Replication archive for "Out of the darkness and into the light? Development effects of rural electrification" (Burlig and Preonas, *Journal of Political Economy*).

This archive facilitates full replication of results in the paper and appendix, beginning with raw data and proceeding all the way through table/figure construction. Full replication of this code requires Stata and ArcGIS with the ArcPy library installed, and requires users to obtain access to proprietary data. As described below, we provide all datasets that are required to generate the final analysis and output, which requires Stata only.

To fully replicate the analysis in the paper, users should start with the provided datasets and code beginning with step A1 in MASTER.do.

Code organization

The *code* folder sequences all code as follows:

- *build*: files that begin with raw data and construct Stata datasets
- *merge*: files that merge/combine "built" datasets to create the datasets that we use in our empirical analysis
- *analyze*: files that execute all regressions, sensitivities, and benefit/cost analysis; and store all results
- *output*: files that construct the tables and figures that appear in the paper and appendix
- *Note*: In order to produce figures as in the paper, users must move "/code/data/analyze/scheme-fb.scheme" into their Stata "/ado/plus/s" folder.
- *Note*: In order to run this code, users must install *unique*, *reghdfe*, and *rdrobust* from SSC.

MASTER.do linearizes all code files within each step. All files are designed to run in the order listed (this is particularly important for "BUILD" and "MERGE" files). *MASTER.do* lists the inputs required for each step (including datasets, auxiliary code, and Stata packages), and the outputs produced by that step.

PATHS.do is called at the top of each .do file, and it defines common global macros that point to subfolders in the *data* folder.

We omit a series of web-scraping scripts in R and Python, which are upstream of the "BUILD" files. Much of this web-scraping is not replicable, since the websites providing the data have changed since we accessed them in 2014–2016. We include the fully scraped raw datasets in the *data* folder, unless otherwise indicated below.

Data organization

The *data* folder presents the folder structure required for full replication. However, since full replication would be cumbersome and time-consuming, we provide all datasets (in their finalized form) that serve as inputs to the "ANALYSIS FILES" and "OUTPUT FILES". This allows researchers to begin replication with step "A1" and proceed (linearly) through step "O7".

The *data* subfolders include many of the raw data files required to perform a full replication (beginning at step "B1.1"). However, we have omitted several large raw data files due to size and licensing issues. Below, we provide instructions for downloading each raw data file that is not included in the replication archive.

Running the "OUTPUT FILES" will populate the *outputs* folder with all tables and figures in the main text and appendix. In *outputs/figures*, we provide two figures in their final form:

- the map from Figure 2, which we coded up in Matlab and assembled in Photoshop (the dataset *data/rggvy data/rggvy_district_progress_X_XI_processed.dta* contains all of the information that assigns districts to Plans)
- the image from Appendix Figure C1, which is a screenshot of zoomed-in nighttime brightness data (manually produced using ArcGIS)

MLInfoMap (inputs to Steps B2, B6, and B8):

- These data are under license, and we are unfortunately unable to put the raw shapefiles into this data archive. A license for the 2001 Indian village boundary shapefiles can be purchased from http://www.mlinfomap.com.
- Users interested in replicating our full project should acquire a 2001 village boundary shapefile for each state (splitting Uttar Pradesh into pieces due to its size).
- As an alternative to MLInfoMap, it might be possible to use the open village maps offered by NASA-SEDAC at Columbia University (https://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001). However, we cannot guarantee that these maps will replicate our results.
- Our archive includes two .DTA files extracted from MLInfoMap village attribute tables, which are inputs into Step M3: "data/village gis maps/attrtables_lat_lon.dta" and "data/nightlights data/attrtables_all.dta".

