Working paper

International Growth Centre

The Socioeconomic High-resolution Rural-Urban Geographic dataset on India (SHRUG)

Sam Asher
Ryu Matsuura
Tobias Lunt
Paul Novosad

February 2019

When citing this paper, please use the title and the following reference number:
C-89414-INC-1

DIRECTED BY LSE UNIVERSITY OF OXFORD

FUNDED BY UKaid from the British people

The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG)*

Sam Asher†
Ryu Matsuura‡
Tobias Lunt§
Paul Novosad¶

February 2019

Abstract

This paper documents the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG), a new administrative data source describing socioeconomic development in India. The first version of the SHRUG describes demographic, socioeconomic, firm and political outcomes at a high geographic resolution for the universe of Indian households and non-farm productive establishments, in both rural and urban India, from 1990–2013. The SHRUG is a platform for future collaboration and data sharing between researchers working with administrative data in India. We have created a set of consistent location identifiers for all geographic locations in India from 1990–2013, and establish a methodology to extend this classification to units from future data sources. Researchers working with geographic variation in India can thus benefit from linking to the SHRUG, and can benefit other researchers by making their data available to others through the SHRUG platform. In this paper, we describe the construction of the data and the strengths and weaknesses of administrative data like these for research on economic development. We then perform several validation exercises to show that the SHRUG is consistent with other data sources. Finally, we present an illustrative data exercise on recent trends in rural and urban development.

*Thanks to Teevrat Garg, Francesca Jensenius, and Dan Keniston for sharing data or information that contributed to this product. This material includes work supported by the IGC (project 89414), and a project funded by the UK Department for International Development (DFID) and the Institute for the Study of Labor (IZA) for the benefit of developing countries. The views expressed are not necessarily those of IGC, DFID or IZA. All errors are our own.

†World Bank, sasher@worldbank.org
‡Dartmouth College, rmatsuura.ryu@gmail.com
§Dartmouth College, tobias.lunt@dartmouth.edu
¶Dartmouth College, paul.novosad@dartmouth.edu

1 Introduction

This paper documents the Socioeconomic High-resolution Rural-Urban Geographic dataset on India (the SHRUG). This is a new dataset that provides multidimensional socioeconomic data on the universe of cities, towns and villages in India from 1990 to 2013, a location panel with over 600,000 constant boundary geographic units.

The SHRUG differs from conventional sample datasets used to study socioeconomic changes in developing countries along several dimensions. First, the SHRUG is a census rather than a sample. This means that it is extensible: any new census dataset describing the universe of locations in India can be directly linked to the SHRUG at a high geographic resolution with minimal loss. In contrast, sample surveys can be reliably linked together only at very high levels of aggregation. For example, India's flagship socioeconomic survey, the National Sample Survey, is representative only at the state level and does not repeatedly sample the same villages. NSS panels are often constructed at the district level, but these are not representative, and no lower level of aggregation can be obtained.

Second, the SHRUG is identified at the town and village level. This enables the large-scale analysis of programs with substantial cross-village variation. To our knowledge, there is no other large scale panel of towns in India (nor in many other developing countries); instead, comparative analysis of cities occurs at a higher level of aggregation, which pools multiple towns and cities of different sizes into the same units.

Third, the SHRUG is multidimensional. We have incorporated data on political outcomes (election results and candidate asset and criminality affidavits), firm outcomes (four Economic Censuses), population demographics (three population censuses), remote sensed measures of forest cover (Vegetation Continuous Fields) and economic activity (night lights), and administrative data on government programs (such as PMGSY, India's national road construction program).

The SHRUG may be beneficial to researchers in a number of dimensions. We provide a few examples in this document, but the list is far from exhaustive. For example, to our knowledge,

the SHRUG is the first socioeconomic panel that is aggregated to the boundaries of legislative constituencies, which are ten times smaller than districts. Another use case is the generation of historical data on a potential sample for a field experiment. Typically, researchers use some form of a national population census as a sampling frame for a new experimental study in the field. The SHRUG will make it possible to obtain a 25-year socioeconomic history of each potential location for a field experiment. This will make it possible, for instance, to test for divergent trends in fields locations even before survey collection has entered the field.

The data underlying the SHRUG has already been used in several research projects, including Adukia et al. (2017), Asher and Novosad (2017), and Asher and Novosad (2018). None of these projects would be possible with conventional sample data, because all of them rely on natural experiments with variation occurring at levels of aggregation that are smaller than districts.

