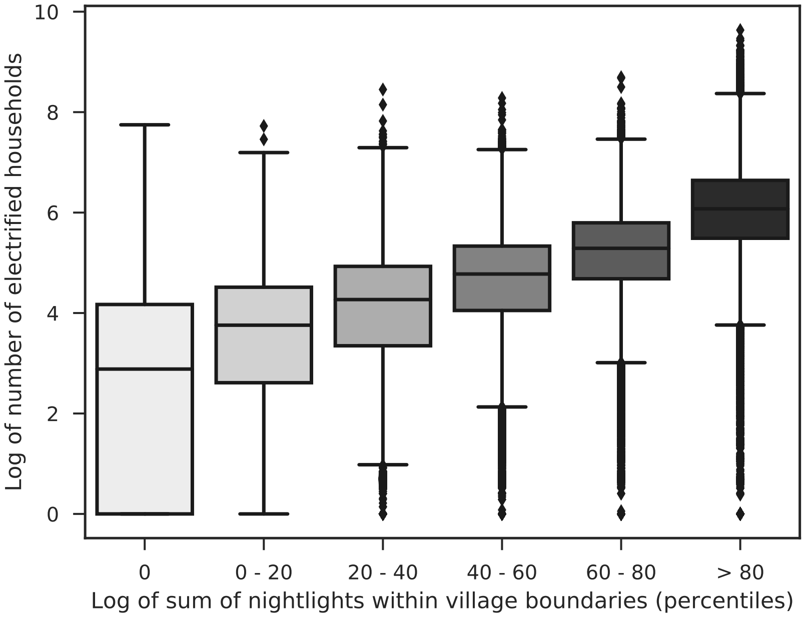


Figure 4: Boxplots of the logarithmized number of electrified households against the logarithmized sum of the 2011 shape file night lights measure, by percentile.

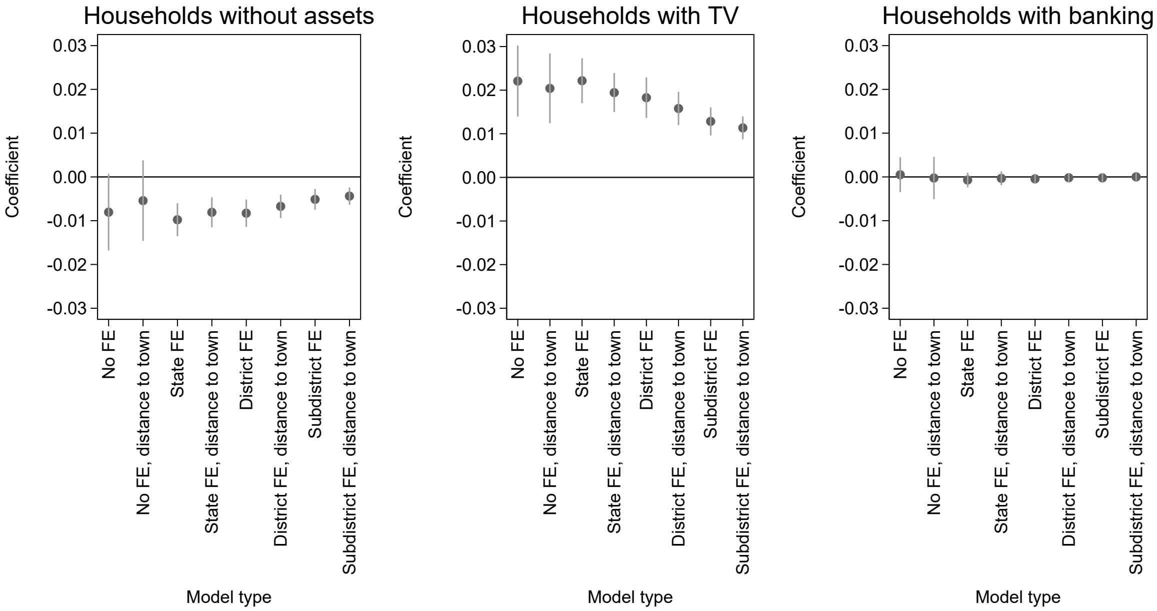


Figure 5: Coefficients and standard errors for models of logarithmized nighttime lights regressed on proxies for poverty and financial inclusion. All models control for the logarithmized numbers of electrified and non-electrified households; FE indicates the inclusion of fixed effects for state, district, or sub-district. Full tables, including control variables, can be found in Tables S10-S12.

Article	Concept	Construction of night lights measure
Chen and Nordhaus (2011)	Economic growth and GDP	Natural log of aggregated DN for all grid cells
Burlig and Preonas (2016)	Rural electrification	Maximum DN pixel
Addison and Stewart (2015)	GDP, manufacturing, electricity consumption and population	# of illuminated pixels, average DN, and sum of DN
Henderson, Storeygard, and Weil (2012)	Economic activity	% change in sum of DN
Filho, Zullo, and Elvidge (2004)	Forced energy shutdowns	Average DN
Doll, Muller, and Morley (2006)	GDP and gross regional product	Sum of DN
Hodler and Raschky (2014)	Regional economic favortism	Log of average DN in a region
Min et al. (2013)	Rural electrification	DN at estimated area of highest brightness
Min and Gaba (2014)	Rural electrification	Sum of DN
Baskaran, Min, and Uppal (2015)	Manipulation of electricity supply	Sum of DN per-capita
Li et al. (2015)	Impact of conflict	Sum of DN within city boundaries
Cole et al. (2017)	Impact of natural disasters	Average DN

Table 2: Summary of construction of night lights measures across recent articles and working papers. DN is the digital number associated with the level of luminosity.

Supplemental Material

Satellite Data for the Social Sciences: Measuring Rural Electrification with Nighttime Lights

Eugenie Dugoua,[1] Ryan Kennedy,[2] Johannes Urpelainen[3*]

[1]School of International and Public Affairs, University of Pittsburgh, Columbia University, 14th Floor New York, NY 10027, USA

Department of Political Science, University of Houston, 447 Philip G. Hoffman Hall, Houston, TX 77204, USA

[3]Department of Political Science, Columbia University, 420 W 118th Street, 712 IAB, New York, NY 10027, USA

[*]Corresponding author, ju2178@columbia.edu

Contents

S1 Materials and Methods

S1.1 Census of India 2011

We use data from the 2011 Census of India to obtain information about electrification and socio-economic characteristics at the village level.[1] Overall, the data contains 596,843 observations (i.e., inhabited villages) once urban areas are dropped. For our main variable of interest, household electricity access, the household amenities module provides the percentage of households using grid electricity as their main source of lighting. The question from the household survey is phrased as follows: "Do you use grid electricity for lighting?" Because this is census data, the percentage is based on the entire village population.

As the primary census abstract also provides information about the number of households, we are able to extract an estimate for the absolute number of households in each village that use grid electricity (share of households electrified multiplied by total households). We use the logarithm of this variable as our main electrification variable.

The census data also contain information about the total number of people living in the village, the number of people belonging to the scheduled caste and scheduled tribe groups, as well as the number of people who are literate. Information regarding the distance to the closest town (km) and land area (hectares) of the village is also available. Finally, economic outcomes at the village level can be measured with the percentage of households with a bank account, a TV, a radio, a mobile phone and other assets.

S1.2 Satellite Data for Lights

The Nighttime Lights dataset is provided by NOAA's National Geophysical Data Center.[2] The satellites responsible for this data were originally tasked by the U.S. Defense Meteorological Satellite Program (DMSP), run by the Department of Defense, to estimate cloud cover by using the level of light from the Earth's surface. Only later was it realized that, by putting together a composite of cloud-free images, one could estimate a digital number (DN) of luminosity around the world.

[1] See `http://censusindia.gov.in/` for additional information.

[2] See `http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`.

We construct nighttime lights proxies in several ways. First, we utilized the India Lights Project's API to download as much village-level data as possible from their system using all the 2001 census codes.[3] The API data is organized around the month of the observation. It usually provides for each month the maximum, minimum, mean, median and standard deviation of the visibility for each village. We constructed yearly data based on these variables by taking the maximum (minimum, mean) of the monthly maximums (minimums, means) to create a yearly maximum (minimum, mean).

Second, we downloaded the stable and cleaned version of the yearly data from NOAA and constructed our own measures using several different GIS files to do the calculations.

- A shapefile of 2011 villages produced by ML InfoMap. The upside to this GIS file is that we have the actual shape of the village with which to calculate zonal statistics. The downside is that there are a few states for which ML InfoMap did not provide a shapefile.

- A pointfile of 2011 villages produced by ML InfoMap. We calculated the village centroids from the previously discussed map and combined it with centroid point data of the states missing in the shapefiles. This increased the number of cases, but it will only allow us to calculate statistics at about 1km around the center point of a village (the size of a pixel in the nightlights data). Following the lead of some other authors, we also calculated the bilinear interpolation values of the point data, which takes into account neighboring pixel values.

- As an attempt to cut the balance between the point file and the shapefile, we also produced datasets were DN values were calculated within a 2km, 3km and 5km circular buffer around the village centerpoint.

The pointfile and the bilinear interpolation methods provide us with the luminosity value of the pixel at the centroid of each village. For the shapefile data and the various circular buffers, we calculated several commonly used values: (1) the mean DN, (2) the sum of DNs, and (3) the maximum DN.

In Section S14 below, we replicate our analysis using the raw lights data that does not correct

[3] Scraping code available from the authors.

for background noise. This analysis is a useful robustness check because the correction procedure used by NOAA is a key reason why the number of villages with zero luminosity is so high in our data.

S1.3 ACCESS Survey

The ACCESS survey [Aklin et al.(2016b), Aklin et al.(2016a)] is a statistically representative survey of 8,568 households in 714 villages across 51 districts in six states in the northern and eastern parts of India: Bihar, Jharkhand, Madhya Pradesh, Uttar Pradesh, Odisha, and West Bengal. The 45-minute survey in the local language (Hindi, Bangla, or Odiya) was conducted between November 2014 and May 2015 by enumerators from the company Morsel Research & Development. The survey modules included basic socio-economic characteristics and information about household energy use. We use the ACCESS survey to supplement our analysis of the 2011 Census of India. Besides the household survey, the dataset also includes a village module on community characteristics based on the responses of a village leader (formal or informal). Although we only have information about 714 villages, we have data on a much larger number of variables than provide by the Census. We collapse the data from the household to the village level and use the 2011 Census village codes to match the survey to our night lights data. See Sections S11-S12 for details.

S1.4 Statistical Methods

Our basic model is a linear regression with the logarithmized number of electrified households as the dependent variable. With i indicating villages, the model can be written as follows:

$$Y_i = \alpha + \beta \text{Night lights (shape, log sum, 2011)}_i + \epsilon_i, \tag{1}$$

where α is a constant and ϵ the error term. The primary explanatory variable is the night lights measure described above. Some models also include state, district, or sub-district fixed effects. We cluster standard errors by state throughout.

We consider the following variants of this model:

- In a placebo model, we replace the logarithmized number of electrified households with the logarithmized number of non-electrified households.

- In some models, we include a separate indicator for zero night lights. In this models, it is thus assumed that the data generation process for zero night lights is different from the data generation process for positive night lights. In other words, besides the intercept for the lowest value of positive night lights, there is a separate point estimate for all villages with zero night lights.

- We also estimate non-linear models that add the second, third, and fourth orders of the polynomial to allow flexible functional forms.

- In some models, we code indicators for the first to the fifth quartiles of positive night lights, with zero night lights as the base category.

- We also estimate models that use the logarithmized mean/maximum for the year 2011 from the India Lights API data instead of the logarithmized sum for the village shape files.

In another batch of models, we use the night lights (shape, log of sum, 2011) as the dependent variable so that we can use multiple explanatory variables in a meaningful and easily interpretable setup. The variables used in this approach are:

- Logarithmized number of electrified households

- Logarithmized number of non-electrified households

- Households with banking access (%)

- Households without any basic assets: radio, television, computer, telephone, bicycle, motorcycle, car

- Households with TV

- Distance to town in logarithmized km

In some models, we also use hours of electricity supply to domestic users. The variable is constructed by taking the average of summer and winter hours to domestic users in the village, with data from the 2011 Census of India. For the state of Karnataka, these values are not available, so we instead use hours of electricity supply for all purposes.

Finally, we estimate LASSO regressions to test for nonlinear associations between night lights and electrification.

S2 Sample Composition

	SH_sum_11	API_mean_2011	BIptval_11	St11
Jammu & Kashmir	5953	6330	6441	6682
Himachal Pradesh	10545	18047	13020	20696
Punjab	12321	12237	12715	12715
Chandigarh	10	0	12	12
Uttarakhand	13585	15755	16542	16845
Haryana	6810	6700	6923	6927
Delhi	211	157	219	222
Rajasthan	43969	43143	44783	44796
Uttar Pradesh	91613	97645	107099	107105
Bihar	35572	38871	44935	44937
Sikkim	442	437	451	452
Arunachal Pradesh	0	5364	5590	5590
Nagaland	0	1397	1435	1435
Manipur	402	2340	441	2611
Mizoram	0	712	830	830
Tripura	898	880	901	901
Meghalaya	0	6599	6851	6851
Assam	20263	24501	22422	26550
West Bengal	34502	37878	40978	40978
Jharkhand	27052	29236	32578	32582
Orissa	43345	47560	51446	51474
Chhattisgarh	17312	20121	17914	20167
Madhya Pradesh	54055	51810	55061	55064
Gujarat	18102	17973	18475	18481
Daman & Diu	20	0	25	25
Dadra & Nagar Haveli	70	0	70	70
Maharastra	43174	41179	43943	43943
Andhra Pradesh	25492	27987	26024	28168
Karnataka	28506	27489	29516	29519
Goa	389	347	397	397
Lakshadweep	0	0	27	27
Kerala	1488	1364	1489	1495
Tamil Nadu	15893	15295	16369	16369
Puducherry	93	0	93	95
Andaman & Nicobar Islands	0	0	559	559
Total	552087	599354	626574	645570

Table S1: Number of villages, by state, in the nighttime lights datasets using village shapefiles, the India Lights Project's API, and the pointfiles of village centers. Note that the largest number of cases is for the pointfiles, followed by the India Lights Project and the shapefiles. This is primarily due to the source of the shapefile, ML Infomap, missing some states.

S2.1 State Abbreviations

Abbreviation	Full Name
JK	Jammu and Kashmir
HP	Himachal Pradesh
PB	Punjab
CH	Chandigarh
UT	Uttarakhand
HR	Haryana
DL	Delhi
RJ	Rajasthan
UP	Uttar Pradesh
BR	Bihar
SK	Sikkim
AR	Arunachal Pradesh
NL	Nagaland
MN	Manipur
MZ	Mizoram
TR	Tripura
ML	Meghalaya
AS	Assam
WB	West Bengal
JH	Jharkhand
OR	Orissa
CT	Chhattisgarh
MP	Madhya Pradesh
GJ	Gujarat
DD	Daman and Diu
DN	Dadra and Nagar Haveli
MH	Maharastra
AP	Andhra Pradesh
KA	Karnataka
GA	Goa
LD	Lakshadweep
KL	Kerala
TN	Tamil Nadu
PY	Puducherry

Table S2: Full state names that correspond with the two letter ISO codes in the chart (a) of Figure 2.

S3 Summary Statistics: Census of India 2011, Satellite Data

- Table S3 shows the summary statistics of the Census of India 2011 variables.
- Table S4 shows the summary statistics for the satellite data.

	count	mean	sd	min	max
Electrified HH (%)	595401	50.32	37.55	0.00	100.00
Electrified HH (nbr)	594020	156.51	330.63	0.00	15267.50
Electrified HH (log nbr)	594020	3.68	2.03	0.00	9.63
Area (ha)	595495	409.97	965.07	0.00	373187.00
Distance town (km)	596878	23.90	25.48	0.00	1717.00
Nbr households	595497	282.25	417.34	0.00	15595.00
Nbr people	595497	1395.70	1959.83	1.00	66062.00
SC population	595497	257.66	462.45	0.00	32621.00
ST population	595497	157.15	408.08	0.00	36026.00
Literate population	595501	808.06	1246.46	0.00	50365.00
Village Electrified (=1)	594380	0.54	0.50	0.00	1.00
HH with banking (%)	595401	55.85	30.77	0.00	100.00
HH without assets (%)	595401	24.87	21.87	0.00	100.00
HH with radio (%)	595401	18.16	18.20	0.00	100.00
HH with TV (%)	595401	27.88	25.54	0.00	100.00

Table S3: Summary statistics: Census of India 2011.

	count	mean	sd	min	max
Night lights (shape, mean, 2011)	555903	5.14	7.12	0.00	63.00
Night lights (shape, sd, 2011)	555903	0.69	1.36	0.00	23.44
Night lights (shape, sum, 2011)	555903	32.05	135.57	0.00	28560.00
Night lights (point value, 2011)	630490	5.03	7.28	0.00	63.00
Night lights (bilinear, 2011)	630490	5.03	7.22	0.00	63.00
Night lights (2km, mean, 2011)	498396	3.90	6.37	0.00	63.00
Night lights (2km, sd, 2011)	498396	0.51	1.16	0.00	18.50
Night lights (2km, sum, 2011)	498396	11.04	21.92	0.00	441.00
Night lights (3km, mean, 2011)	464563	3.90	6.34	0.00	63.00
Night lights (3km, sd, 2011)	464563	0.59	1.32	0.00	21.00
Night lights (3km, sum, 2011)	464563	13.77	31.90	0.00	882.00
Night lights (5km, mean, 2011)	418149	3.94	6.53	0.00	63.00
Night lights (5km, sd, 2011)	418149	0.67	1.48	0.00	25.51
Night lights (5km, sum, 2011)	418149	15.98	44.21	0.00	2329.00
Night lights (API, mean, 2011)	599355	2.43	4.43	0.00	58.32
Night lights (API, sd, 2011)	599355	3.62	2.40	1.33	23.07

Table S4: Summary statistics: satellite data for lights.

S4 Maps

- Figure S1 shows the average bilinear interpolation point values for villages in each districts.
- Figure S2 shows the total village population of all villages in each district according to the 2011 Census.
- Figure S3 shows the total number of villages in each district as of the 2011 Census.
- Figure S4 shows the total number of households that are electrified in each district as of the 2011 Census.
- Figure S5 shows the nighttime lights data for Ghazipur, Uttar Pradesh, India overlaid with a shapefile of the village boundaries and with points placed at the latitude and longitude of the center of the village. This figure illustrates the difference between shapefiles and point file GIS data and the difference it can make in calculating the relevant DN.
- Figure S6 shows a screen capture of the India Lights project data for Ghazipur, Uttar Pradesh, India. It appears that they have used a point data file to calculate the DN of luminosity.

