The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG) Data Description and Codebook

v1.4 (samosa), December 2019

The Socioeconomic High-resolution Rural-Urban Geographic Data Platform for India (SHRUG) is a data platform that facilitates socioeconomic research on India. SHRUG provides a consistent set of location identifiers that can be linked to Indian government data spanning the period 1990 to the present. SHRUG is also an aggregation of administrative and remote sensing datasets that describe socioeconomic and environmental change in India over this period. This codebook describes the mechanics of using the SHRUG. For more information on construction and concepts behind the SHRUG, see Asher, Lunt, Matsuura and Novosad (2019).

Contents

1 Brief Description of Data in the SHRUG

First and foremost, the SHRUG is a set of consistent location keys that allow Indian government data to be synchronized across space from the period 1990 to the present. This makes it possible to study socioeconomic change at a much higher geographic resolution (in over 500,000 villages and 8000 towns) than most prior work. We also aggregate data to the level of the legislative constituency, allowing analysis of links between politics and economics. It is easy for researchers to link their data to the SHRUG, and our hope is that others will contribute to growing this resource.

SHRUG v1.1 currently contains data from and links to the following:

- The Population Censuses of 1991, 2001, 2011, which describe basic demographic characteristics (Primary Census Abstract) and local amenities (Town/Village Directories).
- The Economic Censuses of 1990, 1998, 2005, and 2013, which are full enumerations of all non-farm establishments, including informal firms, service sector firms, and publicly-owned firms.
- The Socioeconomic and Caste Census, an enumeration of assets and additional characteristics for all households in India.
- Administrative data from government programs, like the national rural road program PMGSY.
- Remote sensing data on night lights and forest cover.
- Legislative assembly election results, which are linked to constituency-level data on all of the above.
- Data from politician affidavits, including criminal charges, assets, liabilities, and other characteristics.

2 Location Identifiers in the SHRUG

2.1 Town and Village Identifiers

The backbone of the SHRUG is a set of keys that link all the Indian Population and Economic Censuses to each other from 1990 to 2013 at the smallest consistent geographic unit possible. These keys were developed by matching towns and villages across population censuses, with close attention to splits, merges, and other realignments. Prior to SHRUG, we are aware of no dataset that allows accurate linking across all these datasets at the level of the town and village.

Linking these multiple survey rounds has necessitated merging units at different levels of aggregation depending on how those units have changed. The unit of aggregation in the SHRUG is a SHRUG identifier, or a *shrid*. In many cases, no aggregation was required and shrids can be

matched to single towns and villages in all underlying datasets. However, when two units merge in any population or economic census period, we have merged these units in all periods to allow consistent analysis of the unit. Some of the largest units are Delhi and Chandigarh, for which we were not able to retain any aggregations below the entire metropolis because of changes in unit identifiers across the censuses. Mumbai is aggregated to the district level.

A unique characteristic of location identifiers in the SHRUG is that we treat villages and towns equivalently in cases where similar data is recorded for both. As a result, each SHRUG dataset contains both villages and towns. In contrast, the Indian Population and Economic Censuses use an arbitrary distinction between villages and towns that has no basis in governance, and results in several villages with population over 50,000, and several towns with only a few hundred people. We have included these Population Census classifications if users wish to limit their analysis to census villages and towns, but they should recognize that the distinction is somewhat arbitrary.

Each location in the SHRUG is characterized by a sector code, which takes the value 1 (town), 2 (village), 3 (town and village). These sector codes may change over time, as they reflect the data that was aggregated into that shrid. For example, consider two villages in 1991, where one out of the two was classified as a town in 2001, and both were merged into the same town in 2011. To allow consistent analysis over time, we collapse these data into a single shrid in all periods. The sector variables for this village will thus take the values, `pc91_sector == 2`; `pc01_sector == 3`; `pc11_sector == 1`.

Note that we have *not* attempted to aggregate village and town directory fields across villages and towns. Other fields do not make sense to aggregate, like rural population. For this reason, users should be very cautious when analyzing locations with time-variant sector codes. The example shrid in the paragraph above would show a large decline in rural population from 1991 to 2001, but this would reflect reclassification, not population loss. However, the main population field `pcYY_pca_tot_p` will accurately track total population across the entire sample.

2.2 Legislative Constituency Identifiers

SHRUG includes socioeconomic data aggregated to the level of the legislative constituency, using both boundaries before and after the redelimitation in 2007. Parliamentary constituency boundaries will be available in a future version of the SHRUG.

