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 Political Science, Columbia University, 420 W 118th Street, 712 IAB, New York, NY 10027, 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). 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. 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 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. 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 vary 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 electri-

fied 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. 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.

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.

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.

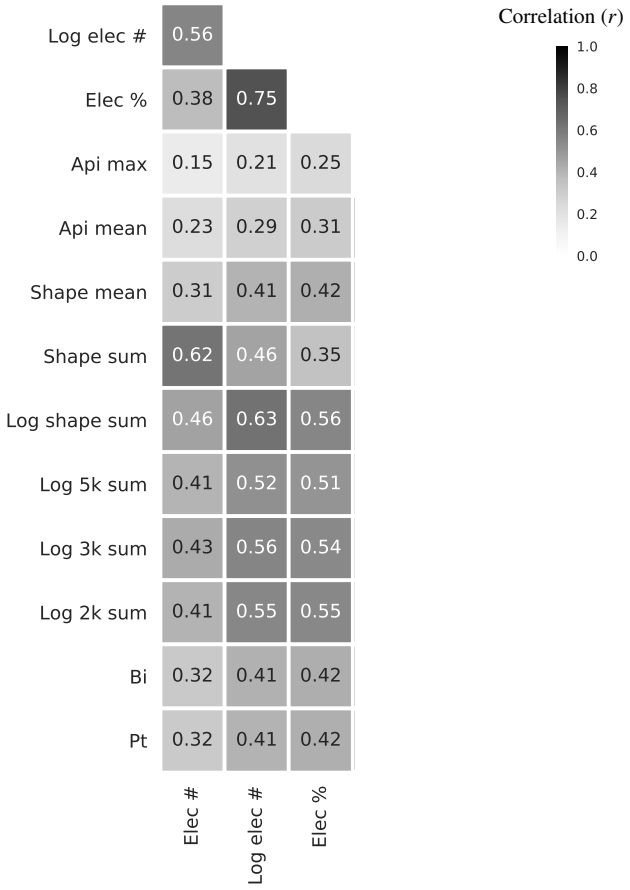

Figure 1: Pearson correlation coefficients (r) between electrification and luminosity variables.

Note: Data used is for the year 2011. Labels for the different variables are indicated below.

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.

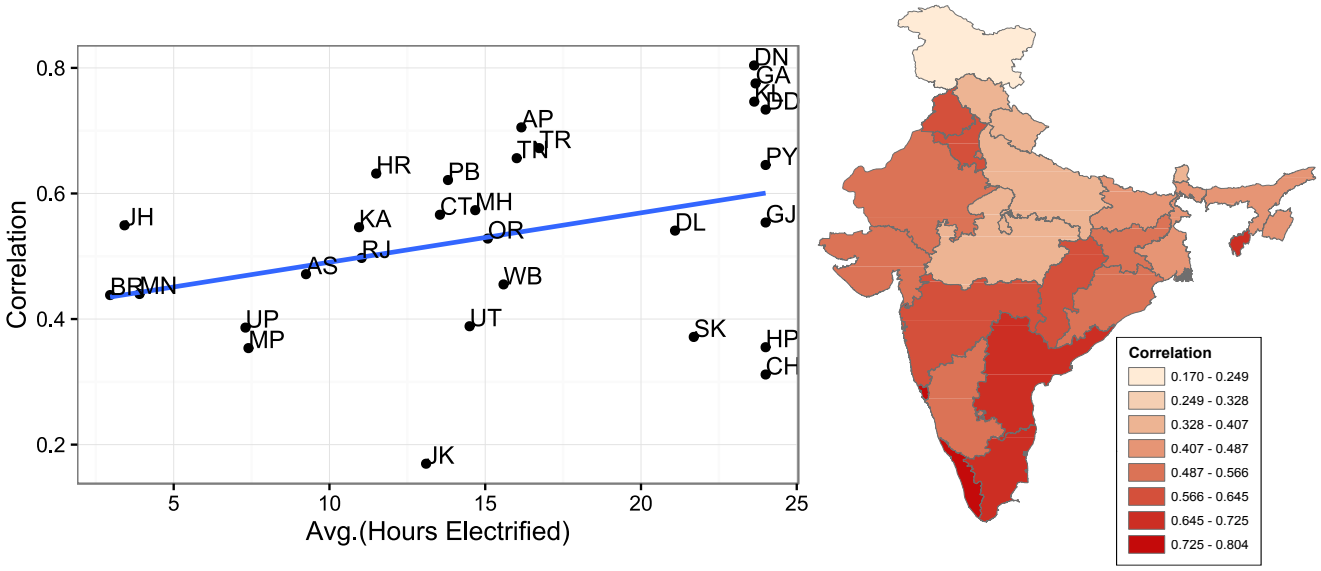

Figure 2: Correlation between the logarithmized number of electrified households and the logarithmized sum of the 2011 shape file night lights measure, state by state. The scatter plot on the left shows the correlation coefficients as a function of average hours of supply (Government of India, 2011); the map on the right places the correlation coefficients on a map of India.