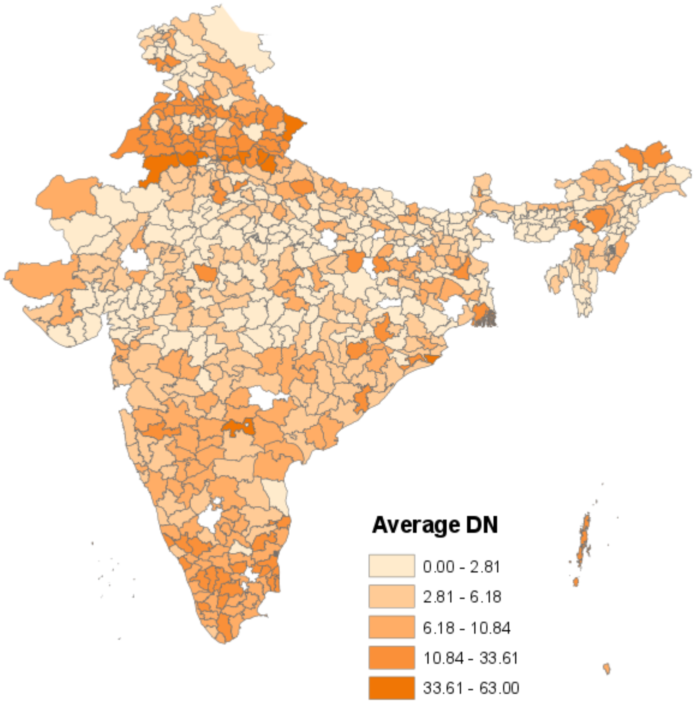


Figure S1: District average DN of nighttime luminosity from bilinear interpolation pointfile village data.

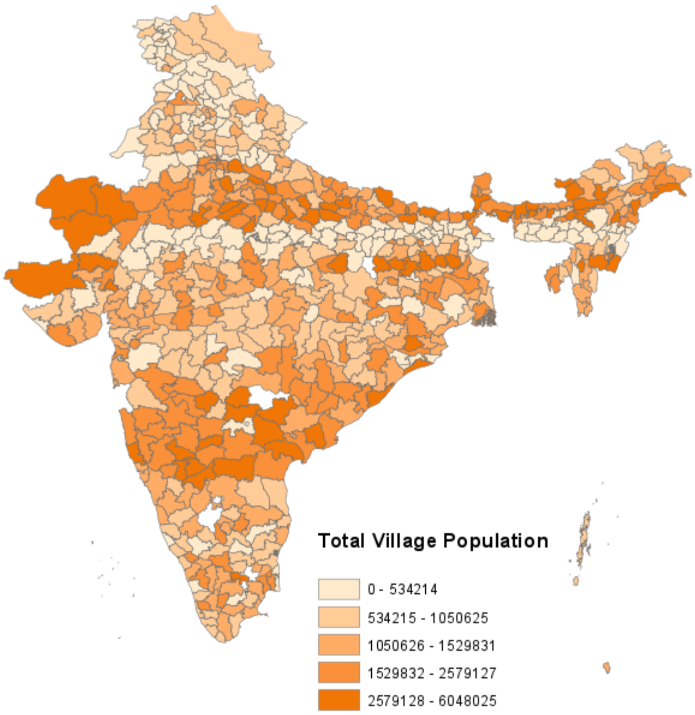


Figure S2: Sum of the total population of all villages in districts in India, 2011 Census.

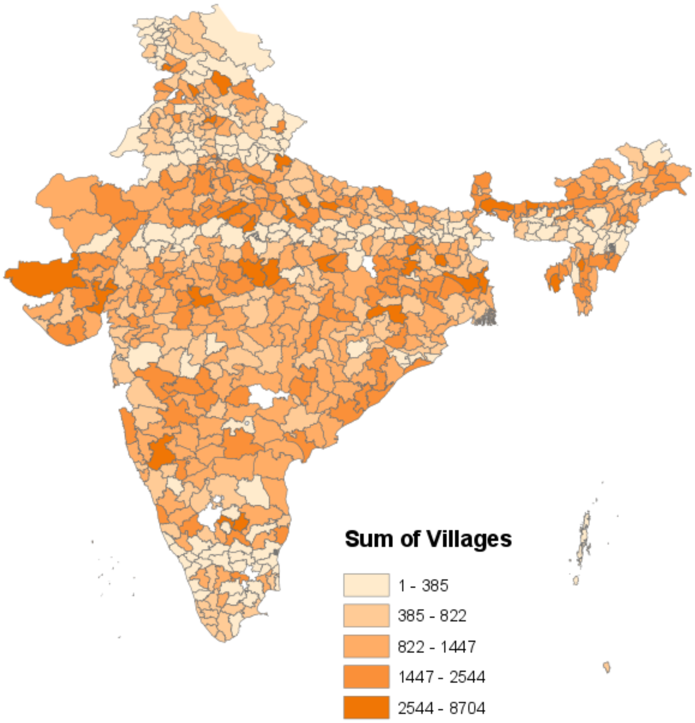


Figure S3: Count of the number of villages in districts in India, 2011 Census.

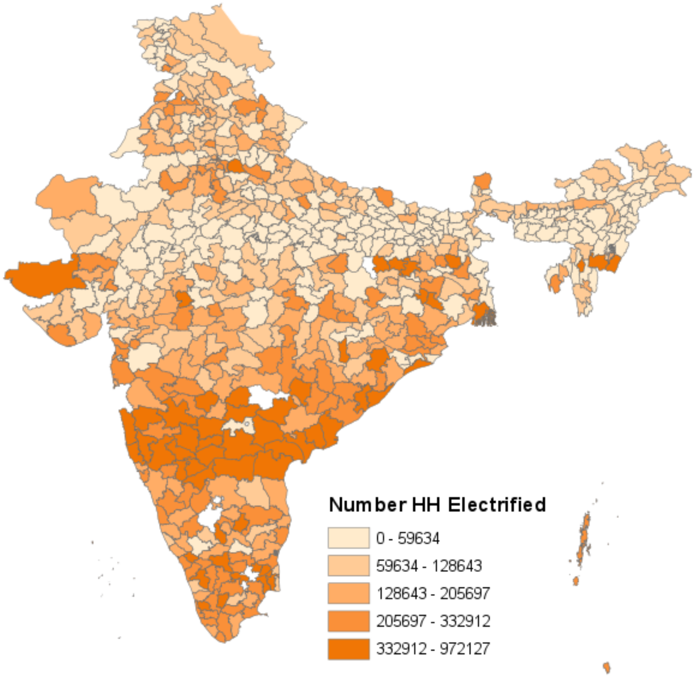


Figure S4: Count of the number of electrified households in districts in India, 2011 Census.

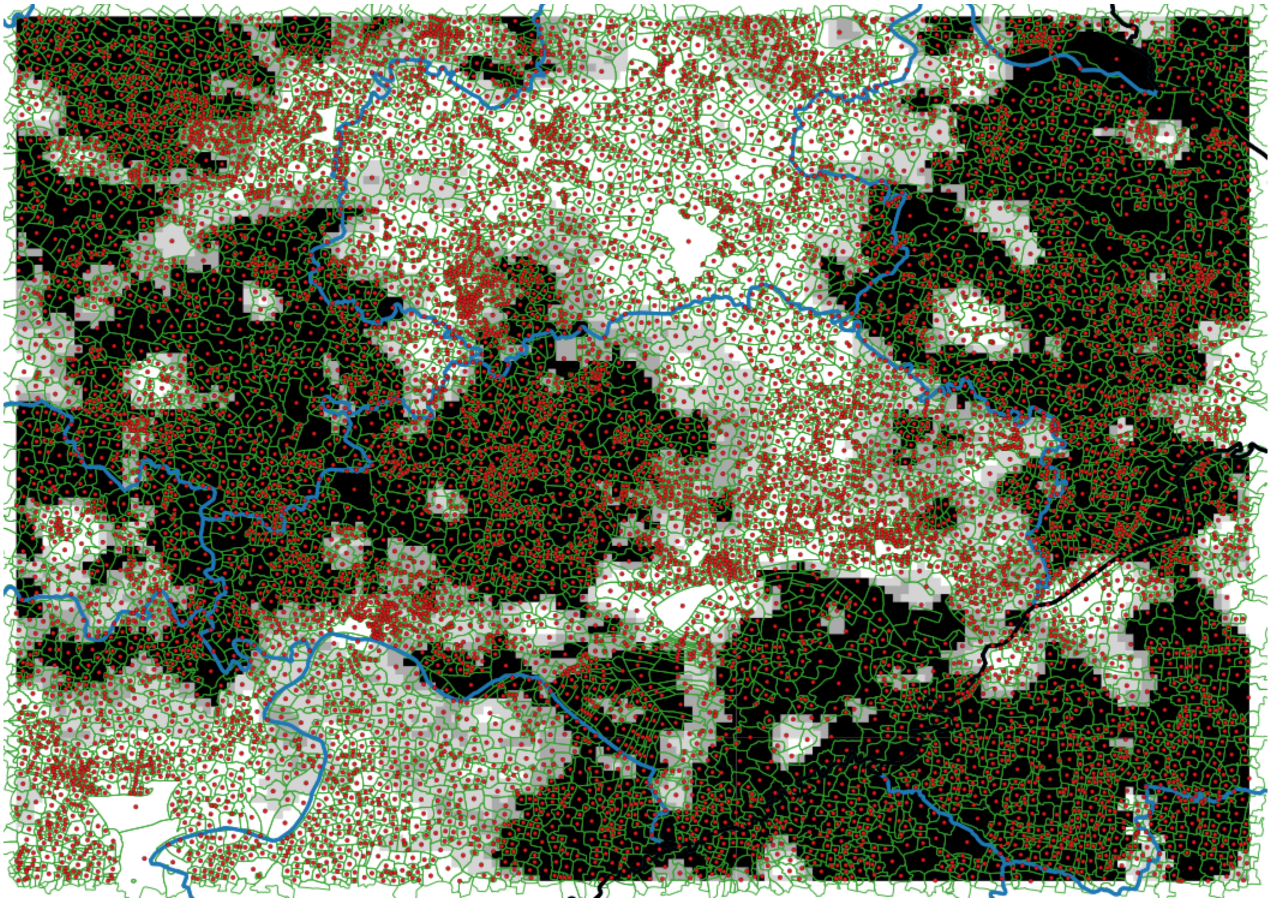
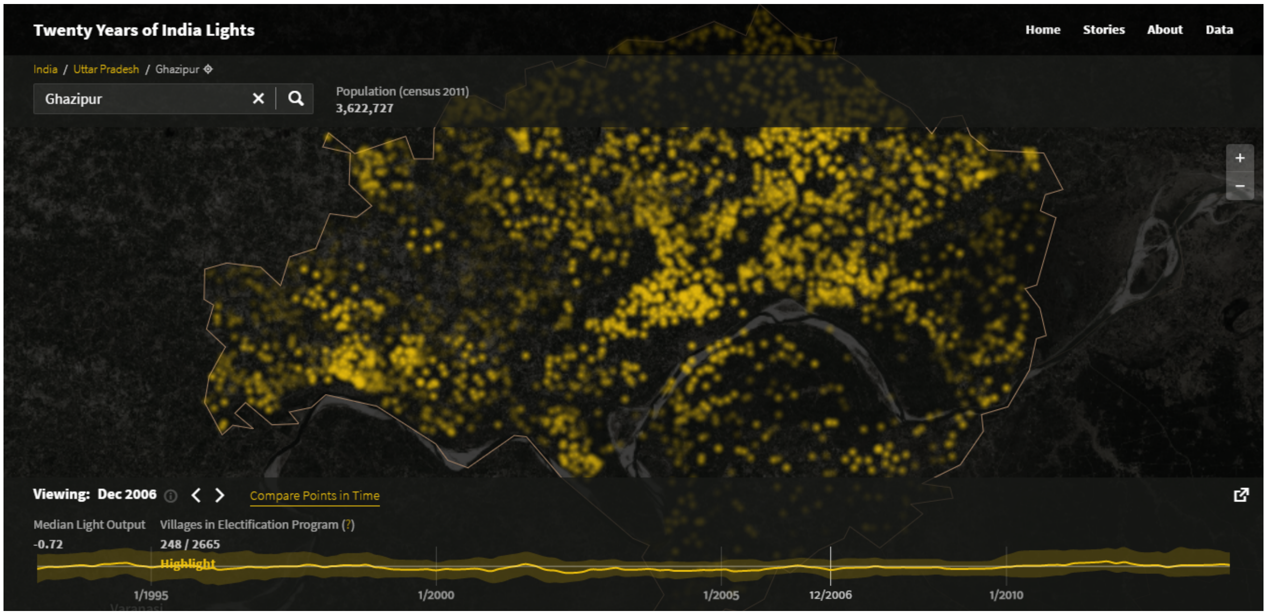


Figure S5: Example of nighttime lights data from Ghazipur, Uttar Pradesh, India in 2011. Two overlays are placed on top of the lights data. The shapefile shows the boundaries of the village. The point data is a point placed at the latitude and longitude of the center of the village.



Twenty Years of India Lights

India / Uttar Pradesh / Ghazipur

Ghazipur

Population (census 2011)
3,622,727

Viewing: Dec 2006

Median Light Output
-0.72

Villages in Electification Program (?)
248 / 2665

Figure S6: Example of India Lights Project data from Ghazipur, Uttar Pradesh, India. The researchers used point data to calculate luminosity DN during two to three months of the year.

S5 Graphical Data Description: Census of India 2011, Satellite Data

- Figure S7 shows a hexabin plot of the percentage of electrified households with the log sum of the 2011 shape file measure.
- Figure S8 shows the correlations with the shape file measures and electrification.
- Figure S9 shows the correlations with the 2 km perimeter measures and electrification.
- Figure S10 shows the correlations with the 3 km perimeter measures and electrification.
- Figure S11 shows the correlations with the 5 km perimeter measures and electrification.
- Figure S12 shows the correlations with the maximum of various measures and electrification.
- Figure S13 shows the correlations with the mean of various measures and electrification.
- Figure S14 shows the correlations with the minimum of various measures and electrification.
- Figure S15 shows the correlations with the sum of various measures and electrification.

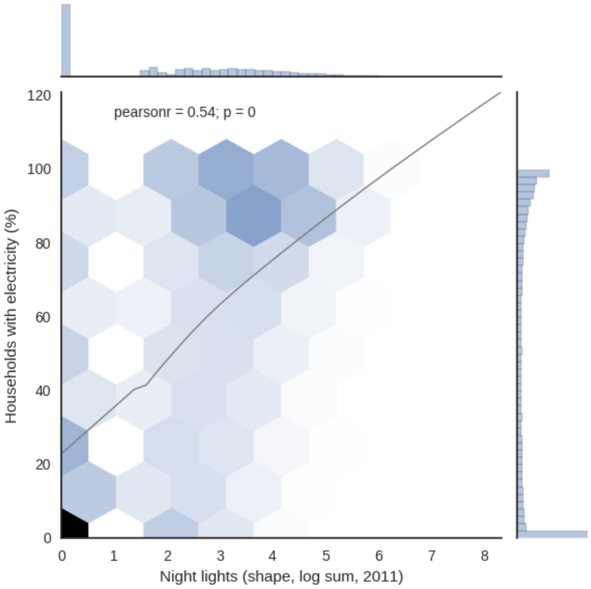


Figure S7: Hexabin plot of the percentage of electrified households with the log sum of the 2011 shape file measure.

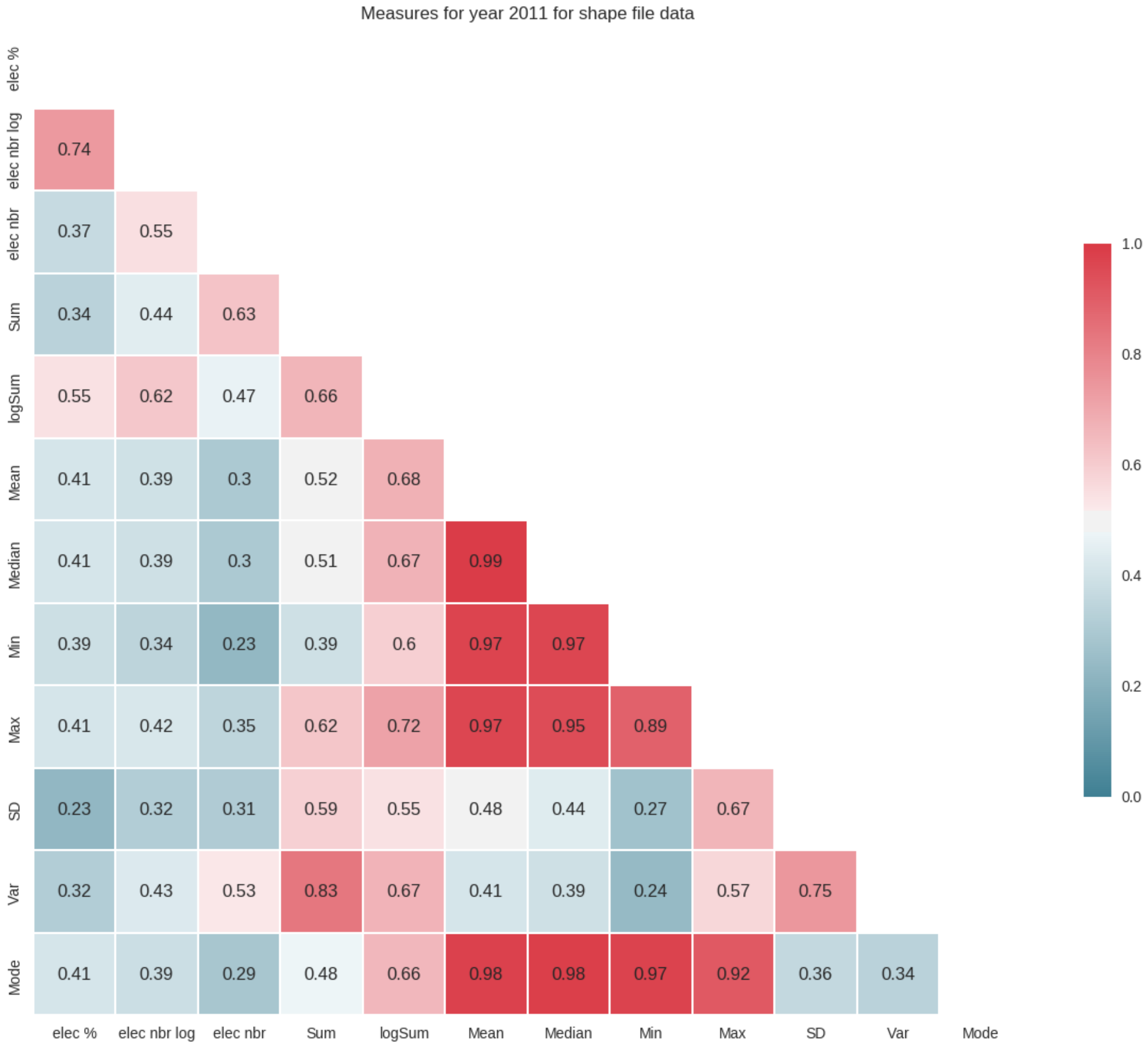


Figure S8: Correlation map for the shape file measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. ‘elec %’ = percentage of households that are electrified in the village; ‘elec nbr’ and ‘elec nbr log’ = the number and log of the number of households that are electrified in the village; ‘Sum’ and ‘log Sum’ = sum and log of the sum of the luminosity of all pixels within the shape boundaries of the village; ‘Mean’, ‘Median’, ‘Max’, ‘Min’, ‘SD’ = mean, median, max, min and standard deviation of the luminosity across all pixels within the shape boundaries of the village. ‘Var’ = the number of unique values of pixel luminosity within the shape boundaries of the village. ‘Maj’ = the most frequently occurring value of pixel luminosity within the shape boundaries of the village. ‘Mino’ = the least frequently occurring value of pixel luminosity within the shape boundaries of the village.