For each legislative constituency, we create a location identifer that is consistent for the entire period of the delimitation. These are labeled `ac07_id` for the 1976–2007 delimitation and `ac08_id` for the 2008–present delimitation. The identifier takes the form `SS-AAA`, where `SS` is the 2011 Population Census state code, and `AAA` is either (i) the last assembly number used internally by the Election Commission (1976–2007); or (ii) the first assembly number used by the Election Commission

after the 2007 delimitation.[1]

Unfortunately constituency identifiers have not been used consistently by the Election Commission of India (ECI), making it sometimes challenging to link constituencies over time. Our approach makes it easy to link a constituency in Jharkhand to the same constituency when it was part of Madhya Pradesh, but this causes some discrepancies between the constituency numbers used by the ECI in some years. We do not include the 20 constituencies in Uttar Pradesh which were reformed into the 70 constituencies of Uttarakhand in 2001 because we could not obtain a high quality map of the prior UP constituencies. However, the 70 Uttarakhand constituencies are included. We also do not include post-delimitation Jharkhand because our constituency map had errors in this state. A future version will correct this.

Similarly, the constituency SHRUG excludes constituencies that are smaller than the smallest identifiable unit in the combined population censuses. This effectively excludes all large cities from the constituency SHRUG (though they are included in the town and village SHRUG). Users of the constituency SHRUG should recognize that the sample has a strong rural bias; however, given the absence of disaggregated census data below the town level in India, this bias exists in most prior analysis of economic characteristics of political constituencies in India as well.

A challenge in assembling constituency level data is dealing with missing and unmatched observations across time. As noted in the SHRUG documentation, village and town match rates to the Economic Census range from 65% to 90%. Further, many Population Census fields are missing for some villages, especially in the village directories. Naively aggregating spatial units with missing data will result in undercounts of population and employment. We have therefore rescaled constituency population and employment counts using an imputation process described here.

We begin by matching complete 2011 PCA data to constituency boundaries. If a village or town is missing in the 2001 population census (because we were unable to match it), we assign a 2001 population by assuming that this location had the same population growth rate as the average location in its 2011 constituency. This gives us a complete set of 2001 populations, and we repeat the process to obtain 1991 populations. For remaining data (e.g. Economic Census, Village Directory, etc.) we use the same process, assuming that missing locations have the same number of employed persons per population as non-missing locations within the same constituency. If a field would require more than 20% of its population or employment to be imputed, we make that field as missing for that constituency. Please also note that share variables, such as the share of people in villages in the constituency connected to power or the road network, are constructed by a population-weighted average within each constituency.

This approach provides what we think is the most accurate measurement of constituency growth over time. Undoubtedly some error is added by the imputation process, but these errors are likely

[1]Note that Telangana locations are indexed with the state code of Andhra Pradesh, as the new state was formed after the most 2011 population census.

to be smaller in importance than the noise that would be caused by treating missing towns and villages at zeroes in some years and not in others.

Please note that constituency identifiers are extremely inconsistent across data sources; often some set of numeric identifiers have excellent overlap, while others within the same state do not. While the numeric codes can be useful for matching, the name matches should always be verified.

Kerala, Goa, Tripura and Sikkim are missing from the 2007 constituency SHRUG because our constituency maps for them were particularly low quality. It was particularly difficult to assign villages to constituencies in these two states because Kerala has very large villages and the other three have very small constituencies and misaligned shapefiles.

3 Linking the SHRUG to Additional Data

The Population and Economic Censuses (among other administrative data sources in India) contain much more potential data than we are able to include in the SHRUG. Some of the data that can be linked to SHRUG via the raw Population and Economic Censuses include:

- Disaggregated data about firms, including firm size, source of finance, and public ownership.
- Additional village characteristics, including post offices, health centers, train stations, and characteristics of agricultural production.
- Additional town characteristics, including district capitals, transportation, and electricity infrastructure.

To make it easy to link the SHRUG to the underlying data, we include keys that link shrids to each Economic and Population Census in a single step. These keys take the form `shrug_[ecpc]_[location]_key.dta`. For instance, `shrug_pc01_pca_key.dta` links to the 2001 PCA at the vilage level, and `shrug_ec98_subdistrict_key.dta` links to the 1998 Economic Census at the subdistrict level.

The keys are unique on Economic and Population Census identifiers but are not necessarily unique on shrids. Researchers wishing to match the SHRUG to multiple rounds of data will need to decide how to deal with these duplicates. We advise collapsing external data sources to the shrid level before merging to the core SHRUG.

