BUILD AND REPLICATION INSTRUCTIONS

"Rural Roads and Local Economic Development"
American Economic Review

By Sam Asher and Paul Novosad

Final version: June 15, 2019

This archive contains the Stata .do files and all data files to reproduce our results. All files were produced with Stata MP version 14.2. To replicate our results:

1. Place all replication files in the same directory on your computer.
2. Run install_external_programs.do to ensure that you have installed all external Stata packages necessary for the replication.
3. Modify lines 7 and 8 in paper_results_aer_final.do to reflect the directories to which you downloaded the replication archive from the ICPSR repository and the SHRUG archive from the Harvard Dataverse (see below).
4. Run paper_results_aer_final.do, which calls all of the individual do files to generate the tables and figures.
5. Please note that while replication of results does not require programs other than Stata, generation of the tables uses the included stata-tex package, which uses Python. Thus full generation of Latex tables should be done on a Unix or Mac system with Python installed.

The following datasets are needed to replicate all major results:

- pmgsy_working_aer.dta: village-level data for all villages that were matched to PMGSY program data (N = 345,797). This contains all variables needed for replication.
- pmgsy_working_aer_mainsample.dta: a smaller version of pmgsy_working_aer.dta, which includes only the sample of villages in the main regression discontinuity specification (N = 11,432).

In addition, the following datasets were used to build the above analysis files. This is the set of all non-proprietary data files used in the paper:

- Data from the Socioeconomic High-resolution Rural Urban Geographic Dataset on India, Version 1.0 -- village/town-level data covering every village and town in India, combining data from the Population and Economic Censuses and using stable geographic units (Shrids). For the full dataset and documentation, see https://doi.org/10.7910/DVN/DPESAK.
    - shrug_pcec.dta: shred-level rural data from the Population and Economic Censuses, including all variables used in this analysis.
    - shrug_*_key.dta: the merge keys that match rural locations in shrug_pcec.dta to Economic Census and Population Census

identifiers using the shrid, a unique stable identifier in the SHRUG database.

- The following data files are included in this archive:
  - pmgsy_2015.dta: data on rural roads built under the PMGSY between the program launch and when we scraped the website in January 2015.
  - comp_roads_DVW2.dta: road-level data on all completed roads by January 2015.
  - evi_pc01.dta: estimates of kharif season agricultural productivity using EVI
  - ndvi_pc01.dta: estimates of kharif season agricultural productivity using NDVI
  - pc11_vd_ag_comm_key.dta: major agricultural commodities by village
  - pc11_hpca_village_pc01.dta: 2011 Population Census houselisting data
  - ec13_collapse_village_pc01.dta: collapse of 2013 Economic Census at the village level, including by major activities
  - village_poly_nl_annual.dta: night lights data
  - village_cropsuit.dta: cereal crop potential production measure (low input usage) from the FAO Global Agro-Ecological Zones (GAEZ)
  - pc01districtkey.dta: district names and codes in 2001 Population Census
  - pc91pc01districtkey_01superdist.dta: 1991 and 2001 Population Census districts aggregated to consistent spatial units
  - ddp_master_nddp_price05_long.dta: district-level gross net domestic product in 2005 prices

The following data sources used in the paper are proprietary and thus are not in the all-village public use data. However, the analysis datasets include fields generated from these datasets. The analysis datasets have keys that can be used to link to these sources for researchers who have access to them, and the build code shows how these datasets can be linked. Please see the paper for more details on these datasets and how to gain access to them.

- IHDS-II
- NSS
- Coordinates for all villages (purchased by the Harvard University Library from ML Infomap)
- Socioeconomic and Caste Census (SECC, 2012)
- Below Poverty Line Census (BPL, 2002)

The following code was used to build and analyze the data in this paper:

- Replication files – those called by paper_results_aer_final.do that generate the tables and figures in the paper:
  - Table 1: balance.do

    - Table 2 and Figure 4: first_stage.do
    - Table 3 and Appendix Table 14: family_index.do
    - Table 4: transportation.do
    - Table 5: labor.do
    - Table 6: firms.do
    - Table 7: agriculture.do
    - Table 8: cons_master_table.do
    - Figure 1: roads_by_year.do
    - Figure 2: balance_fig.do
    - Figure 3: running_var.do
    - Figure 5: family_binscatter.do
    - Figure 6: percentile_plot.do
    - Appendix Table 1: ndvi_validation_table.do
    - Appendix Table 2: table_pc01_sumstats.do
    - Appendix Table 3: nss_manlab_sect_shares.do
    - Appendix Table 4: acre_share.do
    - Appendix Table 5: ag_land_gender.do
    - Appendix Table 6: ihds_impute_coefs.do
    - Appendix Table 7: cons_disaggregated.do
    - Appendix Table 8: cons_ed_occ_table.do
    - Appendix Table 9: placebo_index.do
    - Appendix Table 10: family_index_altspecs.do
    - Appendix Table 11: pop_age_gender.do
    - Appendix Table 12: unemp.do
    - Appendix Table 13: tsc.do
    - Appendix Figure 1: hl_pop_hist.do
    - tables/*.tex: templates for producing tables in paper
- Build files – all other supporting code that is used in the construction of the data and generation of outputs:
    - make_pmgsy.do: master build file, calls everything else
    - create_pmgsy_aer.do: generates main analysis files
    - bootstrap_table_data_prep.do: preps data for bootstrapped consumption estimates
    - gen_night_lights_wide.do: converts night lights to wide for merge with other village-level data
    - gen_pc11_crops.do: generates village-level crop choice measures
    - gen_spillover_data_indexes.do: generates Anderson indexes for spillover catchment areas
    - _gweightave2.ado: generates Anderson indexes
    - impute_consumption_expenditure_secc_rural.do: preps IHDS data for ELL consumption prediction
    - input_block_key.do: inputs block names
    - install_external_programs.do: installs necessary Stata packages using ssc

- label_vars.do: labels variables
- merge_new_scrape.do: merges PMGSY scraped data together
- name_clean_blocks.do: cleans block names
- name_clean.do: cleans village names
- pc01_hab_match_aer.do: matches PMGSY and 2001 Pop Census villages
- pmgsy_include.do: loads all programs used in paper
- prep_pmgsy_aer.do: preps merged data for final analysis
- process_pmgsy_data.do: processes scraped PMGSY data
- settings.do: loads settings for data generation and analysis
- lev.py: Python code for calculative Levenshtein distance between two strings (adapted to Indian languages)
- stata-tex/*: directory of programs used to generate custom tables

Notes:

- Any lines in the code that start with "//do …" refer to do files that are included with this release but run on proprietary data that is not in the public use file. Should anyone wish to replicate these results, the data may be obtained separately and placed in the appropriate directory. All keys necessary for linking these proprietary datasets are included in the public use files.
- All code for generating bootstrapped consumption estimates is included with this release, but it uses proprietary SECC microdata not in the public file. Non-bootstrapped estimates can be generated with the data included, and produce standard errors approximately 10% smaller than those in the paper.
- To convert Stata plots from .eps to .pdf, epstopdf is used and requires a full Latex installation.
- Non-worrisome errors:
  - Some regressions will run but give an error that certain fixed effects "identify no observations in the sample."
  - If you run the replication code on a Windows machine or one without Python installed, the code will run but produce an error saying that it could not create the .tex table.