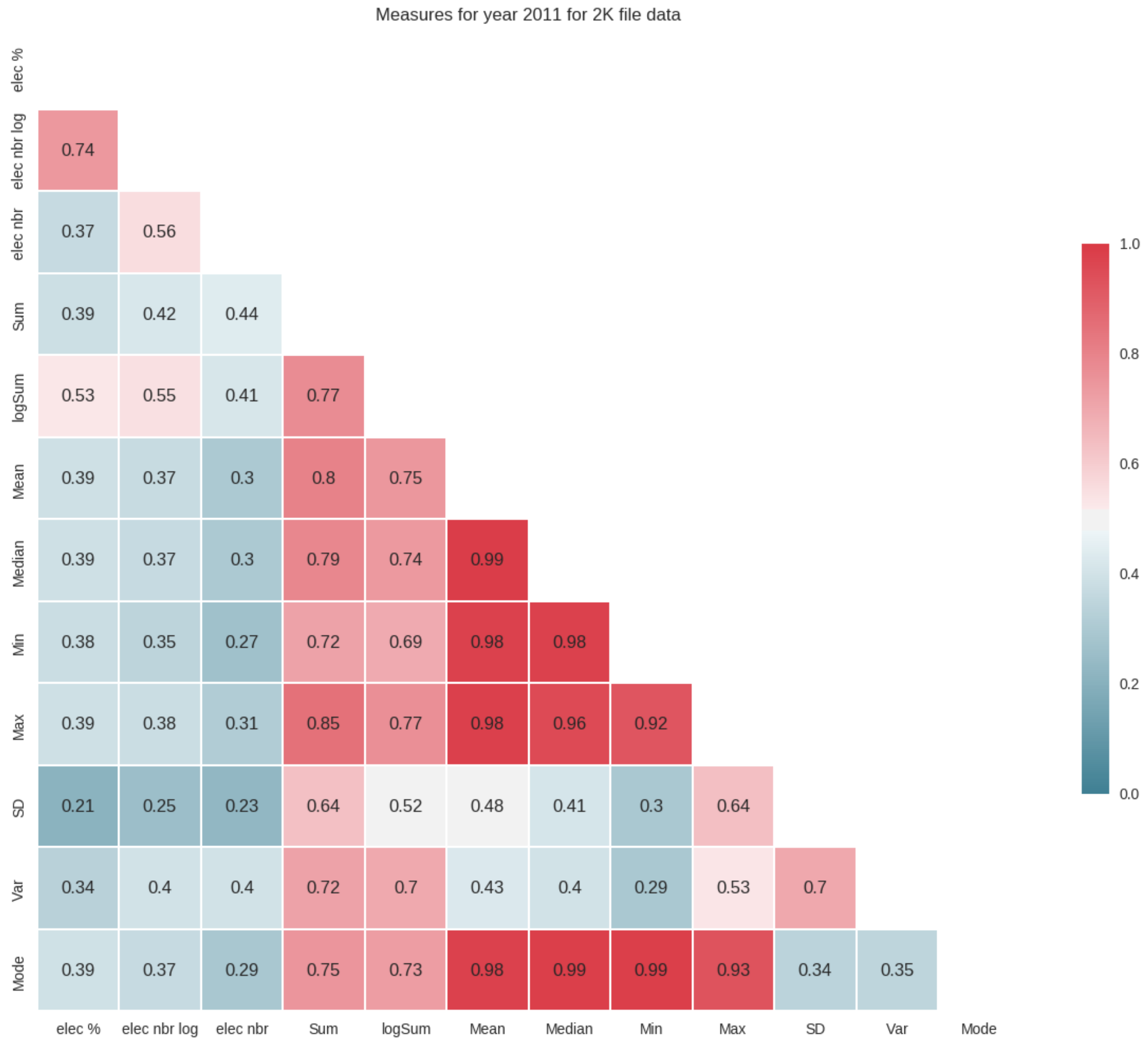


Figure S9: Correlation map for the 2 km perimeter measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. ‘elec %’ = percentage of households that are electrified in the village; ‘elec nbr’ and ‘elec nbr log’ = the number and log of the number of households that are electrified in the village; ‘Sum’ and ‘log Sum’ = sum and log of the sum of the luminosity of all pixels within the 2 km buffer zone around the village centroid; ‘Mean’, ‘Median’, ‘Max’, ‘Min’, ‘SD’ = mean, median, max, min and standard deviation of the luminosity across all pixels within the 2 km buffer zone around the village centroid. ‘Var’ = the number of unique values of pixel luminosity within the 2 km buffer zone around the village centroid. ‘Maj’ = the most frequently occurring value of pixel luminosity within the 2 km buffer zone around the village centroid. ‘Mino’ = the least frequently occurring value of pixel luminosity within the 2 km buffer zone around the village centroid.

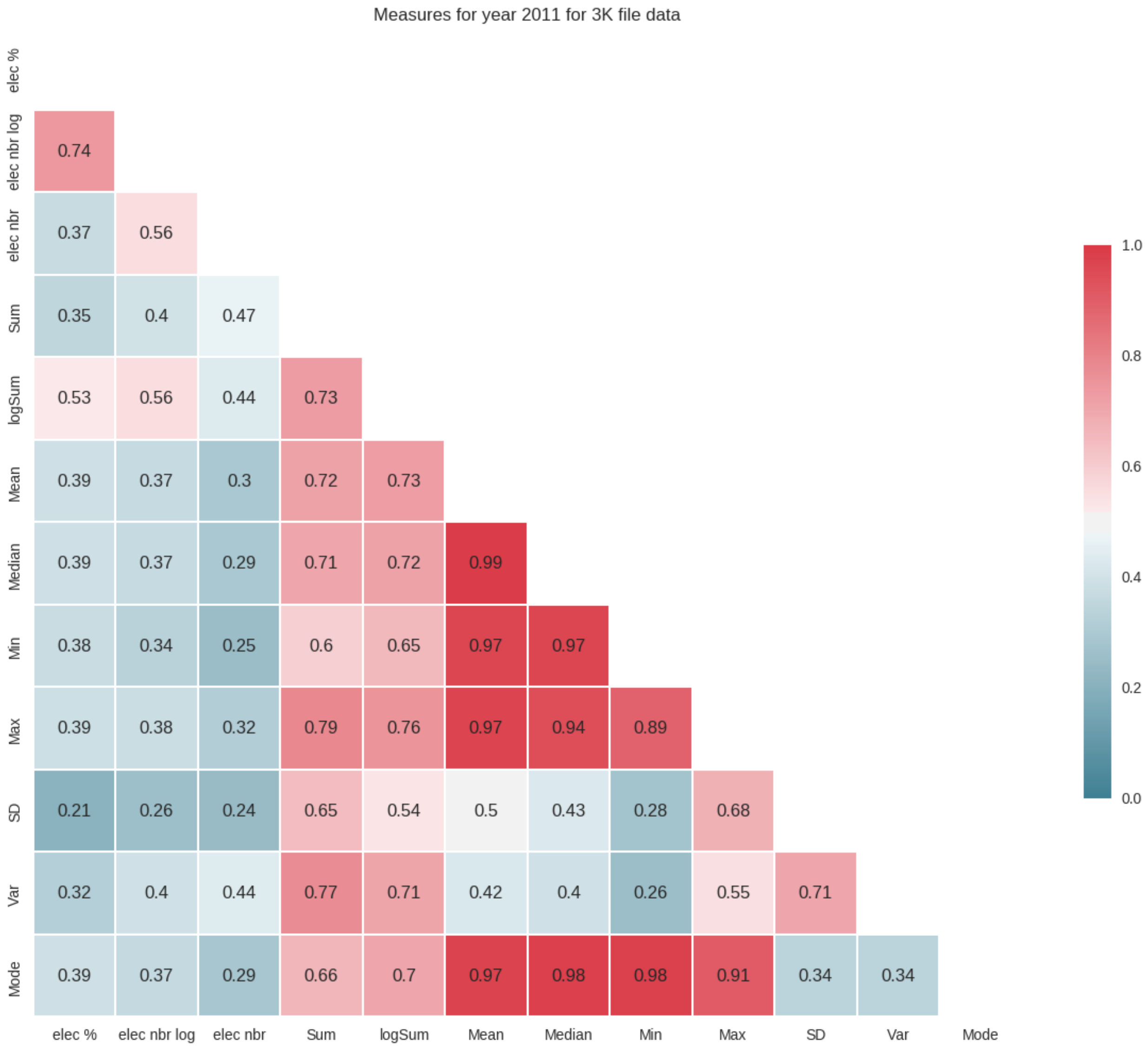


Figure S10: Correlation map for the 3 km perimeter measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. 'elec %' = percentage of households that are electrified in the village; 'elec nbr' and 'elec nbr log' = the number and log of the number of households that are electrified in the village; 'Sum' and 'log Sum' = sum and log of the sum of the luminosity of all pixels within the 3 km buffer zone around the village centroid; 'Mean', 'Median', 'Max', 'Min', 'SD' = mean, median, max, min and standard deviation of the luminosity across all pixels within the 3 km buffer zone around the village centroid. 'Var' = the number of unique values of pixel luminosity within the 3 km buffer zone around the village centroid. 'Maj' = the most frequently occurring value of pixel luminosity within the 3 km buffer zone around the village centroid. 'Mino' = the least frequently occurring value of pixel luminosity within the 3 km buffer zone around the village centroid.

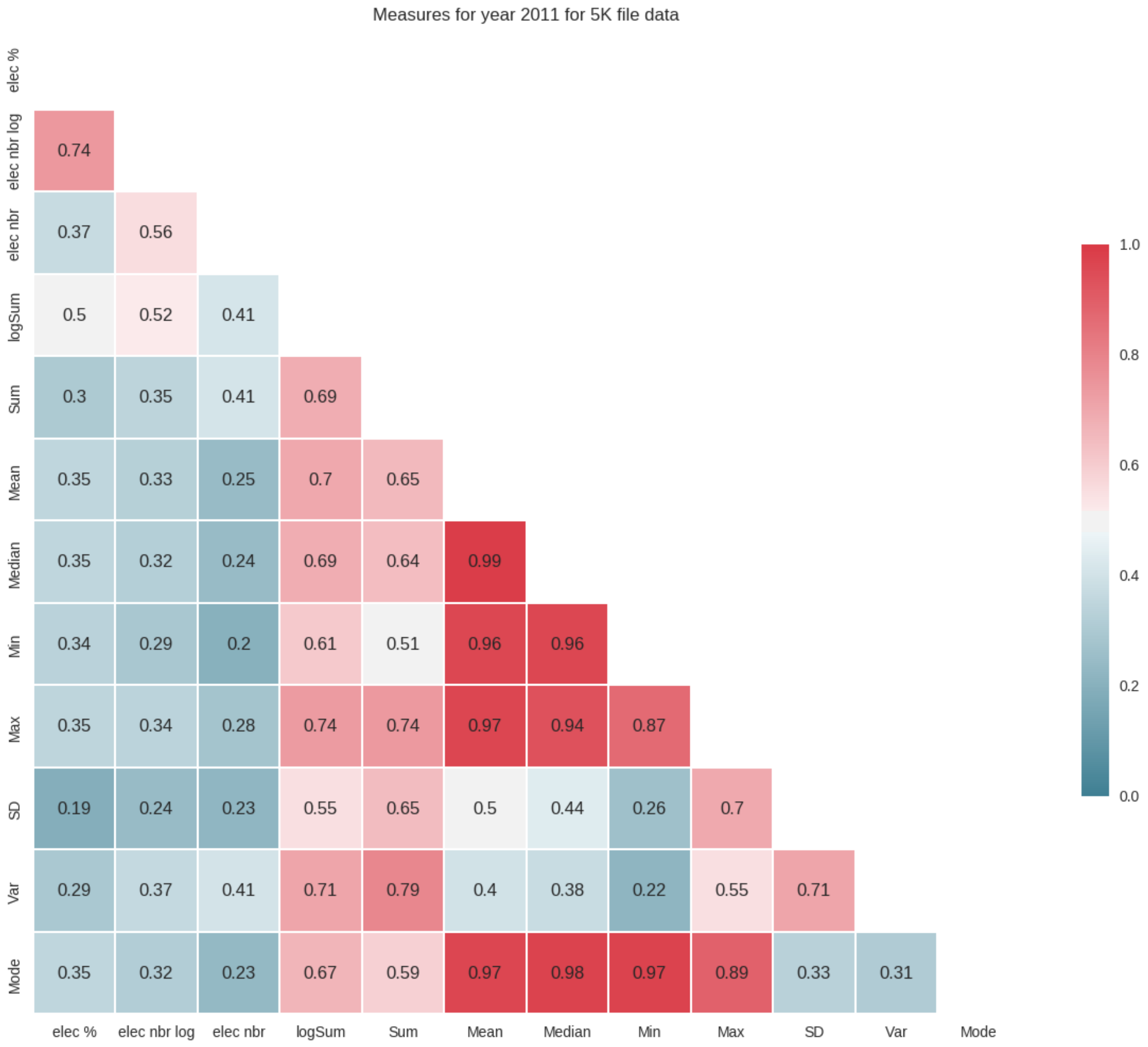


Figure S11: Correlation map for the 5 km perimeter measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. 'elec %' = percentage of households that are electrified in the village; 'elec nbr' and 'elec nbr log' = the number and log of the number of households that are electrified in the village; 'Sum' and 'log Sum' = sum and log of the sum of the luminosity of all pixels within the 5 km buffer zone around the village centroid; 'Mean', 'Median', 'Max', 'Min', 'SD' = mean, median, max, min and standard deviation of the luminosity across all pixels within the 5 km buffer zone around the village centroid. 'Var' = the number of unique values of pixel luminosity within the 5 km buffer zone around the village centroid. 'Maj' = the most frequently occurring value of pixel luminosity within the 5 km buffer zone around the village centroid. 'Mino' = the least frequently occurring value of pixel luminosity within the 5 km buffer zone around the village centroid.

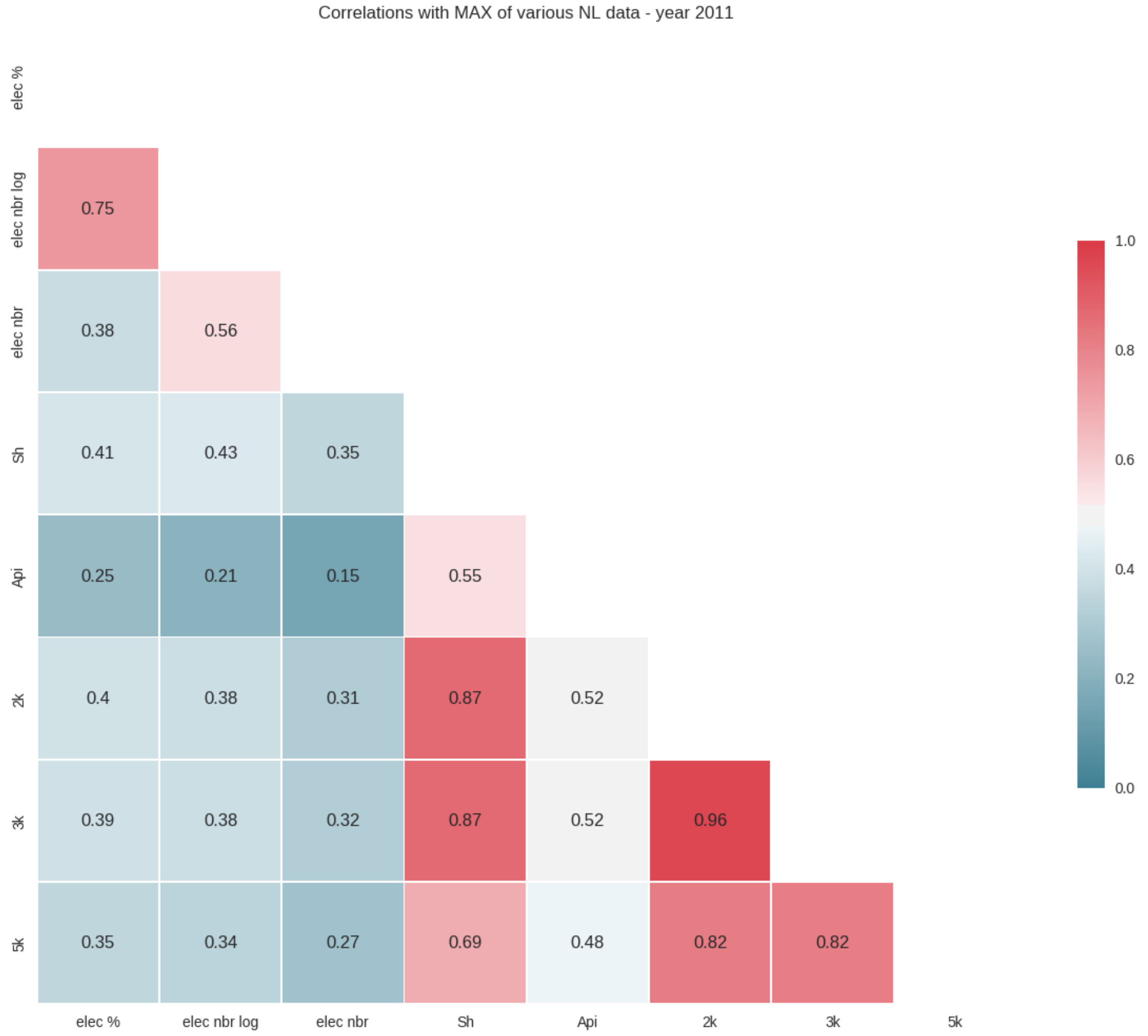


Figure S12: Correlation map for the maximum of various measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. 'elec %' = percentage of households that are electrified in the village; 'elec nbr' and 'elec nbr log' = the number and log of the number of households that are electrified in the village; 'Sh' = maximum of the luminosity across all pixels within the shape boundaries of the village. 'Api' = maximum of the luminosity provided by the India Lights Project API. '2k' = maximum of the luminosity across all pixels within the 2 km buffer zone around the village centroid. '3k' = maximum of the luminosity across all pixels within the 3 km buffer zone around the village centroid. '5k' = maximum of the luminosity across all pixels within the 5 km buffer zone around the village centroid.