Stata code to link SHRUG to additional data in the 1991 and 2001 PCAs would thus take the following form:

```
/* 1. open the SHRUG PCA */
use shrug_pc01_pca.dta, clear


/* 2. prefix shrug data so it does not duplicate */
```

```
ren * sh_*
ren sh_shrid shrid

/* 3. merge to the additional population census data */
merge 1:1 shrid using PCA2001.dta, keepusing(...)

/* 4. collapse PCA back to the shrid level, but don't recollapse SHRUG data */
collapse (sum) pc01_pca_* (firstnm) sh_*, by(shrid)

/* 5. reset names to original format */
ren sh_* *

/* 5. Go back to step 2 in order to merge to additional data */
ren * sh_*
ren sh_shrid shrid
merge 1:1 shrid using PCA1991.dta
[etc...]
```

The dataset `shrug_names.dta/csv` provides the name in the 2011 Population Census of the largest population unit in each shrid. The shrid itself embeds the state code and census code of the largest population unit in the format `YY-SS-CCCCCC`, where `YY` is the most recent census year where we were able to identify this location, `SS` is the state code in that census round, and `CCCCCC` is the town or village code in that census round. This code structure should make it possible to maintain these shrids as constant units as new Indian data are released, such as the 2021 Population Census.

The complete PCAs, Town and Village Directories, and Economic Censuses have been made openly available online by the Indian government. The SHRUG contains a small subset of fields available in these broader datasets, because creating data fields that are consistently described and aggregated across time requires careful attention and cleaning of the raw data. As a result, the SHRUG contains only a small subset of the number of fields available in the raw administrative datasets—those that we have cleaned.

The included keys make it easy to conduct analysis at the shrid level using any additional fields. As we and other research teams clean additional variables and make them consistent across time periods, we will add more variables to the core SHRUG. We caution users bringing in additional data to carefully examine the raw data for miscodings, missing values, outliers, and inconsistent definitions across years, which are common in the raw census data.

4 How to Cite the SHRUG

The SHRUG is above all else a data platform and set of consistent geographic identifiers to facilitate the sharing of data between different research teams. As a convenience, we have included many actual fields with socioeconomic data. But users of these data should cite the researchers who originally prepared, processed, and linked their data to the SHRUG. In other words, if you use many parts of the SHRUG in your analysis, you will need to cite many different research papers – please do not just cite the SHRUG paper, or the work these researchers have put in will go unrecognized.

For example, if you use the SHRUG to study the impact of some political variable on rural road construction, you should additionally cite Jensenius and Verniers (2017) for the political data and Asher and Novosad (2019) for the roads data. When you download the SHRUG data, the platform will helpfully generate for you a list of citations that correspond to the data that you selected for download. References for all the components of the SHRUG can be found at http://devdatalab.org/shrug/refs.html.

5 The SHRUG Open Data License

The terms of use for the SHRUG are based on the Open Data Commons Open Database License (ODbL), which requires derivative works to retain the same open source status as the original work. In practice, this means that if you use the SHRUG, you commit to sharing the non-proprietary data that you link to the SHRUG for the purposes of your research at the time that your research is accepted for publication.

For example, suppose a researcher aggregates administrative data on water quality measures for all villages in Karnataka, and conducts research linking these to village-level economic outcomes using the SHRUG. When that researcher's paper is accepted for publication, she should post the shrid-aggregated water quality data (plus a brief description) for public use.

To make it easier for users to find new SHRUG-linked datasets as they appear, we will post links to them on our web site. With the researchers' consent, all-India datasets that seem particularly general in use may be included in future SHRUG packages. As noted in Section 4, users of the SHRUG are required to cite the original provenance of the data; inclusion in the SHRUG will therefore result in more citations for the reseacher in question. We therefore view this data sharing license as win-win, though recognizing that it does take some additional work to post data in a format that is usable by the public.

We recognize that some data are proprietary and cannot be shared, and the license does not apply to these data sources. Nevertheless, in many cases, location-level aggregates can be shared even if the raw microdata cannot. It is our hope that the user agreement and citation structure of the SHRUG will motivate researchers to release as much data at the shrid level as is possible.

6 Limitations of the SHRUG

The Economic and Population Census components of the SHRUG are aggregations of data collected by the Indian Ministry of Statistics and Programme Implementation, and the Census of India. We have cleaned these data where possible, but errors in data and linking undoubtedly remain. We advise all researchers to run robustness checks with regard to outliers and otherwise unusual units when conducting analyses based on the SHRUG. If errors are found in the SHRUG which are not in the underlying data, please send a detailed error description to `shrug-feedback@devdatalab.com`, and we will do our best to correct them for future versions. If you have questions about using the SHRUG or ideas for improvement, please visit the Shrug-India subreddit (https://www.reddit.com/r/ShrugIndia/), which is dedicated to discussion of research and analysis using the SHRUG.