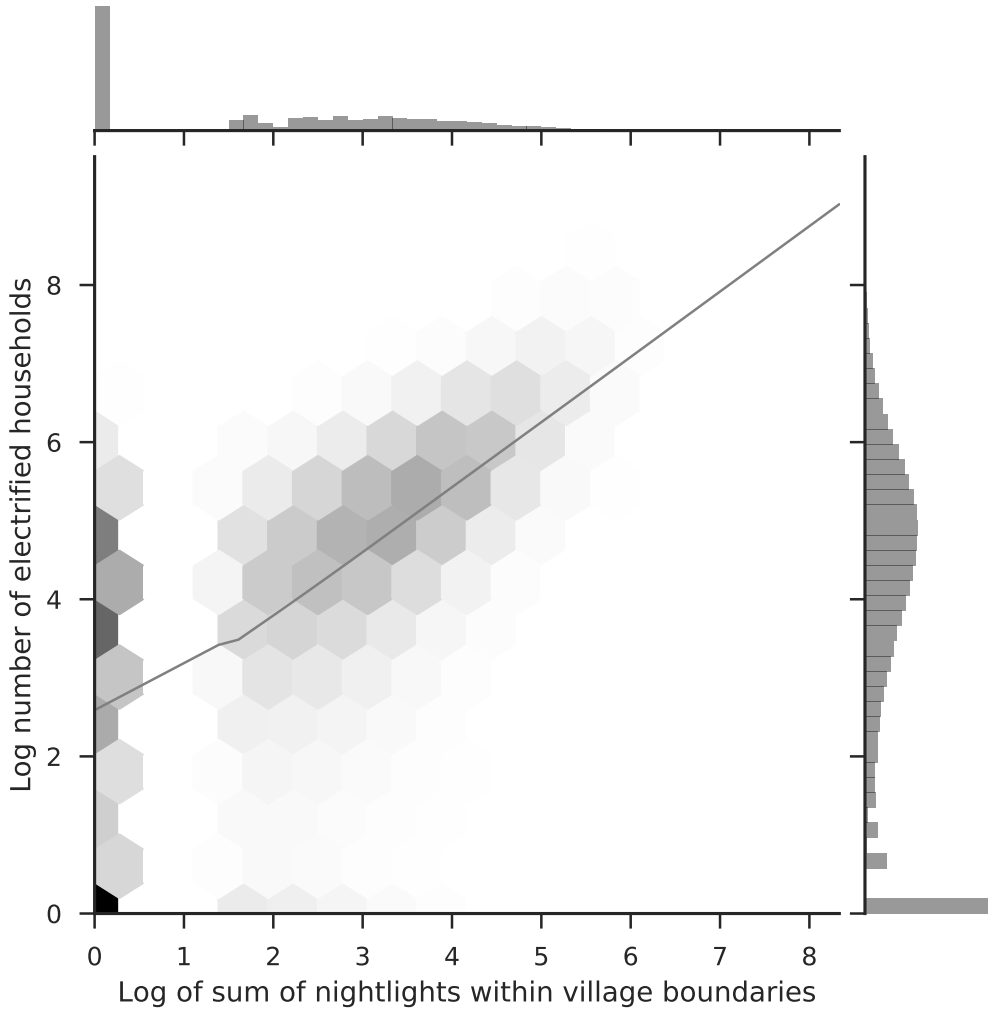

Figure 3: Hexabin plot of the log number of electrified households against the log sum of nightlights for 2011.

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.