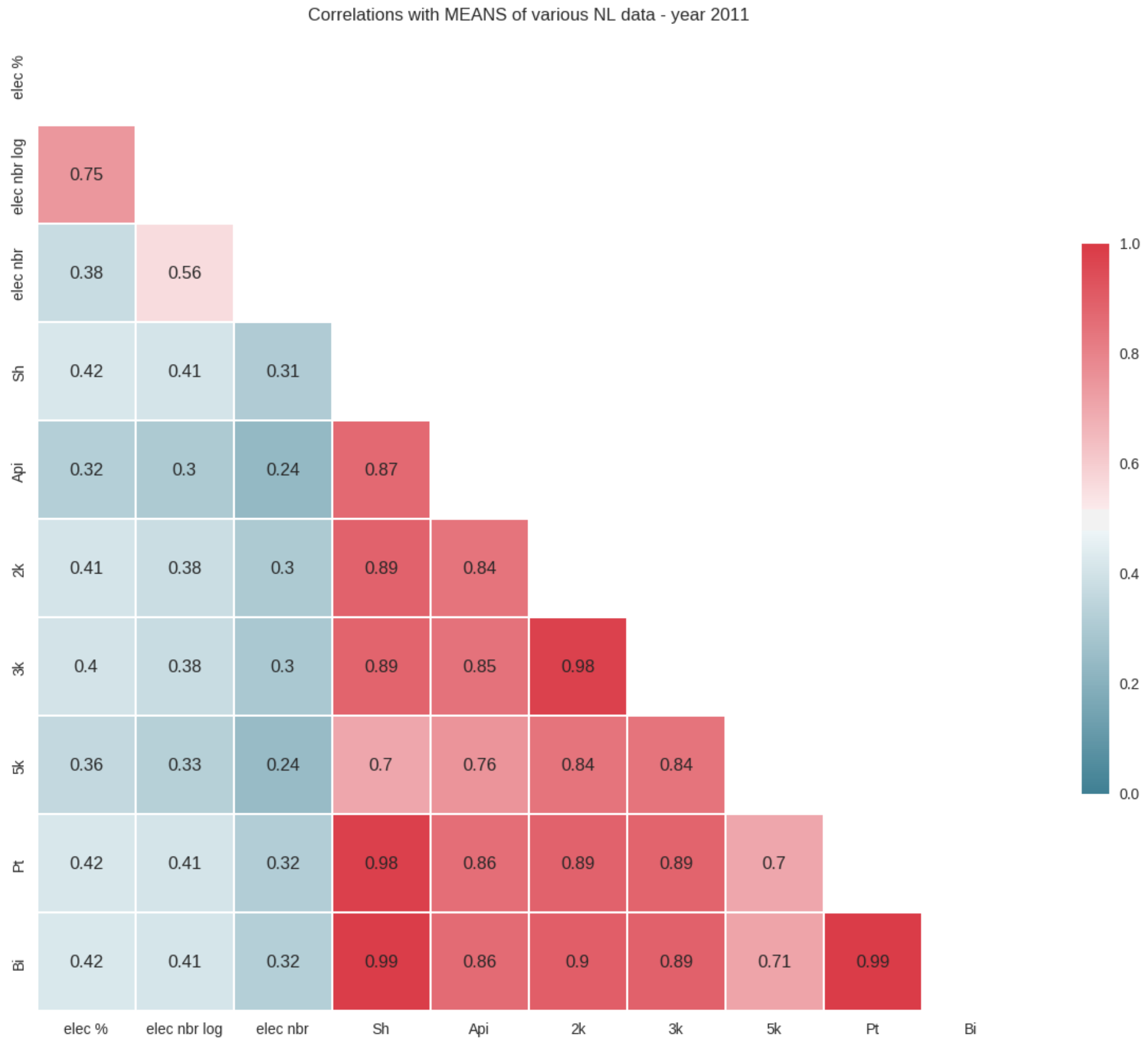


Figure S13: Correlation map for the mean of various measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. 'elec %' = percentage of households that are electrified in the village; 'elec nbr' and 'elec nbr log' = the number and log of the number of households that are electrified in the village; 'Sh' = mean of the luminosity across all pixels within the shape boundaries of the village. 'Api' = mean of the luminosity provided by the India Lights Project API. '2k' = mean of the luminosity across all pixels within the 2 km buffer zone around the village centroid. '3k' = mean of the luminosity across all pixels within the 3 km buffer zone around the village centroid. '5k' = mean of the luminosity across all pixels within the 5 km buffer zone around the village centroid. 'Pt' = luminosity of the pixel at the longitude and latitude of the village centroid; 'Bi' = luminosity of the pixel at the longitude and latitude of the village centroid using linear interpolation of surrounding pixels;

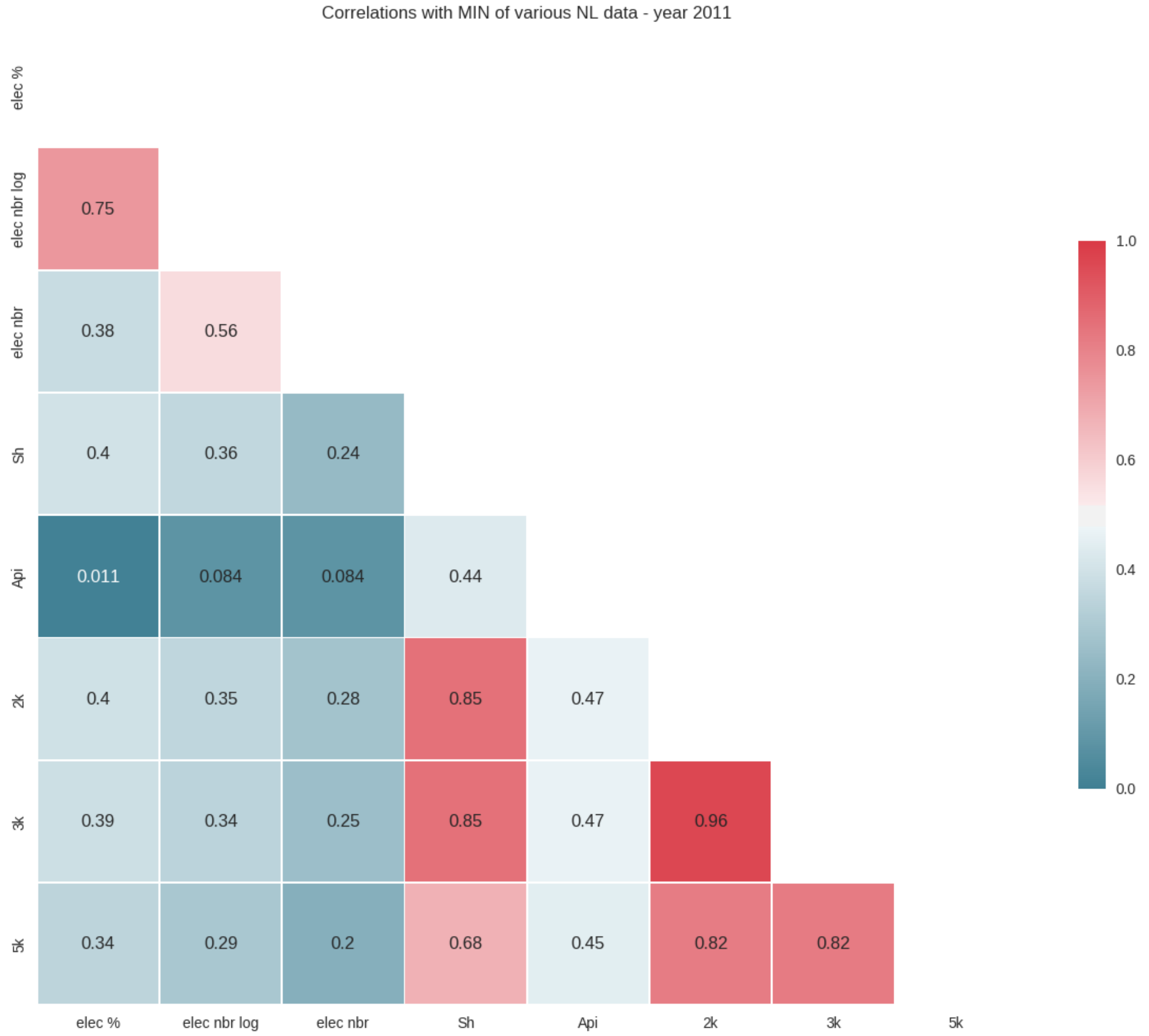


Figure S14: Correlation map for the minimum of various measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. 'elec %' = percentage of households that are electrified in the village; 'elec nbr' and 'elec nbr log' = the number and log of the number of households that are electrified in the village; 'Sh' = minimum of the luminosity across all pixels within the shape boundaries of the village. 'Api' = minimum of the luminosity provided by the India Lights Project API. '2k' = minimum of the luminosity across all pixels within the 2 km buffer zone around the village centroid. '3k' = minimum of the luminosity across all pixels within the 3 km buffer zone around the village centroid. '5k' = minimum of the luminosity across all pixels within the 5 km buffer zone around the village centroid.

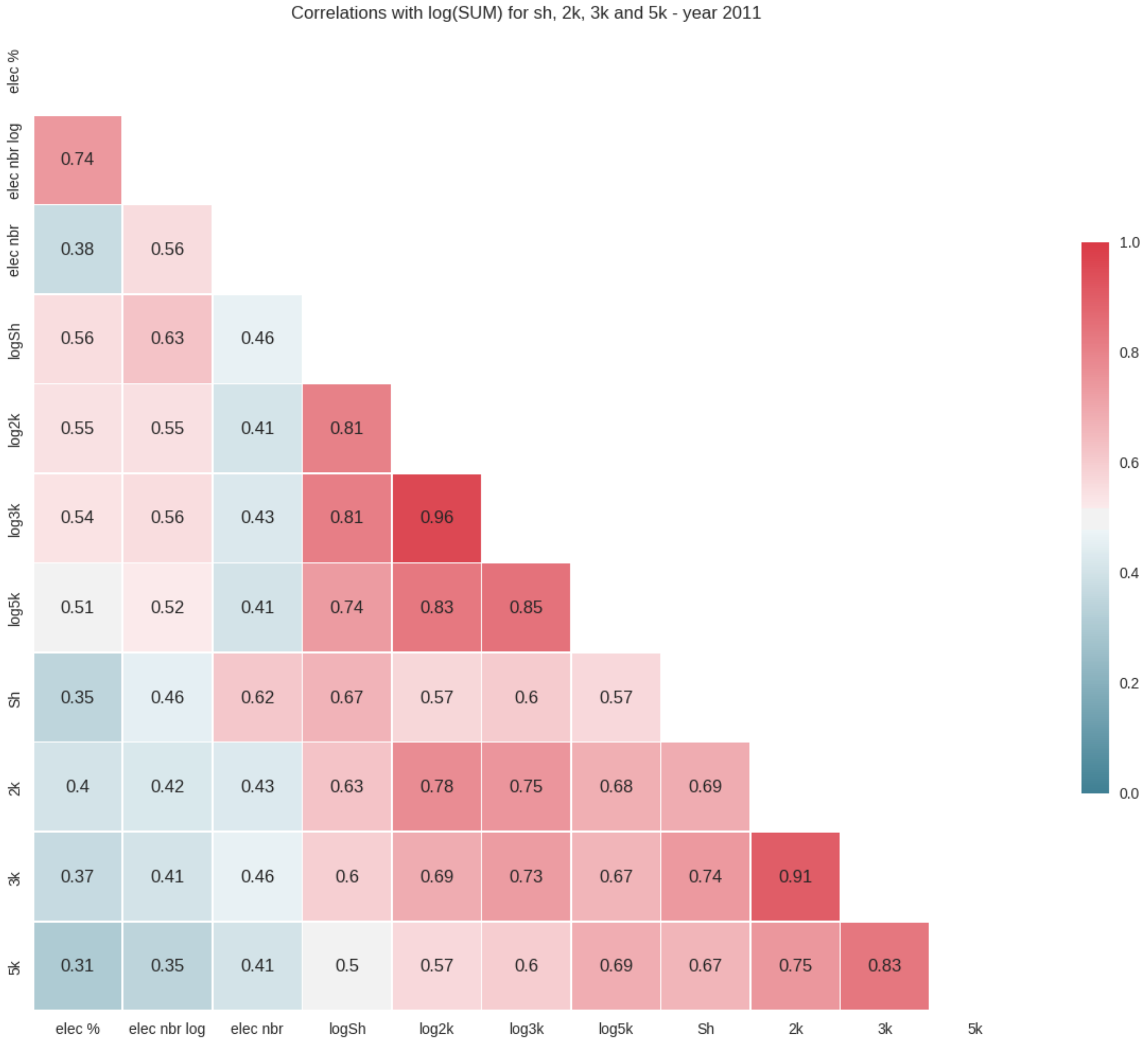
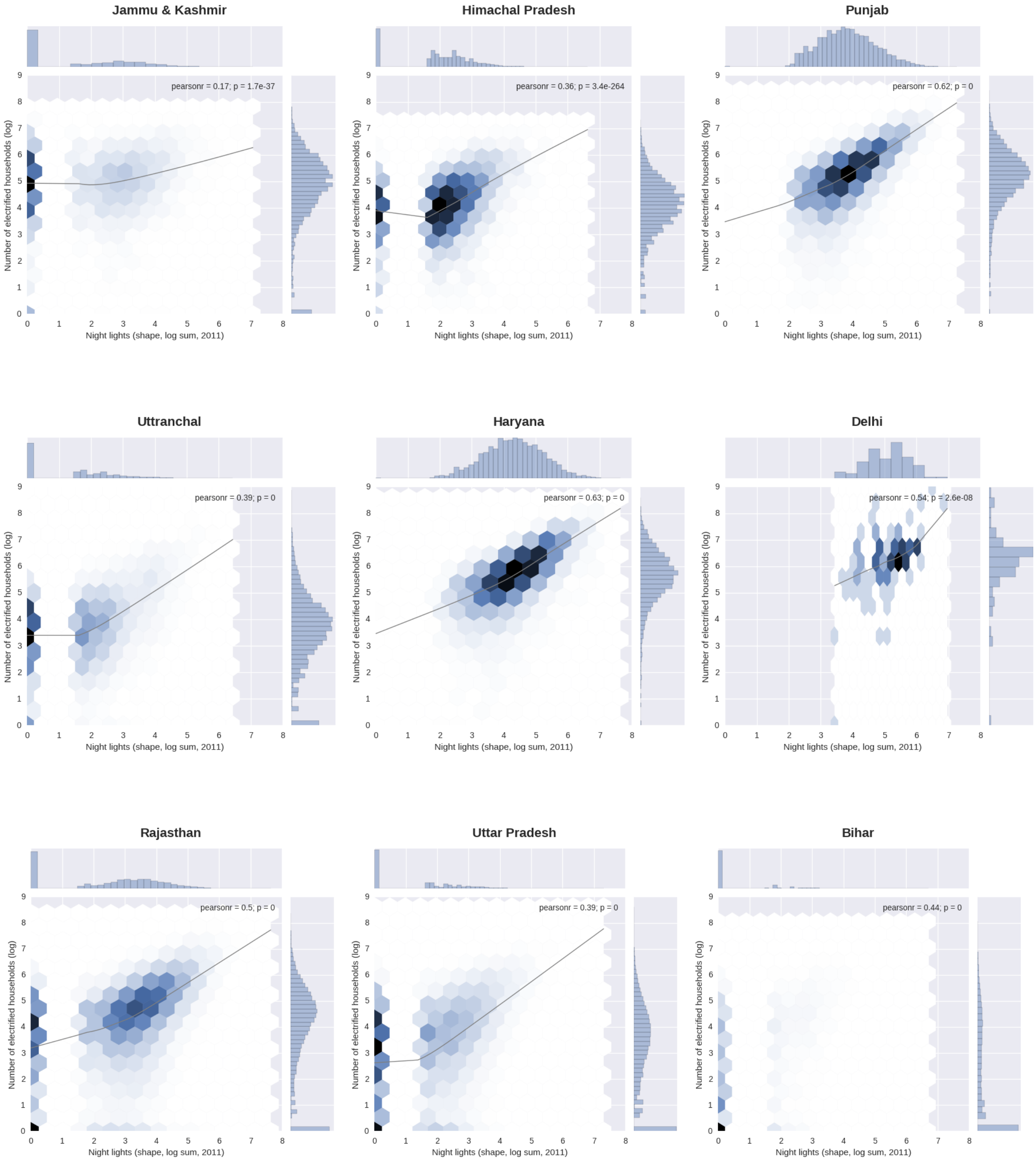


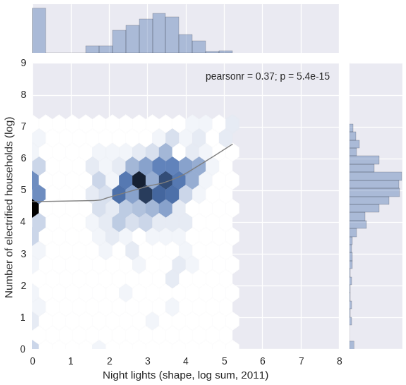
Figure S15: Correlation map for the sum of various measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011. ‘elec %’ = percentage of households that are electrified in the village. ‘elec nbr’ and ‘elec nbr log’ = the number and log of the number of households that are electrified in the village. ‘log ShSum’ = log of the sum of the luminosity of the pixels within the shape boundaries of the village. ‘log 2kSum’, ‘log 3kSum’, and ‘log 5kSum’ = log of the sum of the luminosity of all the pixels within a 2 km, 3 km and resp. 5 km circle centered at the village centroid. ‘Sh’ = sum of the luminosity of the pixels within the shape boundaries of the village. ‘2k’, ‘3k’, and ‘5k’ = sum of the luminosity of all the pixels within a 2 km, 3 km and resp. 5 km circle centered at the village centroid.