The SHRUG is intended to represent the best time series of local socioeconomic data available in India at this time. The strength of the SHRUG is its geographic specificity. The SHRUG is not intended to be used to generate aggregate national statistics. These will be biased downward, because whenever Economic Census locations could not be matched to the Population Census across all periods (due to missing or incomplete location data, for instance), we have excluded their locations from the SHRUG. This is particularly the case for the constituency-level SHRUG, which does not include large cities (see Section 2.2).

The SHRUG does not include geographic data in the form of polygons or shapefiles because we have not yet found a sufficiently accurate data source with open sharing privileges. We are continuing to investigate sources of geographic data and may include shapefiles in a future version of the SHRUG. Users interested in obtaining geocodes or polygons for SHRUG units are advised to examine the open village maps offered by NASA-SEDAC at Columbia University. These can be directly merged to the 2001 Population Census SHRUG keys in `shrug_pc01r_key.dta` and `shrug_pc01u_key.dta`. Our own aggregate data was based on 2011 village polygons which we believe are slightly more accurate but are not made available with an open data license.

7 Variable-by-Variable Description of Core SHRUG Fields

The SHRUG Metadata table (in a separate file) summarizes the variables in this release of the SHRUG, along with their source datasets, year of recording, and the operation used to aggregate them to the village and town level. This section describes additional information on how variables were recorded and aggregated for the SHRUG. Users seeking additional information can refer to

material published by the Indian Census and Ministry of Statistics and Programme Implementation.

Urban/Rural Sector

Variables: `pcYY_sector,ecYY_sector`

These variables describe whether a location's data was aggregated from urban data only (sector=1), rural data only (sector=2), or a combination of urban and rural data (sector=3). As noted above, this can be time-variant, because a location can consist of all villages in one period, but one of those villages can be treated as a town in a future period.

Caution is required when conducting time series analysis of locations that span urban and rural sectors, because these states change over time. The analysis of a village-level measure (like `pcYY_vd_p_sch`, the number of village primary schools) will not be consistent over time in units that combine towns and villages, because the village directory fields are empty in periods after the village has turned into the town. It is therefore necessary to aggregate town and village directory fields for these locations, for example by adding urban and rural primary school numbers. PCA and Economic Census fields have already been aggregated across towns and villages within shrids, so these are consistent in the time series.

Non-Farm Employment

Data file: `shrug_ec.dta`

Variable: `ecYY_emp_all`.

The SHRUG contains data from the 3rd through 6th rounds of the Economic Census, covering 1990, 1998, 2005, and 2013. The Economic Census is a complete enumeration of all non-crop producing economic establishments in India. This includes both public and private firms, firms in manufacturing and service sectors, and both formal and informal firms. The raw data is establishment-level, which have been aggregated to the village/town-level in the SHRUG.

Different rounds of the Economic Census used different inclusion criteria. For instance, the 2013 EC did not include firms in government administration or national defense. To create a consistent set of measures over time, we therefore excluded these industries (NIC2008 Section O) from all prior censuses. Given the apparent inconsistencies in which agricultural firms were included in different locations and different rounds, we also drop all agricultural firms. Variables from each Economic Census have the prefix `ecYY`, where `YY = {90, 98, 05, 13}`and corresponds to the year of data collection.

There are a small number of villages with unusually high or low employment numbers in some periods. Because the Economic Census provides little information about firms beyond their employment count and characteristics of that employment, it is difficult to determine whether these measures are valid or not. We advise testing robustness to excluding outliers as measured by employment over population or improbable swings in employment over population from one census to the next.

These four Economic Censuses used three different rounds of the National Industrial Classification, making it difficult to follow specific industries over time. As a result, SHRUG does not contain industry-specific employment at this time, but we are developing a time-invariant industry classification that will be available in future versions.

The Economic Census is produced by the Ministry of Statistics and Programme Implementation. Documentation for all rounds can be obtained at the official Economic Census website. Raw establishment-level data is now available for free at the National Data Archive and can be readily linked to the SHRUG, as described in Section 3. Additional fields available in the raw Economic Census include four-digit sector, source of finance, source of power, gender and social group of owner, and number of employees of each gender.

Population

Data files: `shrug_pcYY.dta`

Variables: `pcYY_pca_tot_p`: Total Population
`pcYY_pca_tot_p_r`: Rural Population
`pcYY_pca_tot_p_u`: Urban Population
`pcYY_pca_p_sc`: Scheduled Caste Population
`pcYY_pca_p_st`: Scheduled Tribe Population
`pcYY_pca_p_lit`: Literate Population
`pcYY_pca_no_hh`: Number of Households

The decennial Population Census was conducted eight times in British India and, as of 2011, seven times in independent India. The SHRUG contains data from the three rounds of the Population Census for which there is data available at the village/town-level: 1991, 2001, and 2011. These variables are prefixed with `pcYY`, where `YY = {91, 01, 11}`. Each round of the Population Census contains many modules. The SHRUG contains data from the three that are available at the village/town-level, which are the Primary Census Abstracts and the Village and Town directories.