Instructions for downloading and processing raw nightlights data (Step B2):

- Go to https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html, and download the 1998 –2013 files for both the "Average visible, stable lights, & cloud free coverage" and the "Average lights x Pct" categories. If data for multiple satellites are available for a given year (eg "F142000", "F152000"), download both.
- Extract these archives, and place the resulting .TIF files into "data/nightlights data/raw/" with the "avg_vis," "avg_vis_x_pct," and "stable_lights" files going into respective subfolders.
- Using ArcGIS, overlay each .TIF file with village boundary shapefiles from MLInfoMap (see description above), ensuring that the nightlights layer and the boundary layer are in the same projection. Extract the {min, max, mean, mode, median} values for each village, using ArcGIS's *Zonal Statistics as Table* operation. Save a separate .XLS dataset for each state-year-satellite. Repeat for villages containing only one pixel in the nightlights images (assigning the pixel's value for all summary stats), and save an "EXTRA" file for each state-year-satellite.
- Organize all .XLS datasets into the folders "data/nightlights data/[nightlights product]/[state]", as provided in the archive. These files serve as inputs to Step B2.

Raw SECC data (Step B4):

- Go to https://zenodo.org/records/10420499 and https://zenodo.org/records/10420771 to download these raw data. There are 6 separate zipped folders, totaling 58.9 GB.
- Unzip each of the 6 folders and place them in "data/secc data/secc_raw". Their contents are more zipped folders, each containing raw SECC data.
- Step B4 loops over the six (now unzipped) mega-folders, then loops over the zipped folders in each mega-folder automatically unzipping and processing them one-by-one.

Raw NSS data (Step B5):

- Go to https://zenodo.org/records/10416551 to download these raw data, and put them in "data/nss/raw".

Instructions for processing raw elevation data (Step B6):

- The raw elevation data are included in the folder "data/other spatial data/SRTM-elevation".
- Using ArcGIS, overlay the .TIF file with village boundary shapefiles from MLInfoMap

(see description above). Extract the {min, max, mean, mode, median} values for each village, using ArcGIS's *Zonal Statistics as Table* operation. Save a separate .XLS dataset for each state. Repeat for villages containing only one pixel (assigning the pixel's value for all summary stats), and save an "EXTRA" file for each state.

- Organize all .XLS datasets into the folder "data/other spatial data/SRTM-elevation/raw xls", as provided in the archive. These files serve as inputs to Step B6.

Raw Economic Census data (Step B7):

- Go to https://zenodo.org/records/10416621 to download these raw data, and put them in "data/economic census".
- Step B7 reads each wave's raw (text-formatted) data into Stata, using the dictionary files provided.

Instructions for constructing village distances to opposite RGGVY plan (Step B8):

- The inputs required are (i) MLInfoMap village boundaries (see above); (ii) MLInfoMap 2001 district boundaries (see above); and (iii) "data/rggvy data/ rggvy_district_progress_X_XI_processed.dta" (from Step B3).
- Using ArcGIS, overlay (i) and (ii) to assign a district to each village polygon. Next, join with (iii) at the district level, and create subsets of RGGVY 10th-Plan districts and RGGVY 11th-Plan districts. Then, for each group, extract the distance from each village centroid to the nearest boundary of a district in the opposite group. Save a separate .XLS dataset of distances for each state.
- Organize all .XLS datasets into the folder "data/village shapefiles/points_with_dd_dists", as provided in the archive. These files serve as inputs to Step B8.

Raw DISE schools data:

- The raw DISE data were scraped from individual school-x-year specific webpages, whose functionality we cannot guarantee. As a result, in the folder "data/schools data", we provide these data in "raw" Stata format: with all school-years appended, and variables named and labeled. We include only the relevant subset of variables.

JPE Dataverse Notes:

- This replication archive contains a large number of PLACEHOLDER.pdf files, in order to preserve our data structure
- To fully replicate all results from fully raw data, users will need to rename the subfolders currently called “/data/census data/Habitation Populations .2003.” and “/data/census data/Habitation Populations .2009.” to /data/census data/Habitation Populations (2003)” and “/data/census data/Habitation Populations (2009)”
- To fully replicate all results from fully raw data, users will need to rename the subfolders currently called “/data/census data/hpca11/excel scraped xlses/jammu . kashmir” to “/data/census data/hpca11/excel scraped xlses/jammu . kashmir”