S6 Graphical Data Description by State

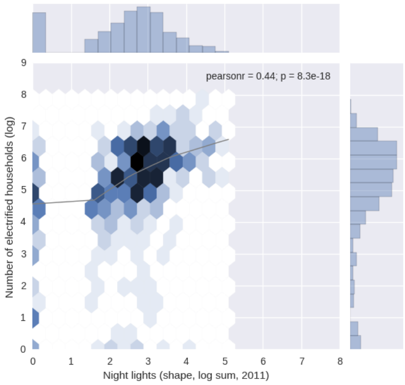
- Figures S6-S6 show the hexabin plots for the logarithmized DN sum for village shapefiles against the number of electrified households (logarithmized).



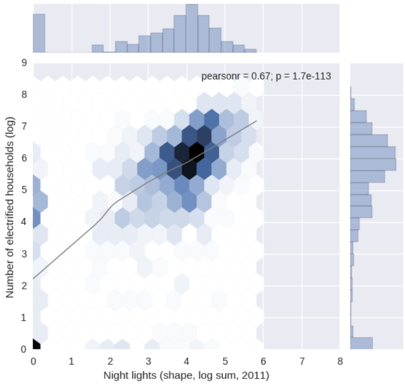
Sikkim



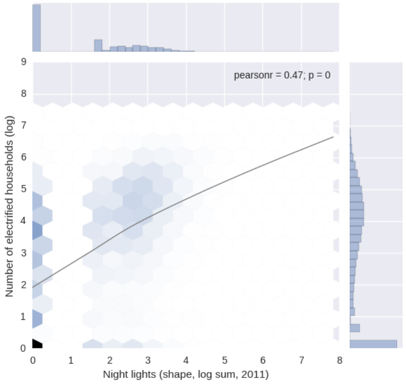
Manipur



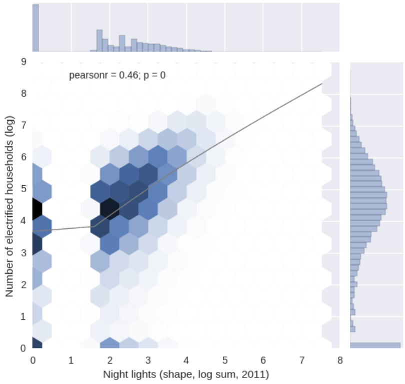
Tripura



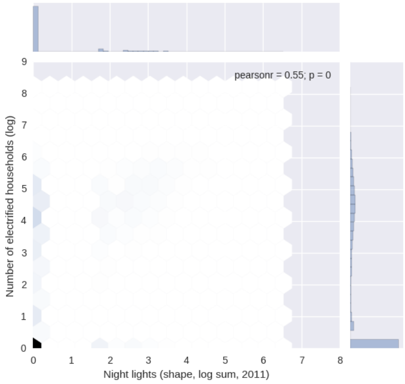
Assam



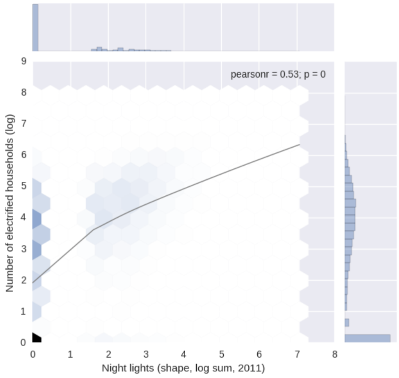
West Bengal



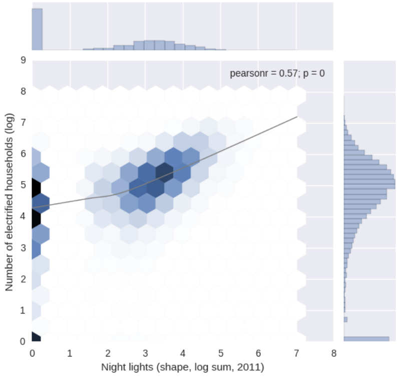
Jharkhand



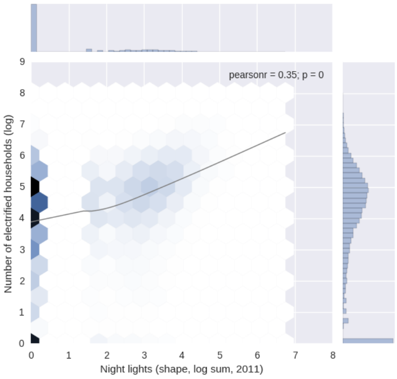
Orissa



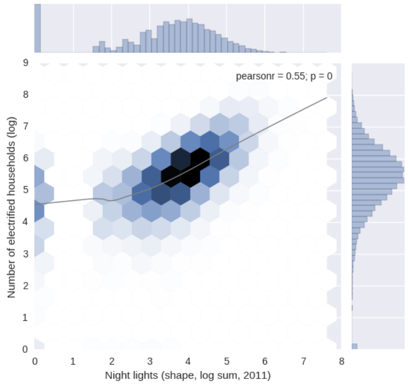
Chhattisgarh



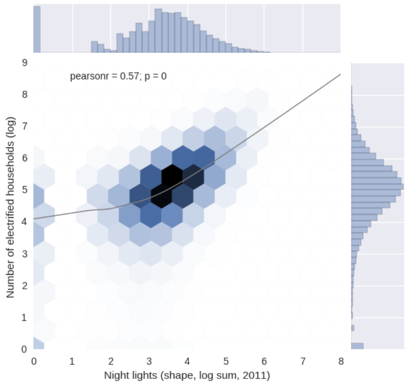
Madhya Pradesh



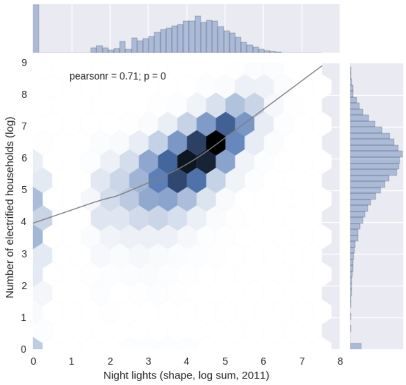
Gujarat



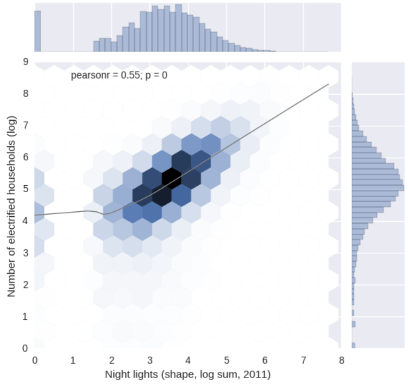
Maharashtra



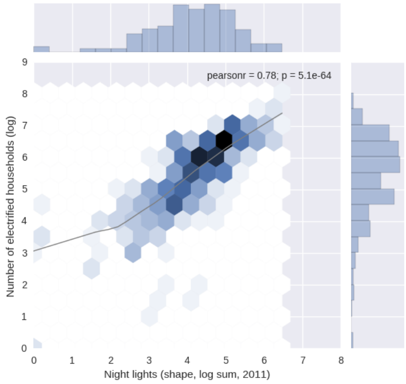
Andhra Pradesh



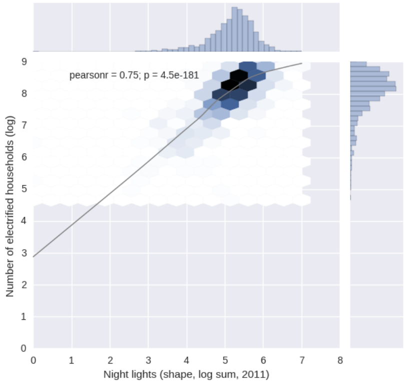
Karnataka



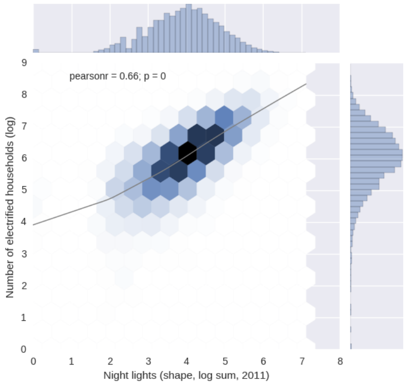
Goa



Kerala



Tamil Nadu



S7 Regression Output

- Table S5 shows the placebo regression results for non-electrified households.

- Table S6 shows the regressions results for polynomials up to the fourth degree.

- Table S7 shows the regressions results for quantiles. The baseline category is zero night lights.

- Table S8 shows the linear regression results using the yearly mean DN values from the India Lights Project.

- Table S9 shows the linear regression results using the yearly maximum DN values from the India Lights Project.

- Table S10 shows the linear regressions results with night lights as the dependent variable, including percentage of households without any assets as an independent variable.

- Table S11 shows the linear regressions results with night lights as the dependent variable, including percentage of households with a TV as an independent variable.

- Table S12 shows the linear regressions results with night lights as the dependent variable, including percentage of households with a bank account as an independent variable.

- Figure S16 replicates our models of asset ownership, as a proxy for poverty, but without the controls for the number of electrified households. As the reader will note, this increases the relationship between nightlights and individuals without assets or a TV, although the result for banking remain similar. This emphasizes the patterns we highlight in the main paper – to the extent that energy poverty represents a different concept than income poverty, much of nighttime luminosity's ability to pick up on poverty is through electrification.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Night lights (shape, log sum, 2011)	-0.101*	0.089**	0.131***	0.184***	0.144	0.456***	0.457***	0.497***
	(0.051)	(0.042)	(0.031)	(0.031)	(0.120)	(0.050)	(0.036)	(0.033)
Night lights (shape, sum) (=0)					1.015**	1.422***	1.253***	1.220***
					(0.444)	(0.163)	(0.116)	(0.095)
Fixed effects: state	No	Yes	No	No	No	Yes	No	No
Fixed effects: district	No	No	Yes	No	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes	No	No	No	Yes
R-squared	0.012	0.009	0.018	0.030	0.032	0.062	0.062	0.070
Observations	516769	516769	516769	516769	516769	516769	516769	516769

Standard errors in parentheses
Dependent variable: Number of households without electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S5: Linear regressions, with standard errors clustered by state.

	(1)	(2)	(3)	(4)
Night lights (shape, log sum, 2011)	0.701***	0.450***	0.180	0.214
	(0.065)	(0.150)	(0.256)	(0.358)
Night lights 2 (shape, log of sum, 2011)		0.059**	0.204**	0.175
		(0.024)	(0.094)	(0.197)
Night lights 3 (shape, log of sum, 2011)			-0.019*	-0.011
			(0.010)	(0.039)
Night lights 4 (shape, log of sum, 2011)				-0.001
				(0.003)
Fixed effects: state	No	No	No	No
Fixed effects: district	No	No	No	No
Fixed effects: subdistrict	No	No	No	No
R-squared	0.381	0.386	0.388	0.388
Observations	516769	516769	516769	516769

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S6: Linear regressions with polynomials, with standard errors clustered by state.

	(1)	(2)	(3)	(4)
Night lights 0-20th percentile	0.937***	0.775***	0.557***	0.452***
	(0.249)	(0.183)	(0.116)	(0.081)
Night lights 20th-40th percentile	1.490***	1.204***	0.962***	0.840***
	(0.247)	(0.166)	(0.115)	(0.085)
Night lights 40th-60th percentile	2.034***	1.596***	1.343***	1.216***
	(0.252)	(0.156)	(0.117)	(0.095)
Night lights 60th-80th percentile	2.589***	2.008***	1.756***	1.625***
	(0.262)	(0.144)	(0.112)	(0.095)
Night lights 80th-100th percentile	3.471***	2.725***	2.462***	2.311***
	(0.279)	(0.140)	(0.116)	(0.093)
Fixed effects: state	No	Yes	No	No
Fixed effects: district	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes
R-squared	0.379	0.212	0.160	0.125
Observations	516769	516769	516769	516769

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
Night lights: Omitted category is when night light values are null.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S7: Linear regressions with dummy variables, with standard errors clustered by state.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
India Lights (Log Mean Yearly Value)	1.093***	0.755***	0.630***	0.548***	0.828***	0.550***	0.491***	0.458***
	(0.003)	(0.073)	(0.052)	(0.043)	(0.097)	(0.047)	(0.031)	(0.030)
India Lights (Mean) (=0)					-1.052***	-0.859***	-0.649***	-0.454***
					(0.268)	(0.130)	(0.078)	(0.056)
Fixed effects: state	No	Yes	No	No	No	Yes	No	No
Fixed effects: district	No	No	Yes	No	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes	No	No	No	Yes
R-squared	0.188	0.100	0.062	0.033	0.218	0.128	0.078	0.040
Observations	512918	512918	512918	512918	512918	512918	512918	512918

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S8: Linear regressions using India Lights API data (yearly mean of reported values), with standard errors clustered by state.

	(1)	(2)	(3)	(4)
India Lights (Log Maximum Yearly Value)	0.723***	0.409***	0.294***	0.193***
	(0.004)	(0.056)	(0.045)	(0.045)
Fixed effects: state	No	Yes	No	No
Fixed effects: district	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes
R-squared	0.066	0.028	0.014	0.005
Observations	512918	512918	512918	512918

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S9: Linear regressions using India Lights API data (yearly maximum of reported values), with standard errors clustered by state. Models 5-8 omitted since there are no zero values.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Electrified HH (log nbr)	0.525***	0.520***	0.364***	0.337***	0.300***	0.277***	0.239***	0.228***
	(0.048)	(0.049)	(0.022)	(0.019)	(0.021)	(0.020)	(0.017)	(0.017)
Non-electrified HH (log nbr)	-0.086**	-0.090**	0.019	0.041	0.049***	0.069***	0.080***	0.095***
	(0.041)	(0.044)	(0.026)	(0.024)	(0.017)	(0.016)	(0.014)	(0.013)
HH without assets (%)	-0.008*	-0.005	-0.010***	-0.008***	-0.008***	-0.007***	-0.005***	-0.004***
	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Distance town (log km)		-0.225***		-0.378***		-0.364***		-0.329***
		(0.066)		(0.035)		(0.030)		(0.034)
Fixed effects: state	No	No	Yes	Yes	No	No	No	No
Fixed effects: district	No	No	No	No	Yes	Yes	No	No
Fixed effects: subdistrict	No	No	No	No	No	No	Yes	Yes
R-squared	0.397	0.409	0.230	0.272	0.170	0.215	0.131	0.166
Observations	516769	516769	516769	516769	516769	516769	516769	516769

Standard errors in parentheses
Dependent variable: Night lights (shape, log of sum, 2011).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S10: Linear regressions for night lights and percentage of households with no assets. Standard errors clustered by state.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Electrified HH (log nbr)	0.398***	0.402***	0.281***	0.265***	0.238***	0.225***	0.197***	0.191***
	(0.052)	(0.054)	(0.023)	(0.022)	(0.020)	(0.020)	(0.016)	(0.017)
Non-electrified HH (log nbr)	0.042	0.031	0.094***	0.105***	0.099***	0.112***	0.112***	0.123***
	(0.046)	(0.048)	(0.023)	(0.021)	(0.016)	(0.015)	(0.014)	(0.012)
HH with TV (%)	0.022***	0.020***	0.022***	0.019***	0.018***	0.016***	0.013***	0.011***
	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Distance town (log km)		-0.159**		-0.338***		-0.339***		-0.313***
		(0.067)		(0.029)		(0.026)		(0.031)
Fixed effects: state	No	No	Yes	Yes	No	No	No	No
Fixed effects: district	No	No	No	No	Yes	Yes	No	No
Fixed effects: subdistrict	No	No	No	No	No	No	Yes	Yes
R-squared	0.445	0.451	0.269	0.302	0.197	0.235	0.147	0.179
Observations	516769	516769	516769	516769	516769	516769	516769	516769

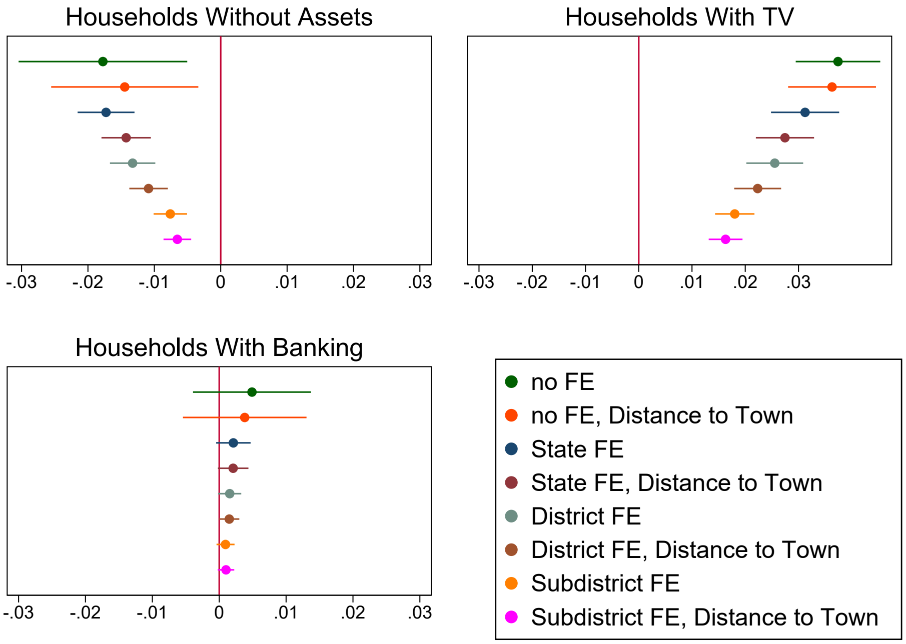
Standard errors in parentheses
Dependent variable: Night lights (shape, log of sum, 2011).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S11: Linear regressions for night lights and percentage of households with a TV. Standard errors clustered by state.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Electrified HH (log nbr)	0.540***	0.529***	0.397***	0.361***	0.323***	0.295***	0.251***	0.237***
	(0.047)	(0.048)	(0.028)	(0.024)	(0.025)	(0.023)	(0.019)	(0.019)
Non-electrified HH (log nbr)	-0.095**	-0.098**	-0.013	0.017	0.024	0.050***	0.066***	0.084***
	(0.038)	(0.042)	(0.026)	(0.024)	(0.016)	(0.015)	(0.014)	(0.013)
HH with banking (%)	0.001	-0.000	-0.001	-0.000	-0.000	-0.000	-0.000	0.000
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
Distance town (log km)		-0.261***		-0.397***		-0.377***		-0.335***
		(0.071)		(0.039)		(0.033)		(0.036)
Fixed effects: state	No	No	Yes	Yes	No	No	No	No
Fixed effects: district	No	No	No	No	Yes	Yes	No	No
Fixed effects: subdistrict	No	No	No	No	No	No	Yes	Yes
R-squared	0.389	0.405	0.216	0.262	0.160	0.208	0.127	0.162
Observations	516769	516769	516769	516769	516769	516769	516769	516769

Standard errors in parentheses
Dependent variable: Night lights (shape, log of sum, 2011).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S12: Linear regressions for night lights and percentage of households with a bank account. Standard errors clustered by state.