The PCAs (prefixed with `pcYY_pca` in the data) contain the basic demographic structure of every town and village in India, which are aggregated from individual census respondents. Currently included variables are the total population of the location, number of literate people, and number of individuals coming from scheduled castes and scheduled tribes.

The Village and Town Directories (prefixed with `pcYY_vd` and `pcYY_td`) provide descriptions of village- and town-level characteristics and public goods, such as the availability of electricity (or number of hours in 2011), number of schools, and presence of a paved road. These measures are recorded based on discussion with a village leader.[2]

[2]For the official description of the variables in the original PCA and Village/Town Directories, please see the

In some cases, the Census shows inconsistencies between rural PCA population and village directory population, or between urban PCA and town directory. A future version of the SHRUG will combine these fields into a best estimate of each location's population; at this time we advise using the population from the PCA. Note that the town and village directory population fields will not be consistent in the time series for places whose sector classification changes. If a location is a village in 1991 and a town in 2001, then it will have non-missing village directory population in 1991 and missing village directory population in 2001 (but non-missing town directory population). However, the PCA population will be consistent across all periods.

Zero population values may indicate abandoned villages or data entry errors at the Census. We advise dropping from analysis all locations with zero population values in any census year.

Town and Village Schools

Data files: `shrug_pcYY.dta`

Variables:
`pcYY_vd_p_sch,pcYY_td_p_sch`: Number of village/town primary schools
`pcYY_vd_m_sch,pcYY_td_m_sch`: Number of village/town middle schools
`pcYY_vd_s_sch,pcYY_td_s_sch`: Number of village/town lower secondary schools
`pcYY_vd_s_s_sch,pcYY_td_s_s_sch`: Number of village/town senior secondary schools
`pcYY_vd_college,pcYY_td_college`: Number of village/town colleges

These variables describe the number of schools at various levels as reported in the town and village directories. To obtain the total number of schools in locations that combine both urban and rural status, it is necessary to add the urban and rural measures together.

In 2011, the amenities tables separately reports the number of public and private schools; we add these together to obtain a count of the number of schools that is consistent with prior years. Users interested in the disaggregated public/private data can link the SHRUG to the raw 2011 Population Census data. Alternately, the raw Economic Census includes data on schools and includes a variable for whether it is government or private. Note that the distribution of primary schools in the Rajasthan village directory in 2011 suggests severe data errors; these data are aggregated directly from the Indian government data and there may be additional errors.[3]

The `college` variable adds up all the post-secondary colleges list in the town and village directories, which includes colleges, engineering schools, law schools, medical schools, but does include vocational schools. More granular information on schools can be obtained by merging these data

2011 Population Census Instruction Manual and Concepts and Definitions, respectively.

[3]Specifically, Rajasthan reports zero primary schools in nearly all villages with 2011 population less than 500, and then one primary school in villages with population greater than 500. The majority of villages under 500 reported schools in 2001, so these data are likely to be erroneous.

with the raw village and town directories available on the Census web sites.

Note that village and town directory fields are not consistent descriptors of locations with changing sector classifications; see the descriptions of the pcXX_sector variables and Subsection 2.1 on town/village identifiers.

PMGSY and Other Rural Roads

Data files: `shrug_pc.dta, shrug_ancillary.dta`

Variables:
`pcYY_vd_tar_road`: Indicator for paved road (1=Yes)
`pcYY_vd_dirt_road`: Indicator for dirt road or better (1=Yes)
`road_award_date_new`: Date new PMGSY contract awarded
`road_award_date_upg`: Date upgraded PMGSY contract awarded
`road_comp_date_new`: Date new PMGSY road completed
`road_comp_date_upg`: Date upgraded PMGSY road completed
`road_comp_date_stip_new`: Stipulated completion date of new road
`road_comp_date_stip_upg`: Stipulated completion date of upgraded road
`road_sanc_year_new`: Year new PMGSY road was sanctioned
`road_sanc_year_upg`: Year upgraded PMGSY road was sanctioned
`road_length_new`: Length of new PMGSY road
`road_length_upg`: Length of upgraded PMGSY road
`road_cost_new`: Cost of new PMGSY road
`road_cost_upg`: Cost of upgraded PMGSY road
`road_cost_sanc_new`: Sanctioned cost of new PMGSY road
`road_cost_sanc_upg`: Sanctioned cost of upgraded PMGSY road
`road_cost_state_new`: State-paid cost of new PMGSY road
`road_cost_state_upg`: State-paid cost of upgraded PMGSY road