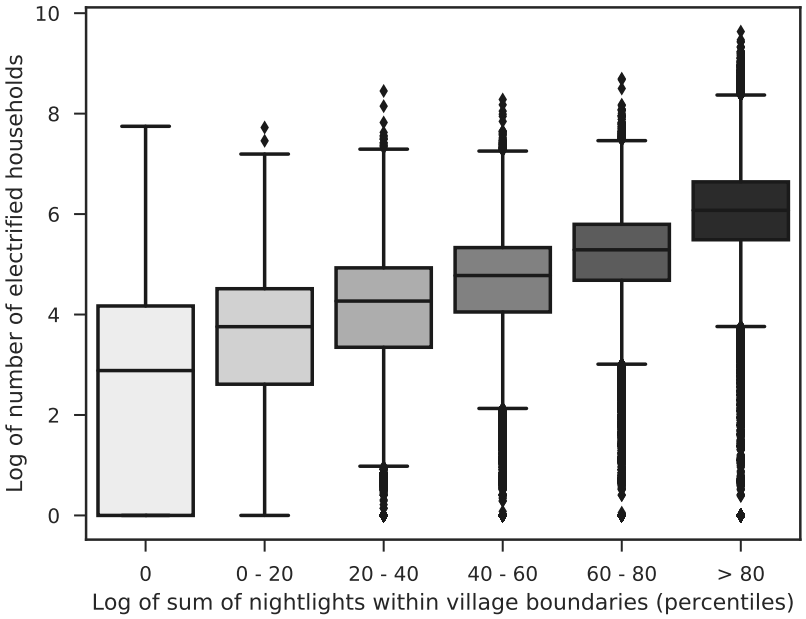

Figure 4: Boxplots of the logarithmized number of electrified households against the logarithmized sum of the 2011 shape file night lights measure, by percentile.

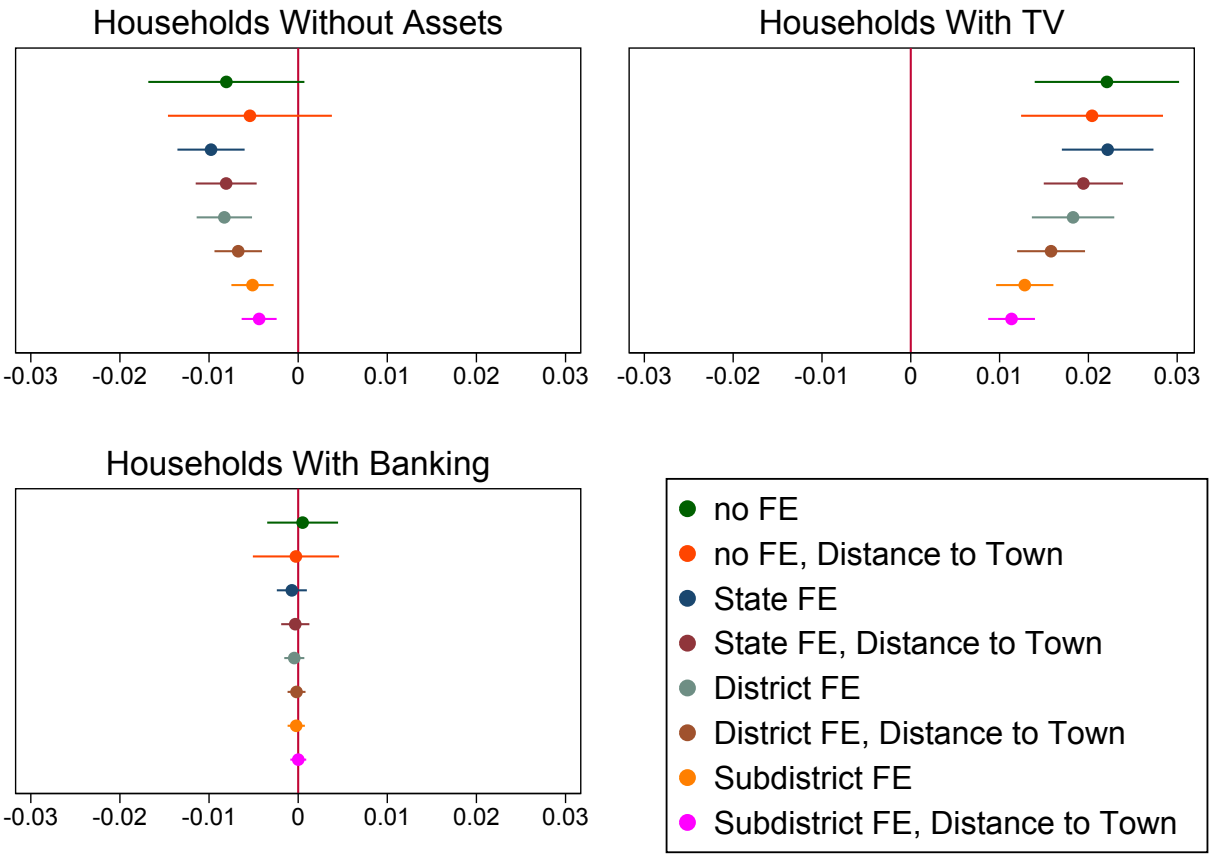

Figure 5: Coefficients and standard errors for models in Tables S9-S11. The dependent variable is the logarithmized nighttime lights. All models control for the logarithmized numbers of electrified and non-electrified households; some models also control for distance to nearest town. For the same coefficients withouot controls for the logarithmized numbers of electrified and non-electrified households, see SI Figure S16.

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

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.