Households Without Assets

Households With TV

Households With Banking

Figure S16: Coefficients and standard errors for models equivalent to those in Tables S10-S12, but without control for the logarithmized number of electrified and non-electrified households. Some models control for distance to nearest town.

S8 LASSO Regressions

- Figure S17 displays the predicted values from linear regressions with and without polynomials.
- Table S13 displays the coefficients from various LASSO models.
- Figure S18 displays the mean squared errors path on each of the fold for the cross-validation.
- Figure S19 displays predicted values from a LASSO model with $\alpha = 10^{-10}$.
- Figure S20 displays predicted values from a LASSO model with $\alpha = 10^{-3}$.

Motivated by the possibility of a non-linear relationship between the log number of electrified households in a village and the village nighttime lights (shape, log sum), we investigate the usefulness of adding various polynomials in the regression of night lights on log number of electrified households. To that end, we implement the least absolute shrinkage and selection operator (LASSO), a statistical method used for model selection and regularization.

LASSO penalizes the magnitude of the regressors' coefficient. This is a good idea because in a linear regression with many high order polynomials, coefficients tend to be very large and result in overfitting: large coefficients allow for capturing more of the variation in the data and results in a lower sum of squared errors. This leads to overfitting and graphically to a very wiggly line through the data. Intuitively, LASSO finds the regressors that explain well the variation in the data without requiring large coefficients. Figure S17 illustrates the problem of over-fitting when using many polynomials. This is also illustrated in the second column of table S13 where the coefficients are particularly large.

Technically, the idea is implemented by adding to the objective function of the optimization problem a term that penalizes the magnitude of coefficients. In the usual linear regression, the optimization problem consists in minimizing the residual sum of squares (RSS). In a lasso regression, we minimize the following: $RSS + \alpha \sum_k |\beta_k|$, where β_k is the coefficient for regressor k and α is a parameter, fixed by the experimenter, that controls for the trade-off between minimizing the RSS and minimizing the sum of the absolute value of the coefficients. Adding such a penalizing term for the magnitude of coefficients is useful because when regressing many variables, the minimization of the RSS will tend to overfit the data. Intuitively, the larger the coefficients, the larger the overfitting and that is exactly what the $\alpha \sum_k |\beta_k|$ penalization term help avoid. This is called regularization. The second useful aspect of LASSO is that, not only will the coefficients be regularized (i.e. reduced), they will effectively be dragged to zero, and as a consequence LASSO performs model selection. This is extremely useful when researchers find themselves to choose among a lot of regressors.

In our case, we want to check for any non-linear relationship between the log number of electrified households and the village luminosity (shape, log sum). We use a polynomial of degree 15 in the luminosity variable: $\sum_{k=1}^{15} x^k$ where x is the luminosity. We also add lower order powers: $\sum_{k=1}^{9} x^{k/10}$. We estimate several LASSO models varying alpha from 10^{-10} to 10^{10} and we select the best model using k-fold cross-validation. We created 100 stratified folds in our data with each fold preserving the percentage of samples within the following categories: $\{y = 0\}$, $\{0 < y < 2\}$, $\{2 \leq y < 3\}$, $\{3 \leq y < 4\}$, $\{4 \leq y < 5\}$, $\{5 \leq y < 6\}$, $\{6 \leq y\}$ where y is the dependent variable (log number of electrified households in a village).

Figure S18 displays the path of the mean square error (MSE) on each of the 100 folds. We see that as we decrease the value of alpha the MSE keeps decreasing. The value of alpha for which the average MSE accross folds is the smallest is 10^{-10}. Although the MSE path looks flat on the left-hand side of the graph, it is decreasing slightly. Table S13 displays the magnitude of the coefficients for different models. We note that as alpha increases, most regressors are taken to zero: this is because the penalization term dominates. Figure S19 and S20 display predicted values from a LASSO model with alpha $= 10^{-10}$ and 10^{-3} respectively. As we compared these graphs to figure S17, we see that the smaller coefficients obtained through LASSO removed excessive variations.

Comparing the MSE from the simple univariate regression of x on y with the average MSE from the LASSO model with $\alpha = 10^{-10}$ shows that the non-linearity in the relationship is not very strong, and a simple linear description should be considered as a good approximation, while the LASSO model seems to be marginally useful.

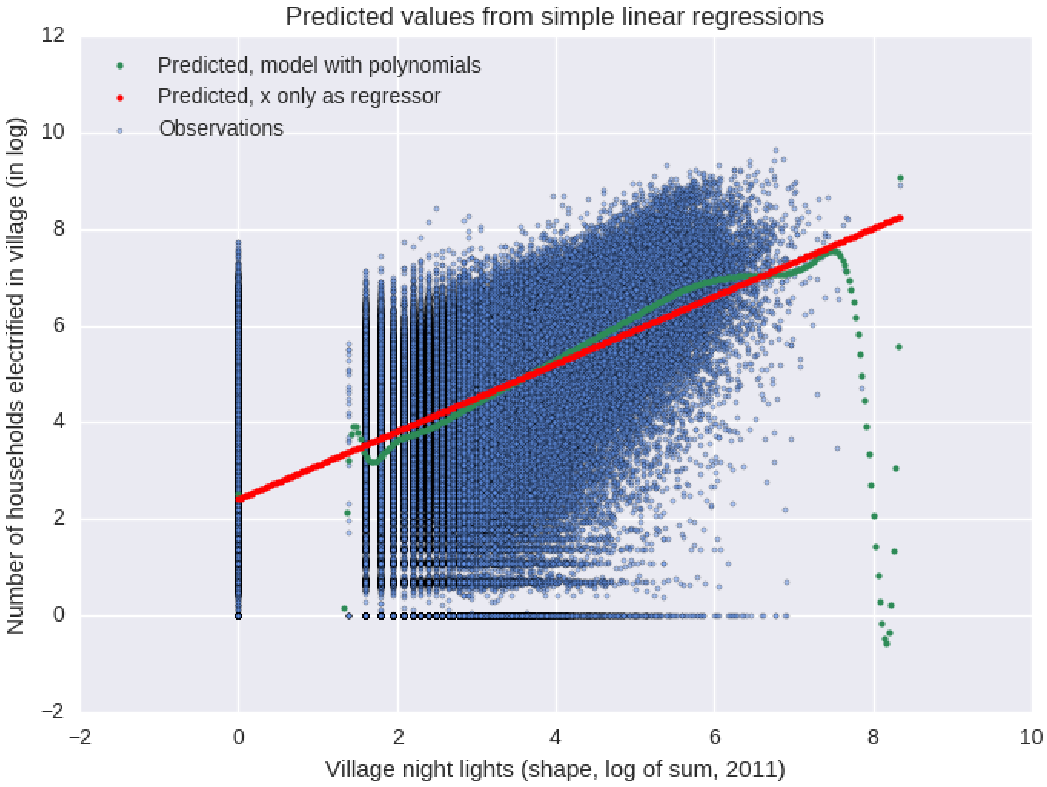


Figure S17: Predicted values from a linear model with many polynomials.

	$\alpha = 0$	$\alpha = 0$	$\alpha = 10^{-10}$	$\alpha = 10^{-3}$	$\alpha = 10^{+10}$
Regressors					
x	0.701	-3.3e+09	0.59	0.28	0
$x^{0.1}$	-	-3.5e+09	-0.75	-0	0
$x^{0.2}$	-	3.7e+09	0.079	-0	0
$x^{0.3}$	-	1.1e+10	0.069	0	0
$x^{0.4}$	-	-1.1e+10	0.06	0	0
$x^{0.5}$	-	-1.3e+10	0.051	0	0
$x^{0.6}$	-	1.1e+10	0.043	0	0
$x^{0.7}$	-	3.3e+09	0.035	0	0
$x^{0.8}$	-	-2.8e+08	0.029	0	0
$x^{0.9}$	-	1.7e+09	0.024	0	0
x^{2}	-	1.6e+08	0.03	0.13	0
x^{3}	-	-3.5e+07	0.0012	-0.00074	0
x^{4}	-	8.8e+06	2e-06	-0.00088	0
x^{5}	-	-2e+06	-3.3e-06	-9.2e-05	0
x^{6}	-	3.8e+05	-1.7e-07	-1.9e-06	0
x^{7}	-	-5.9e+04	-1.1e-07	7.7e-07	0
x^{8}	-	6.9e+03	-5.3e-08	1.3e-07	0
x^{9}	-	-5.6e+02	-1.4e-08	6.8e-09	0
x^{10}	-	24	-2.4e-09	-1.2e-09	0
x^{11}	-	0.8	-3.1e-10	-3.6e-10	0
x^{12}	-	-0.21	-2.4e-11	-5e-11	0
x^{13}	-	0.015	4.1e-13	-3.3e-12	0
x^{14}	-	-0.00055	5.8e-13	3.1e-13	0
x^{15}	-	8.6e-06	1.4e-13	1.5e-13	6.2e-13
MSE	2.4548	2.4264	2.4275	2.4286	3.9452

Table S13: Coefficients from LASSO regressions with different values of alpha. Cases where $\alpha = 0$ implies simple linear regressions. For the first two columns, MSE corresponds to the mean squared error of a simple regression. For the three last columns, MSE corresponds to the average MSE over the k folds created for the cross-validation.

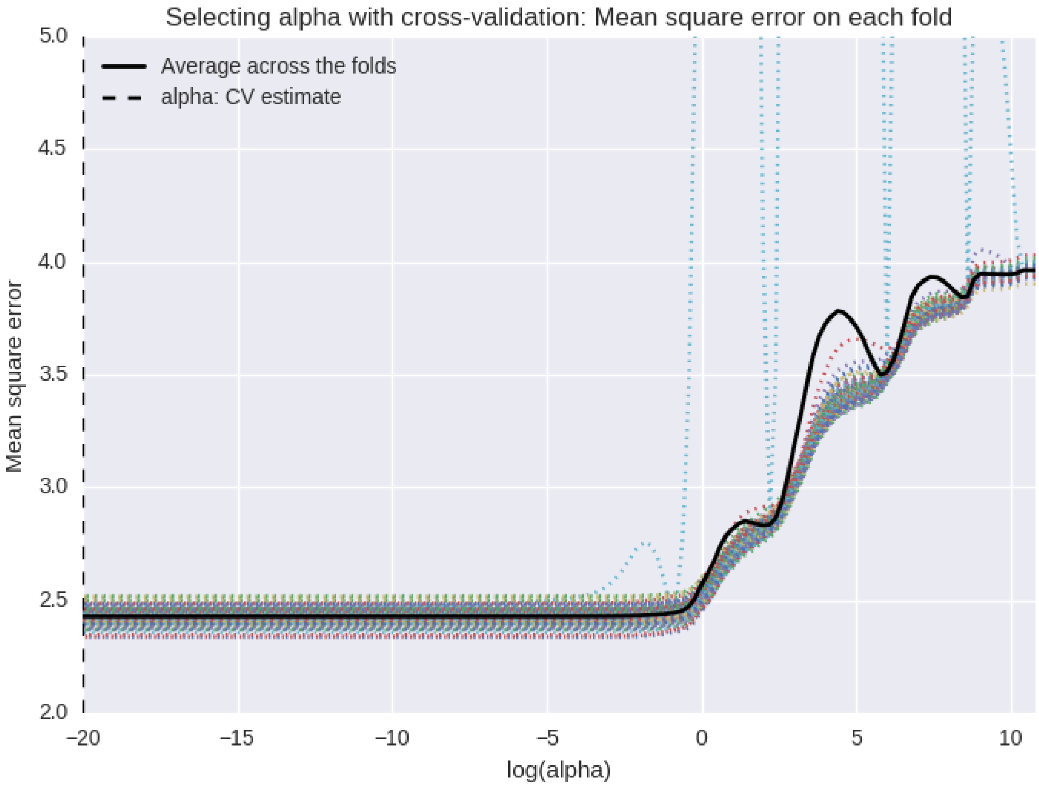


Figure S18: Mean squared errors on each fold of the 100 folds.

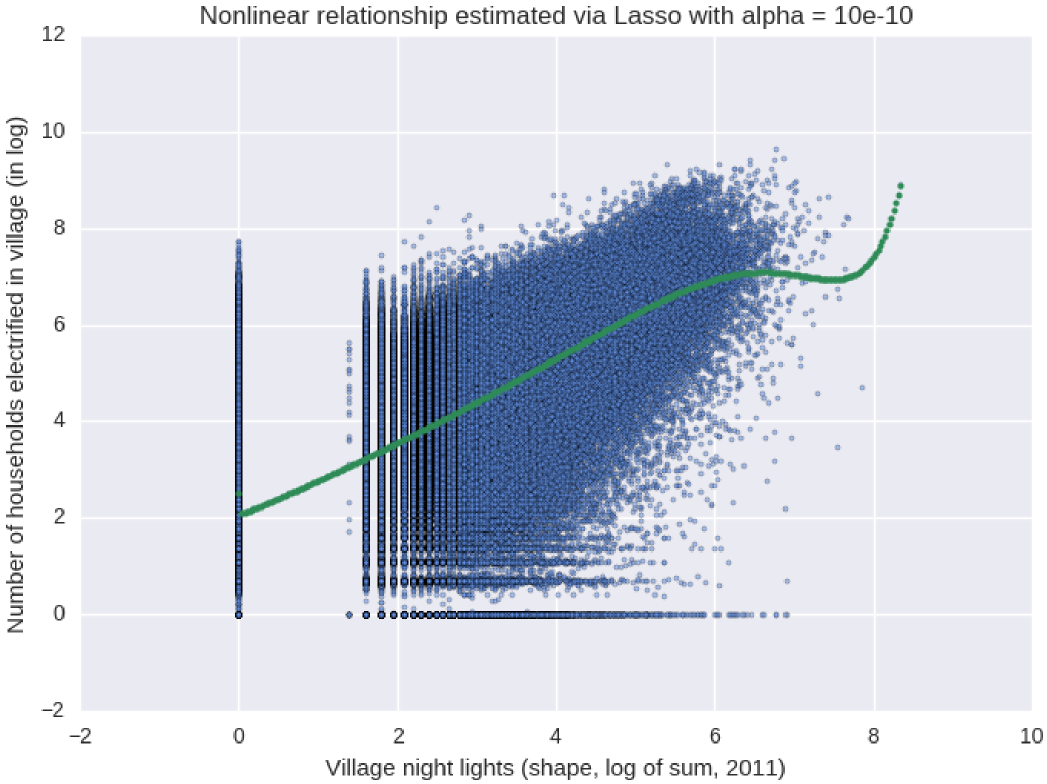


Figure S19: Predicted values from a LASSO model with alpha = 10^{-10}.

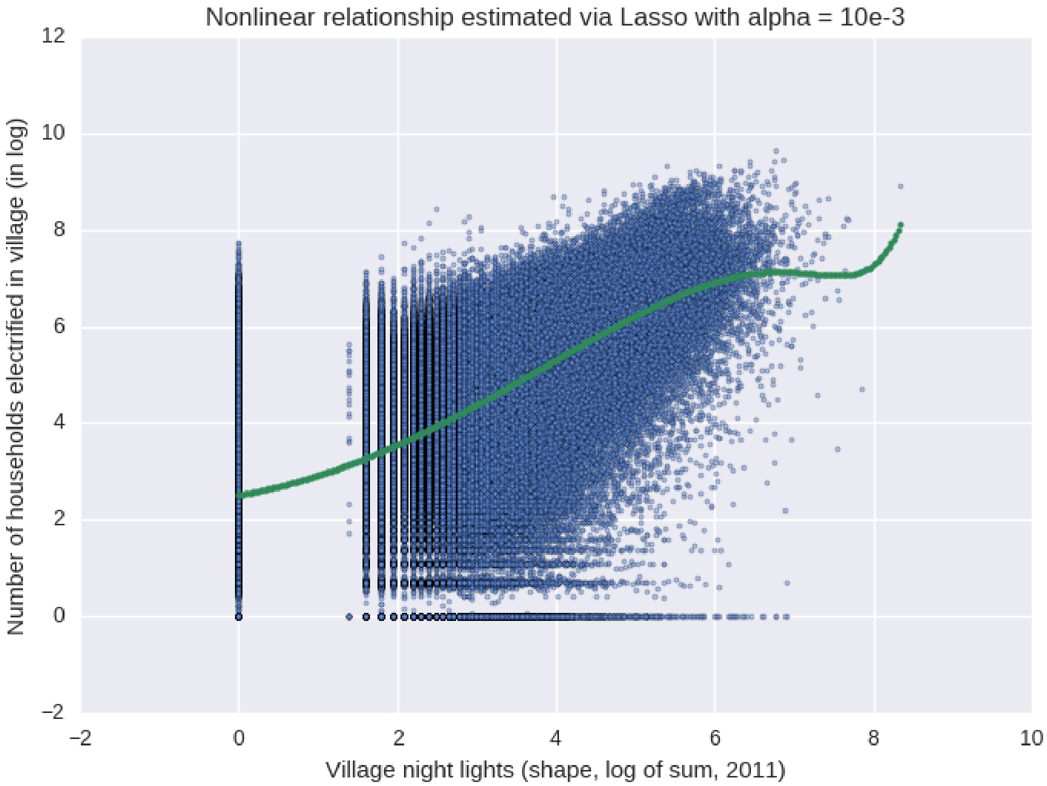


Figure S20: Predicted values from a LASSO model with alpha = 10^{-3}.