Indicators for paved and dirt roads are recorded in the village amenities table for the population census. These binary variables take the value 1 if any of the components of the location (shrid) were reported as have a paved road. The variables released by the Census to describe roads changed in 2011. A paved road in the SHRUG is defined as a pucca road in 1991 or 2001, and as a black-topped road in 2011. A dirt road in the SHRUG is defined as a kuchha road in 1991 and 2001, and as a gravel road in 2011. Users interested in alternate measures of road completion included in the 2011 Village Directory can readily merge the SHRUG to the original Census data.

Based on our cross-checking of the data, we believe that some states reported missing values in the 1991 and 2001 village directories to indicate the absence of roads (rather than zeroes). We inferred this from examining the correlation with other variables and reports of roads in other years.

These zeroes have been filled in for the SHRUG. The original data can be accessed by merging these data with the raw Population Census data.

PMGSY data comes from the Online Monitoring and Management System (hosted at `omms.nic.in` at the time of release), and was downloaded in 2016. More recent data are now available but have not been integrated into SHRUG. PMGSY data originate at the habitation level; habitations are smaller than villages. For each village, we construct the variables denoted above from both the earliest new and upgraded roads completed within a shrid. Additional fields from OMMS can be obtained with the data repository included in Asher and Novosad (2019) at the Harvard Dataverse.

Rural Electrification

Data files: `shrug_pcYY.dta`

Variables: `pcYY_vd_power_dom`: Indicator for power for domestic uses (1=Yes)
`pcYY_vd_power_agr`: Indicator for power for agricultural uses (1=Yes)
`pcYY_vd_power_all`: Indicator for power for all uses (1=Yes)
`pc11_vd_power_[type]_sum`: Daily hours of power for [type] use in summer
`pc11_vd_power_[type]_win`: Daily hours of power for [type] use in winter

Village electrification is reported in 1991 and 2001 as a binary measure indicating whether power was available in the village for various uses. We include domestic, agricultural and all, which are consistently reported across all years. The meaning of "all" is not made clear in the village directories.

The binary variables take the value 1 if any of the components of the location (shrid) were reported as having the given type of electricity. Some of these values are known to have a questionable relationship with the actual availability of electricity in the field, especially in 1991 and 2001.

In 2001, the electrification variables were missing for about 60% of villages. The constituency-level data marks average village electrification as missing when electrification information is missing for more than 20% of the constituency population. Users who wish to impute more of these data will need to refer to the original village directories.

In 2011, the village directory began reporting electricity availability in number of hours available per day in summer and winter for each use. These values appears to be more useful as proxies for electricity availability at the village level, but are unfortunately not available in prior years.

Town/Village Size

Data files: `shrug_pcYY.dta`

Variables: `pcYY_vd_area`: Village area in hectares
`pcYY_td_area`: Town area in square kilometers

These variables report the amenities table description of village and town size. Some numbers

from the Census here are substantial outliers and are likely to represent data entry errors.

SECC: Assets and Agriculture

Data file: `shrug_ancillary.dta, bootstrapped_cons`

Variables: `secc_rural_cons_pc`: Small-area estimate of per capita consumption

`secc_inc_cultiv`: Share of households where cultivation is main income source

`secc_rural_cons_pc[1-1000]`: 1000 bootstrapped consumption estimates

Beginning in 1992, the Government of India has conducted multiple household censuses in order to determine eligibility for various government programs; the SECC is the latest of these, conducted mostly in 2012, and the first to cover all households in the country. The SECC was based on the National Population Register (NPR) from the 2011 Population Census.

Variables in the SHRUG are aggregated from individual level responses to the Socioeconomic Caste Census (SECC), undertaken primarily in 2011-12. Consumption is not recorded by the SECC, so we generate small-area estimates following the methodology in Elbers et al. (2003). We use data from the 2011-12 India Human Development Survey (IHDS-II), and regress total household consumption on dummy variables that are equivalent to all asset and earnings information contained in the SECC. We then use the coefficients to predict household-level consumption in the SECC microdata. This is used to generate consumption per capita at the individual level, which is in turn used to produce village level statistics for mean predicted consumption per capita, which is included in the SHRUG. Note that rural consumption is estimated, not urban consumption, thus shrids with both urban and rural components only report rural consumption. SECC households above 20 individuals are dropped from the small-area estimation. Consumption is reported as individual annual consumption expenditure.[4]

Researchers may wish to account for the fact that the coefficients from the model that generate these consumption numbers are estimated with error. To make this possible, we used a bootstrap approach, drawing a full-size sample households from IHDS with replacement and re-estimating village-level per capita consumption 1000 times. These 1000 draws are available as a SHRUG package for download and reflect the distribution of per capita consumption that arises from the first stage estimation process. Researchers can use these 1000 draws in a second bootstrap process to account for the consumption estimation error.

Rural households directly report whether their primary income source is in agriculture, small enterprise, wage work, or another source. We aggregate this and report the share of households in a village that draw their income from agriculture.

[4]The IHDS asks questions with 30-day and annual recall, which are aggregated to annual consumption expenditure. Note that as of the present version of data from ICPSR, the IHDS manual incorrectly states that the data are reported as 30-day consumption measures. SHRUG consumption is left as 365-day per capita consumption for consistency with the IHDS data. Users desiring monthly expenditure should divide by 12.

Future versions of the SHRUG will include additional asset variables from the SECC and small area estimates of urban consumption. In the interim, users can use the provided keys to match the SHRUG to the house listing of the 2011 Population Census, which provides village-level ownership data for a range of assets.

Spatial and Remote Sensing Measures

Data file: `shrug_ancillary.dta`

Variables: `tdist_NNNk`: Distance to nearest town with at least NNN thousand people
`total_light,total_light_cal,max_light`: Night-Time Luminosity
`total_forest,max_forest`: Forest Cover and Forest Loss
`num_cells`: Number of pixels in night light or forest polygon
`thiessen_polygon`: Indicator that shrid polygon was autogenerated

Night lights are from the DMSP-OLS annual measures of night time luminosity, measured at 1/120 degree. `total_light` is the sum of the luminosity values (0–63) of all pixels in the unit. `total_light_cal` is the same value, but calibrated for consistent estimation across time using the method of Elvidge et al. (2014). Average pixel luminosity in a geographic unit can be calculated by dividing one of the total light variables by `num_cells`. The data span 1994 to 2013.

Forest cover 2000–2014 is aggregated from Vegetation Continuous Fields (VCF) 250m resolution data, which provides annual tree cover in the form of the percentage of each pixel under forest cover, generated from a machine learning model based on a combination of images from MODIS and samples from higher resolution satellites (Townshend et al., 2011) and used in (Asher et al., 2019a). By using broad-spectrum measures, VCF is better at distinguishing plantation from primary and secondary forest, which tends to be characterized by cooler temperatures. As with night lights, we report the maximum and total forest cover value in each geographic unit. We use VCF rather than the widely used Global Forest Change for India, because GFC has limited reporting of forest cover gains, and India has gained forest over the sample period according to most accounts.

Each VCF pixel takes a value between 0 and 100 reflecting the percent of that pixel covered by forest. The variable `total_forest` adds up these pixel values for all pixels in a village, town, or constituency. Divide by `num_cells` to get the average forest cover in the location. The variable `max_forest` contains the highest pixel value in that location's polygon.

About 90% of SHRUG locations were georeferenced with polygons, permitting accurate measurement of night lights and forest cover. About 10% of locations, especially in the Northeast, were georeferenced only by points; we constructed Thiessen polygons to match these to the forest cover and night light rasters. These locations are identified by the `thiessen_polygon` field; users of the spatial data may want to verify robustness to excluding these locations.

Distance from each shrid to the nearest location with population over some threshold value in

2011 (the `tdist` variables) were calculated from centroid to centroid.

Remotely sensed measures of air pollution, agricultural production, village assets and consumption, slums, as well as annual precipitation data are slated for inclusion in future versions of the SHRUG.

Election Results

Data files: `assembly_candidates_clean.dta, assembly_elections_clean.dta`

Variables: `tr_ac_id,tr_ac_name`: ECI / Trivedi constituency identifiers

`ac07_id,ac08_id`: SHRUG constituency identifiers

`sh_election_id`: Election identifier

`sh_cand_id`: Election identifier

We include election results from 1990 to the present, kindly posted by the Trivedi Center for Political Analysis at Ashoka University. Users of these data should cite Jensenius and Verniers (2017).

The electoral data include both Trivedi identifiers (`tr_ac_id`) and SHRUG constituency identifiers (`ac07_id` and `ac08_id`). The SHRUG identifiers and constituency names consistently identify a constituency over time; the Trivedi identifiers, which are consistent with the ECI's own numbering, do not.