S9 Additional Box Plots

- Figure S21 plots the residuals across different levels of the DN measures used in the models above. All of these plot are generated from the direct relationship between the DN measures and the measure of electrification, without fixed effects. Overall, the results show that, for each level of the DV, the median error is centered around 0. With the exception of the lowest area of the DN measure, however, the shapefile generated measures tends to produce smaller IQR ranges for the errors, especially at the higher levels of the DN measure. This corresponds with what was observed in the descriptive analysis above – while the India Lights measures seem to do a good job of capturing the initial move from no electrification to some electrification, the measure does not do as well when we move to higher levels of village electrification.

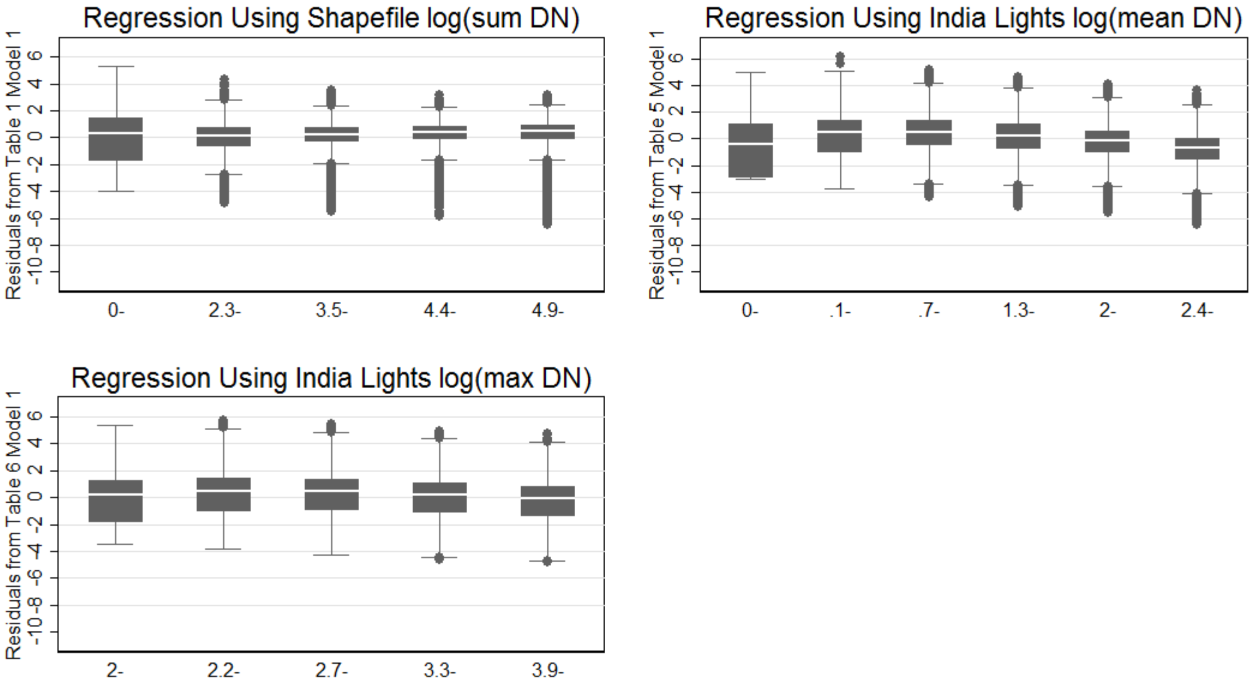


Figure S21: Boxplots of residuals from models using different measures of nighttime lights.

S10 Hours of Electricity

- In this section, we use data from the 2011 Census of India to better understand the relationship between nighttime lights and various measures of rural electrification. The measure of hours is the average value of summer (May-September) and winter (October-April) supply of power for domestic uses; in Karnataka, this variable is not available, so we instead use the supply of power for any uses. The scale is 0-24 hours.

- Figures S22 and S23 display hexabins plots of the number of electrified households (logarithmized) with the number of hours per day of grid electricity for households with no night lights (S22) and villages with high values of night lights (S23). The two hexabins contrast sharply. Among villages with no night light, those with medium to high levels of electrification have a small number of actual hourrs of electricity (around 5-10 hours a day). This might partially explain why some villages with medium to high electrification rates display no night lights: they might not be supplied with electricity in the evening and at night.

- Figure S24 is to be compared with Figure S25. In both figures, the x-axis is the same: night lights (shape, log sum, 2011). In Figure S25, the y-axis is the number of electrified households (in log) while in Figure S24, it is the number of electrified households (in log) weighted by the number of hours per day of grid electricity. We see that the two graphs are very similar, but Figure S24 shows a somewhat more precise relationship, with a more narrow spread at zero.

- Table S14 displays regressions of night lights (shape, log sum, 2011) with number of the electrified households interacted with hours of electricity per day. The results confirm that the number of hours of electricity correlates significantly with the intensity of the night light signal. indeed, the coefficient on the interaction 'Electrified HH (log nbr) X Hours per day of elec.' is positive with tight confidence bounds.

- Table S15 investigate to what extend the heterogeneity in hours per day of electricity contributes in the over or under prediction of the number of electrified households. In column

2, the coefficient for ‘Night lights X If elec < 5 h’ is negative implying that the number of electrified households is overpredicted for villages with only a few hours of electricity.

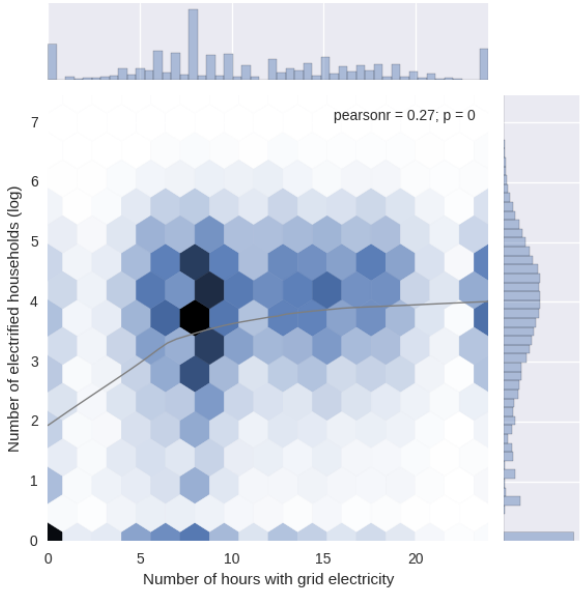


Figure S22: Hexabin plots of the number of electrified households (in log) with the number of hours per day of grid electricity for villages with a null value of log sum of the 2011 shape file measure.

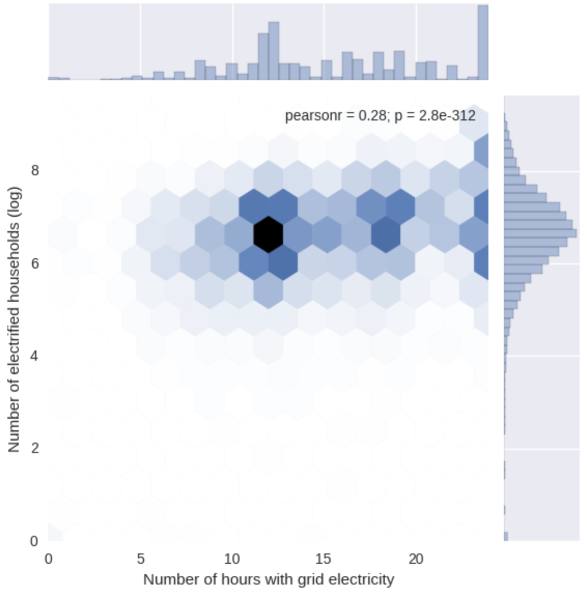


Figure S23: Hexabin plots of the number of electrified households (in log) with the number of hours per day of grid electricity for villages with a value of log sum of the 2011 shape file measure between 5 and 8.

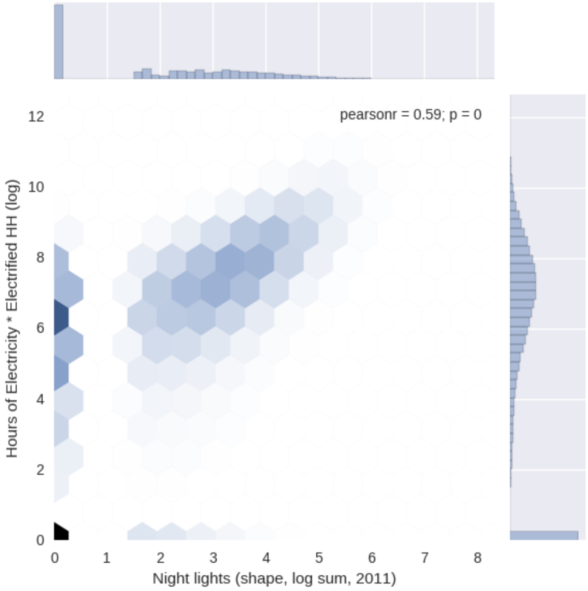


Figure S24: Hexabin plots of Hours of Electricity * Electrified HH (log) with the log sum of the 2011 shape file measure.

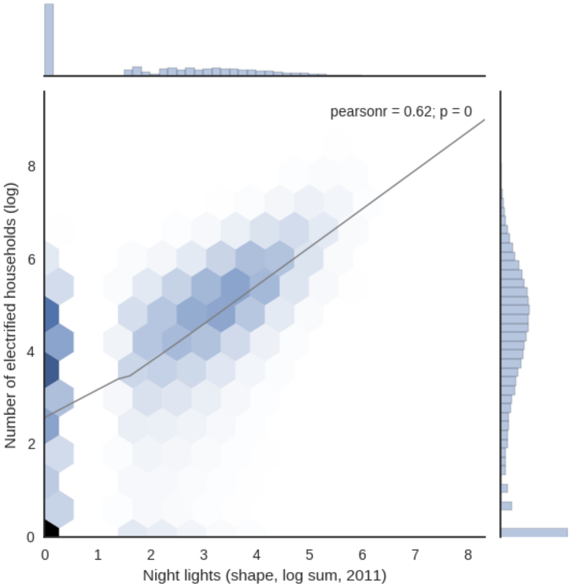


Figure S25: Hexabin plot of the number of electrified households (in log) with the log sum of the 2011 shape file measure.

	(1)	(2)	(3)	(4)	(5)	(6)
Electrified HH (log nbr)	0.543***		0.485***	0.337***	0.366***	
	(0.046)		(0.061)	(0.079)	(0.081)	
Hours per day of electricity		0.106***	0.040**	-0.020*		
		(0.018)	(0.016)	(0.011)		
Electrified HH (log nbr) X Hours per day of elec.				0.016***	0.012***	0.026***
				(0.004)	(0.004)	(0.002)
R-squared	0.381	0.153	0.390	0.404	0.403	0.330
Observations	516769	495466	492908	492908	492908	492908

Standard errors in parentheses
Dependent variable: Night lights (shape, log sum, 2011).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S14: Linear regressions, with standard errors clustered by state.

	(1)	(2)	(3)	(4)	(5)
Night lights (shape, log sum, 2011)	0.812***	0.859***	0.878***	0.875***	0.858***
	(0.046)	(0.060)	(0.061)	(0.063)	(0.057)
Hours per day of electricity	0.052***				
	(0.011)				
Night lights X If elec < 5 h		-0.400***			
		(0.044)			
Night lights X If elec between 5 and 10 h			-0.075**		
			(0.031)		
Night lights X If elec between 10 and 15 h				0.005	
				(0.018)	
Night lights X If elec between 15 and 20 h					0.059**
					(0.028)
R-squared	0.365	0.351	0.330	0.326	0.329
Observations	320015	330521	330521	330521	330521

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
Sub0sample of villages with night lights striclty positive
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S15: Linear regressions, with standard errors clustered by state.

S11 Summary Statistics: ACCESS Survey Data

- Figures S26-S29 display scatter plots of the relationship between night lights ($x - axis$) and other variables (y-axis).

- Table S16 summarizes the variables used in the analysis of the ACCESS survey. The unit of analysis is a village. The variables Night lights and Electrified HH are from the satellite data and the 2011 Census of India, respectively.

- The variable Average HH Expenditure (INR) is the average self-reported monthly expenditure in Indian rupees from the surveys of households within that village (12 per village).

- The variable Street Lights (nbr) is from the village module of the survey. It reports the number of street lights in the village, as reported by the (formal or informal) village leader.

- The variables Day Electricity Hours / Night Electricity Hours are the average self-reported hours of access to electricity supply by households during day (between sunrise and sunset) and at night (other times). We multiple these numbers of hours by the number of electrified households to arrive at the total supply of electricity to the village during day and night.

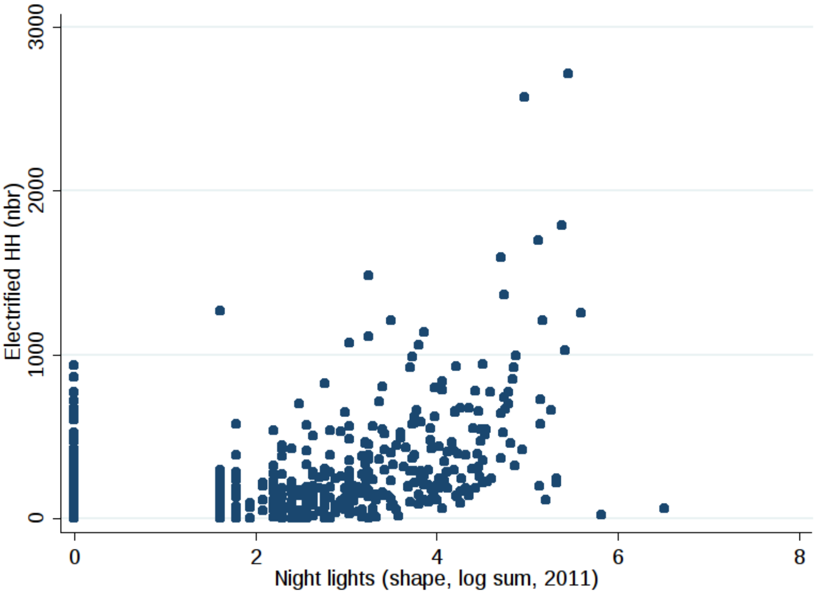


Figure S26: Scatter plot of the number of electrified households (in log) against the log sum of the 2011 shape file measure.

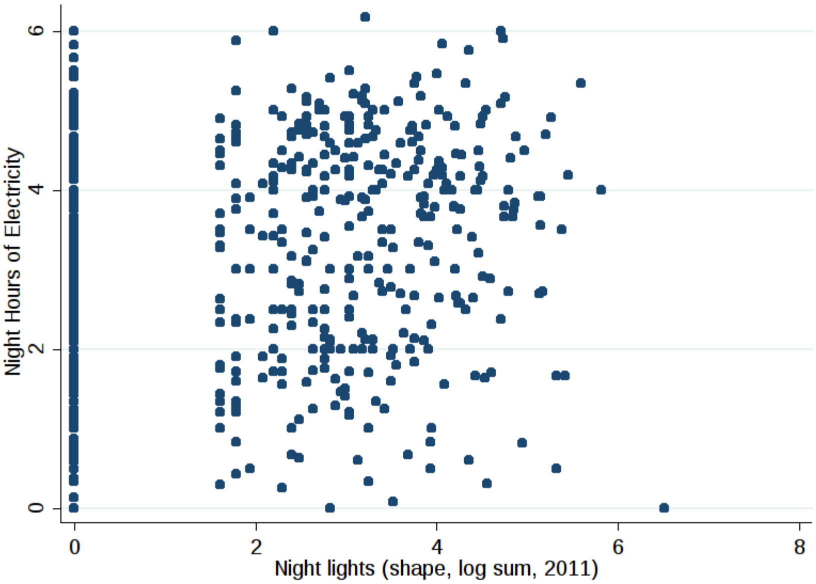


Figure S27: Scatter plot of the number of hours of electricity at night against the log sum of the 2011 shape file measure.

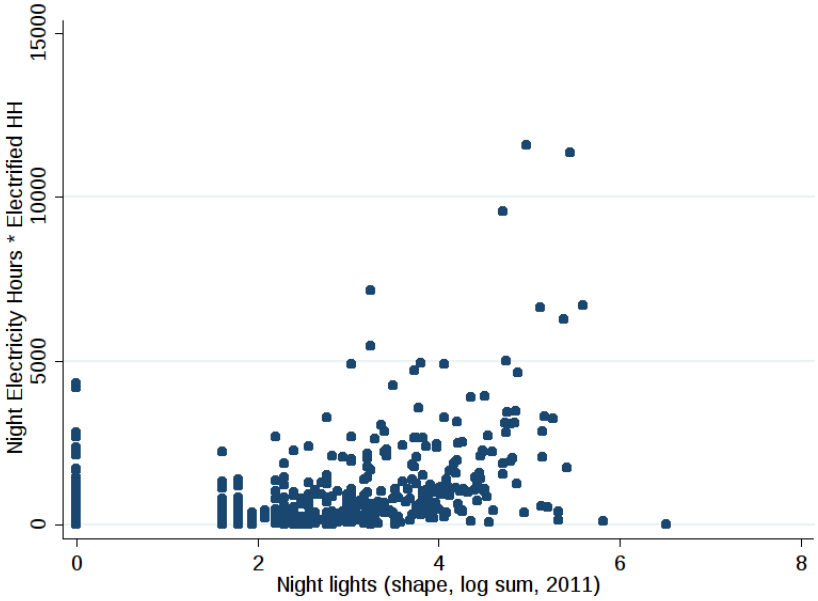


Figure S28: Scatter plot of the number of hours of electricity at night multiplied by the number of households against the log sum of the 2011 shape file measure.

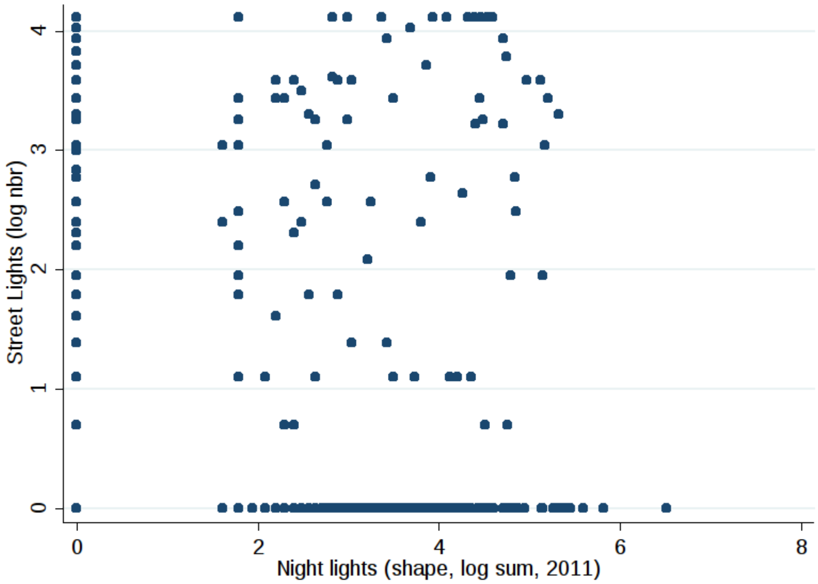


Figure S29: Scatter plot of the number of street lights (in log) against the log sum of the 2011 shape file measure.

	count	mean	sd	min	max
Night lights (shape, log sum, 2011)	681	1.69	1.74	0.00	6.51
Electrified HH (nbr)	713	198.80	287.23	0.00	2715.19
Average HH Expenditure (INR)	713	5301.58	1635.50	2125.00	11333.33
Street Lights (nbr)	712	4.45	12.45	0.00	60.00
Day Electricity Hours * Electrified HH	703	1941.30	3476.32	0.00	37765.80
Night Electricity Hours * Electrified HH	703	697.27	1206.55	0.00	11584.27
Night lights (shape, log sum, 2011)	681	1.69	1.74	0.00	6.51
Electrified HH (log nbr)	713	4.14	1.95	0.00	7.91
Average HH Expenditure (log INR)	713	8.53	0.30	7.66	9.34
Street Lights (log nbr)	712	0.56	1.17	0.00	4.11
Day Electricity Hours * Electrified HH (log)	703	5.92	2.60	0.00	10.54
Night Hours of Electricity * Electrified HH (log)	703	5.04	2.32	0.00	9.36

Table S16: Summary statistics for ACCESS data.

S12 Regression Output: ACCESS Survey Data

- Table S17 reports results from linear regressions at the village level. The dependent variable is Night lights (shape, log sum, 2011) throughout. The only differences between the models are the inclusion of different explanatory variables, all logarithmized to arrive at a log-log specification. Models 1-3 include numbers of electrified or non-electrified households, street lights, and expenditure separately; model 4 includes all of them. Models 5-7 include day/night hours either separately or together.

- The estimation results from models 1-4 reveal a few important patterns. First, even controlling for average household expenditure, the number of electrified households remains a strong and robust predictor of night lights. Second, average household expenditure is a strong and robust predictor of night lights, but the coefficient is approximately halved when we control for the number of electrified households. Third, non-electrified households do not predict night lights. Finally, the number of streetlights does not predict night lights.

- The estimation results from models 5-7 show that day electricity hours are a worse predictor than night electricity hours, as one would expect. Although both are strongly correlated in bivariate regressions, when they are included together, the night hours of electricity variable has almost twice the coefficient of the day hours of electricity variable.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Electrified HH (log nbr)	0.469***			0.460***			
	(0.029)			(0.029)			
Non-electrified HH (log nbr)	0.033			0.025			
	(0.039)			(0.040)			
Street Lights (log nbr)		0.100*		-0.038			
		(0.057)		(0.049)			
Average HH Expenditure (log INR)			0.890***	0.478**			
			(0.216)	(0.190)			
Day Electricity Hours * Electrified HH (log)					0.332***		0.134**
					(0.022)		(0.059)
Night Hours of Electricity * Electrified HH (log)						0.378***	0.240***
						(0.025)	(0.066)
R-squared	0.280	0.005	0.024	0.286	0.248	0.257	0.263
Observations	681	680	681	680	672	672	672

Standard errors in parentheses
Dependent variable: Night lights (shape, log, sum, 2011).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S17: Linear regressions with village-level data from the ACCESS survey.

S13 Distance to Cities

- We investigate whether light from big cities possibly spills over to nearby cities leading to an overprediction of the number of electrified households for villages within short distance to a major city, defined as a locality with a population greater than 100,000. In Table S18 we restrict the sample to the observations with non null values of night lights because we ask whether, conditional on seeing lights, the intensity of this light may be the results of a nearby major city as opposed to the village itself. The coefficient on night lights for cities within 15 km is negative and significant at the 99% level, indicating that using only night lights to predict electrification would lead to overprediction for these villages. The coefficient is tiny, however, suggesting that the bias is small.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Night lights (shape, log sum, 2011)	0.877***	0.877***	0.879***	0.878***	0.878***	0.877***	0.864***	0.701***	0.881***
	(0.065)	(0.064)	(0.065)	(0.064)	(0.064)	(0.064)	(0.062)	(0.066)	(0.065)
Distance (km) to major city in log	0.001								
	(0.040)								
Lights X If distance < 5 km		-0.080**						-0.037	
		(0.033)						(0.035)	
Lights X If distance between 5 and 10 km			-0.046**					-0.015	
			(0.018)					(0.020)	
Lights X If distance between 10 and 15 km				-0.034***				-0.014	
				(0.012)				(0.016)	
Lights X If distance between 15 and 20 km					-0.021*			-0.008	
					(0.011)			(0.013)	
Lights X If distance between 20 and 30 km						0.001		0.010	
						(0.010)		(0.011)	
Lights X If distance > 30 km							0.018		
							(0.011)		
Lights X If distance < 15 km									-0.042***
									(0.013)
R-squared	0.326	0.326	0.326	0.326	0.326	0.326	0.326	0.381	0.326
Observations	330520	330521	330521	330521	330521	330521	330521	516769	330521

Standard errors in parentheses
Observations with strictly positive values of night lights
Dependent variable: Number of households with electricity (log).
No fixed effects. Sub-sample of villages with night lights strictly positive.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S18: Linear regressions with distance to major cities - null values of night lights excluded.

S14 Analysis of Raw Night Lights

The National Oceanic and Atmospheric Administration (NOAA, `https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`) provides three versions of the nighttime lights data for each satellite/year. For the main part of our paper, we follow the general convention in the literature and use the cleaned stable lights file, which removes ephemeral events such as fires [Lowe(2014)]. This raster file replaces the values for what is identified as background noise with a value of zero. However, given that much of the error observed in the main paper is found where villages have electrification, but the nighttime lights data assigns them values of zero, one may rightly wonder if the process of removing this background noise has caused more problems than it has solved.

In this section, we reproduce our analysis using the raw version of the nighttime lights data, with the identified background noise still present. As expected, we find that this reduces the number of zero value observations in the dataset substantially. It does not, however, reduce the amount of error, nor does it change the general conclusions of the main paper.

- Figure S30 shows a hexabin plot of the logarithmized number of electrified households with the log sum of the 2011 shape raw file measure.
- Figure S31 shows a boxplot of the logarithmized number of electrified households against the logarithmized sum of the 2011 shape raw file night lights measure, by percentile.
- Figure S32 shows the correlations with the shape raw file measures and electrification.
- Table S19 shows regression results for raw night lights, with standard errors clustered by state.
- Table S20 shows regressions results for raw night lights with quantiles. The baseline category is night lights below the 20th percentile.
- Table S21 shows regressions for raw night lights with polynomials, with standard errors clustered by state.

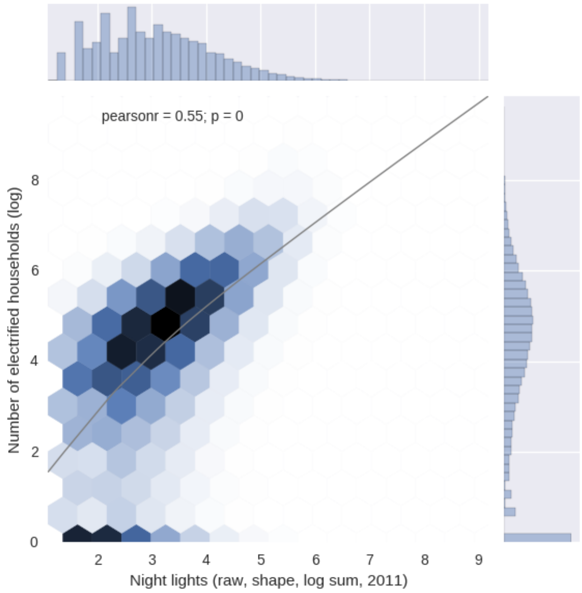


Figure S30: Hexabin plot of the logarithmized number of electrified households against the logarithmized sum of the 2011 shape file raw night lights measure. Dark colors indicate more observations in each hexabin.

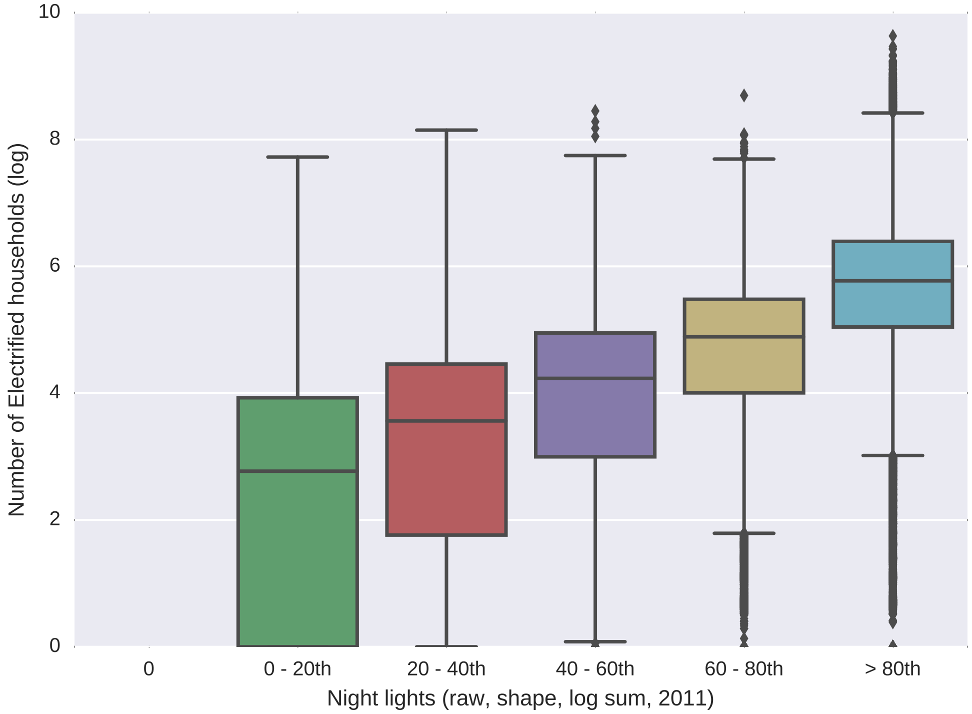


Figure S31: Boxplots of the logarithmized number of electrified households against the logarithmized sum of the 2011 shape raw file night lights measure, by percentile.

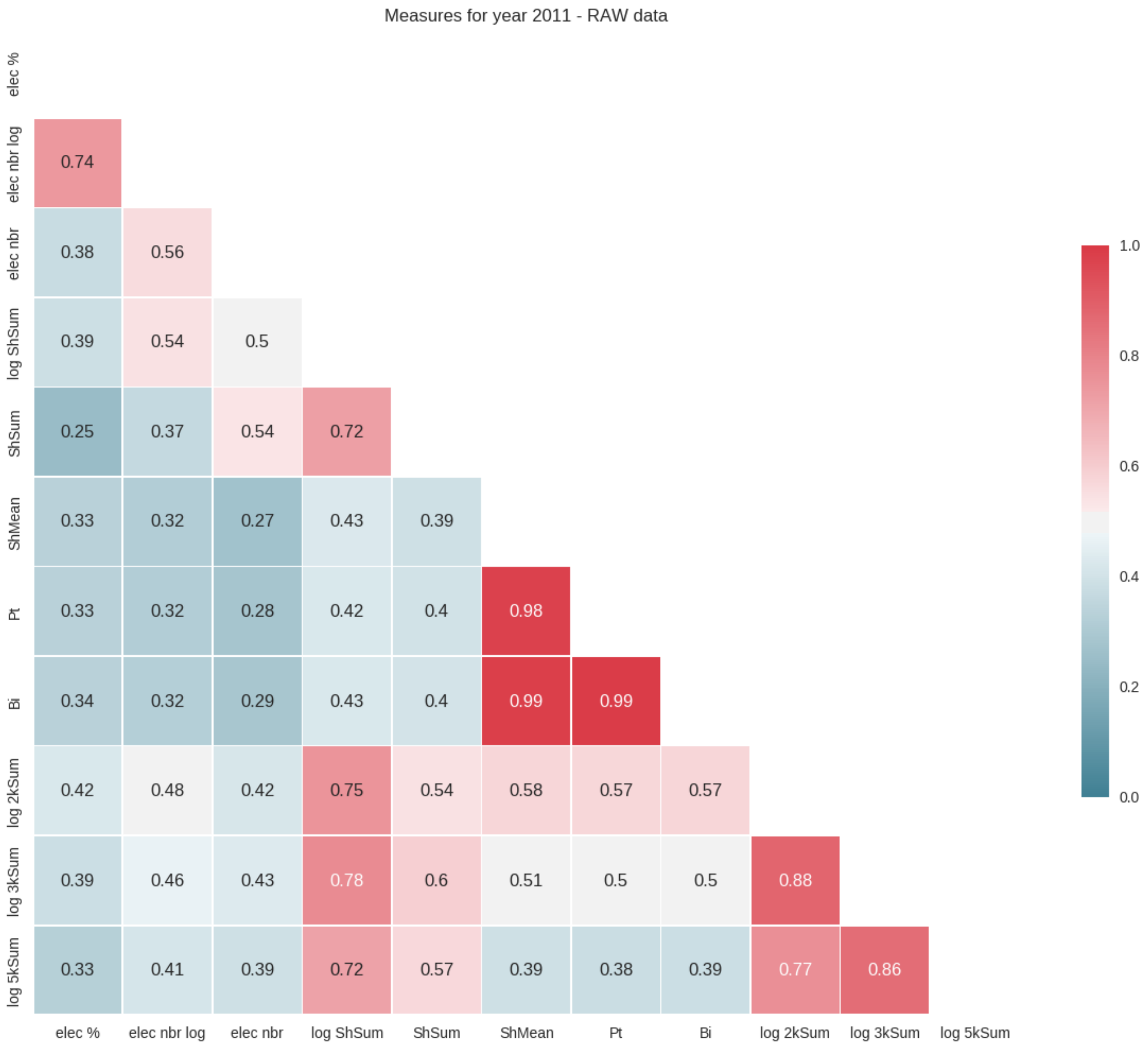


Figure S32: Correlation map for the shape raw file measures and electrification. Variable labels correspond to what follows. Note that all variables are for the year 2011.‘elec %’ = percentage of households that are electrified in the village; ‘elec nbr’ and ‘elec nbr log’ = number and logarithmized number of households that are electrified in the village, respectively; ‘ShSum’ and ‘log ShSum’ = sum and log of the sum of the luminosity of the pixels within the shape boundaries of the village; ‘ShMean’ = mean luminosity of pixels inside the shape boundaries of the village; ‘Pt’ = luminosity of the pixel at the longitude and latitude of the village centroid; ‘Bi’ = luminosity of the pixel at the longitude and latitude of the village centroid using linear interpolation of the surrounding pixels; ‘log 2kSum’, ‘log 3kSum’, and ‘log 5kSum’ = log of the sum of the luminosity of all the pixels within a 2-km, 3-km, and 5-km circle, respectively, centered at the village centroid.

	(1)	(2)	(3)	(4)
Night lights (raw, shape, log sum, 2011)	1.080***	0.779***	0.761***	0.724***
	(0.085)	(0.031)	(0.029)	(0.028)
Fixed effects: state	No	Yes	No	No
Fixed effects: district	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes
R-squared	0.304	0.154	0.145	0.131
Observations	516818	516818	516818	516818

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S19: Linear regressions with raw night lights.

	(1)	(2)	(3)	(4)
Night lights 20th-40th percentile	0.620***	0.488***	0.456***	0.410***
	(0.084)	(0.041)	(0.029)	(0.027)
Night lights 40th-60th percentile	1.282***	0.958***	0.898***	0.821***
	(0.142)	(0.057)	(0.055)	(0.058)
Night lights 60th-80th percentile	2.021***	1.451***	1.364***	1.262***
	(0.194)	(0.069)	(0.069)	(0.075)
Night lights 80th-100th percentile	3.058***	2.156***	2.052***	1.913***
	(0.263)	(0.084)	(0.086)	(0.082)
Fixed effects: state	No	Yes	No	No
Fixed effects: district	No	No	Yes	No
Fixed effects: subdistrict	No	No	No	Yes
R-squared	0.282	0.138	0.126	0.111
Observations	516818	516818	516818	516818

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
Night lights omitted category: values are below the 20th percentile.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S20: Linear regressions with dummy variables, with standard errors clustered by state.

	(1)	(2)	(3)	(4)
Night lights (raw, shape, log sum, 2011)	1.080***	1.371***	0.877	4.072***
	(0.085)	(0.231)	(0.574)	(1.294)
Night lights 2 (raw, shape, log sum, 2011)		-0.044	0.108	-1.356**
		(0.031)	(0.152)	(0.524)
Night lights 3 (raw, shape, log sum, 2011)			-0.014	0.263***
			(0.013)	(0.090)
Night lights 4 (raw, shape, log sum, 2011)				-0.018***
				(0.006)
Fixed effects: state	No	No	No	No
Fixed effects: district	No	No	No	No
Fixed effects: subdistrict	No	No	No	No
R-squared	0.304	0.305	0.305	0.306
Observations	516818	516818	516818	516818

Standard errors in parentheses
Dependent variable: Number of households with electricity (log).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S21: Linear regressions with polynomials, with standard errors clustered by state.

Supplementary Appendix: References

[Aklin et al.(2016a)] Aklin, Michaël, Chao-yo Cheng, Karthik Ganesan, Abhishek Jain, Johannes Urpelainen, and Council on Energy, Environment and Water. 2016a. "Access to Clean Cooking Energy and Electricity: Survey of States in India (ACCESS)." Harvard Dataverse, V1. http://dx.doi.org/10.7910/DVN/0NV9LF.

[Aklin et al.(2016b)] Aklin, Michaël, Chao-yo Cheng, Johannes Urpelainen, Karthik Ganesan, and Abhishek Jain. 2016b. "Factors Affecting Household Satisfaction with Electricity Supply in Rural India." *Nature Energy* 1: 16170.

[Lowe(2014)] Lowe, Matt. 2014. "Night Lights and ArcGIS: A Brief Guide." `http://economics.mit.edu/files/8945`.