The variable `sh_election_id` uniquely identifies an election and takes the form `ss-nn-aaa[-p]`, where: `ss` is the state code, `nn` is the assembly number (i.e. 1st assembly, 2nd assembly, etc.) `aaa` is the SHRUG constituency identifier, and `p` is the optional poll number, used only for bye-elections.

The variable `sh_cand_id` uniquely identifies a candidate in the dataset; it takes the value of the candidate vote rank appended to the `sh_election_id`.

The posted election data is missing a few constituencies that can be found in the original Trivedi dataset:

- A few constituencies that only appeared in election results in very early years and not again;
- By-elections that appeared before the first election in Jharkhand and Telangana;
- UP constituencies that ended up in Uttarakhand after 2001; and
- Constituencies with redundant names and changing ECI identifiers; we could not determine which pre/post-split constituencies to match and thus excluded these.

A handful of constituency names are non-unique within states, which is a minor annoyance when merging constituencies to other data. To mitigate this, the SHRUG names (`ac07_name` and `ac08_name`) have been prefixed with the either the 3-digit ECI code (3rd delimitation) or the first four letters of the district name (4th delimitation). For example, `ac08_name` takes the value of `rajk-jetpur` and `vado-jetpur` for the two different `jetpur` constituencies in Gujarat. The field `tr_ac_name` contains the name as described in the ECI data.

Subsection 2.2 further describes the identifiers used in the constituency-level data. For detailed description of the variables used in this dataset, see Jensenius and Verniers (2017).

Candidate Affidavits

Data file: affidavits_clean.dta

Variables: ac07_id,ac08_id: SHRUG constituency identifiers
sh_election_id: SHRUG election identifier
sh_cand_id: SHRUG candidate identifier for ECI-matched data
age, ed, assets, liabilities, etc.: Candidate characteristics
winner: Indicates ADR site shows candidate is election winner
punishment: Years of punishment for most serious criminal charge (top-coded to 50)
num_crim: Number of criminal charges faced
adr_major_crime: Indicates ADR coded at least one crime as serious
adr_cand_id: Candidate identifier used by ADR

Data on political candidate characteristics originate in the affidavits filed by each candidate with the Electoral Commission of India. These were hand-entered from scanned PDFs by the Association for Democratic Reform (ADR). Prakash et al. (2019) re-entered the data for winners and runners-up from 2004 to 2007 to validate the ADR data. The re-entered data corresponds very closely with the original data, lending credence to the work by ADR; type 1 and type 2 errors in presence of criminal accusations in the ADR data occur less than 5% of the time. However, Prakash et al. (2019) code a much higher number of criminal accusations against most candidates. The source variable describes whether a given entry is from the Prakash et al. (2019) coding or from the ADR. Data was collected from ADR in 2013 and 2018, also indicated in the source variable.

Data in the SHRUG uses the Prakash et al. (2019) results where available, and the ADR results where they are not. Users of these data should cite Prakash et al. (2019), who collected and cleaned almost all of the data from ADR. We have added a measure of the severity of the most serious section of the Indian Penal Code under which a candidate has been charged; the punishment variable reports the maximum number of years imprisonment that can result from the charge. This number is top-coded at 50.

We made an effort to match winners and runners up in the affidavit data to the electoral data on the basis of candidate names and other characteristics. This process was not comprehensive and could be improved. The variables sh_cand_id and sh_election_id link these affidavits back to the SHRUG election data. If you link additional candidates, please let us know so we can improve the breadth of these links.

The SHRUG constituency identifiers acYY_id link each constituency to the SHRUG economic constituency data.

References

Asher, Sam and Paul Novosad, "Rural Roads and Local Economic Development," *American Economic Review (forthcoming)*, 2019.

____ , Teevrat Garg, and Paul Novosad, "The Ecological Footprint of Transportation Infrastructure," *The Economic Journal (forthcoming)*, 2019.

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

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

Elvidge, Christopher D, Feng-Chi Hsu, Kimberly E Baugh, and Tilottama Ghosh, "National trends in satellite-observed lighting," *Global urban monitoring and assessment through earth observation*, 2014, *23*.

Jensenius, Francesca R and Gilles Verniers, "Studying Indian politics with large-scale data: Indian election data 1961–today," *Studies in Indian Politics*, 2017, *5* (2).

Prakash, Nishith, Marc Rockmore, and Yogesh Uppal, "Do criminally accused politicians affect economic outcomes? Evidence from India," *Journal of Development Economics*, 2019.

Townshend, J., M. Hansen, M. Carroll, C. DiMiceli, R Sohlberg, and C. Huang, "User Guide for thet MODIS Vegetation Continuous Fields product, Collection 5 Version 1," *Collection 5, University of Maryland, College Park, Maryland*, 2011.