The SHRUG is based on a combination of census data collected by the Indian government, supplemented with several varieties of remote sensing data. The foundation of the SHRUG is a core set of censuses: the population censuses of 1991, 2001 and 2011, and the economic censuses of 1990, 1998, 2005 and 2013. While each of these datasets contains information on individuals and firms in the universe of towns and villages, to our knowledge these datasets have never before been linked together and aggregated into constant geographic units.[1] The linking process consists of fuzzy merging on the basis of names and identifiers at various levels of aggregation, with merge information supplemented by research in the physical volumes describing aggregations of towns and villages across the different census periods. While the process is relatively straightforward, it is extremely labor intensive; we estimate that over 5000 person hours of work were involved in linking and cleaning all of these datasets.

In its present condition, the SHRUG describes: (i) demographic data on every town and village in India from 1991 to 2011; (ii) employment and industry data on every firm in India from 1990 to 2013; (iii) legislative election results and party history from 1980 to 2013; (iv)

[1]Various research teams have assembled partial matches of these units at various times. See, for instance, Muralidharan et al. (2017), Lehne et al. (2018), and Burlig and Preonas (2016).

electoral candidate affidavit information from 2004 to 2013; (v) remotely sensed night lights from 1992 to 2014; and (vi) remotely sensed forest cover from 2000 to 2014. Because of the nature of census data, the breadth of this panel will continue to grow, as each new census or remote sensing dataset can be fully linked to all the previous datasets.

There is now a wealth of data on the implementation of government programs that is beyond the ability of any single research team to exploit. In India, many of these datasets are organized according to population census locations (either by name or identifier), making the SHRUG the ideal complement to their use. Our hope is that researchers will enrich their work by bringing in fields from the SHRUG, and will simultaneously improve the SHRUG by posting their administrative data with SHRUG identifiers when their work is published.

We recognize that when researchers share data, they face the risk of being scooped on future potential projects with that data. We propose that researchers post their comprehensive administrative datasets with SHRUG identifiers at the time of publication of their first paper with that administrative data. Given the time lag from project completion to publication in Economics, this will give them a huge lead on any potential projects, while simultaneously creating a public good in a timely enough fashion for other researchers to benefit from it.

In this paper, we describe the construction and contents of the SHRUG (Sections 2 and 3), the strengths and weaknesses of this dataset relative to traditional sample datasets like the NSS or Annual Survey of Industry (Section 4), and illustrative analyses that are possible with the SHRUG but difficult with any other data source (Section 5). Section 6 concludes with a discussion of data to be added to the SHRUG in the future, and proposes a framework for sharing of future census data among members of the research community.

2 Data Contents

Table 1 describes the components of the SHRUG in summary terms. The following subsections describe the different components in detail.

2.1 SHRUG Identifiers

Each village or town unit in the SHRUG is identified by a SHRUG identifier, or a *shrid*. A shrid describes a geographical unit that can be mapped consistently across multiple rounds of the Indian population and economic censuses. In the majority of cases, a shrid is a village or town. When villages or towns have merged or separated between 1990 and 2013, we have aggregated them in the periods where they are separated, such that the aggregation is represented by a single shrid in each period. Some shrids are thus composed of multiple population census villages or towns, or a combination of villages and towns. The shrid consists of a two digit census year based on the census year in which the shrid was defined, then a two digit state identifier and a multidigit village or town identifier based on the lowest numbered code for a unit in that shrid in that census year.[2]

The use of a single consistent unit for each geographic location over time creates a tremendous simplification for the researcher, as it can take a substantial amount of time to identify how components of various census units have changed over time.[3]

2.2 Population Census and Amenities Tables

The Indian Population Census, undertaken in 1991, 2001 and 2011 is a complete enumeration of households in India. Tabulations are provided by the government at the village and town level. The Population Census Abstract (PCA) includes the number of households and population of men and women at various age groups and in various social groups, number of workers in different occupation classes.

The Population Census also published a village and town directory, which describes an increasingly large set of public goods. For example, the amenities include but are not limited to data on paved road and electrification, the distance of villages to the nearest town, the

[2]For instance, 11-03-123456 identifies a location in state 03 (Punjab), that had the town or village code 123456 in census year 2011. This approach is taken so that new agglomerations can be created without conflicting names in the future as census towns and villages are newly split or combined in future rounds of the Economic and Population Censuses.

[3]To put a fine point on it, thousands of person hours have been spent by our research team alone studying the consistency of location units over time.

presence of post offices and various medical facilities, the number of market days, and the source of water. A different set of amenities is recorded for census villages and census towns, the latter of which are defined as units with population greater than 5000 with 75% of the employed workforce outside of the agricultural sector. When villages and towns are pooled, the SHRUG reports both the village and town amenities for each unit.

The downloadable village/town SHRUG dataset includes only total male and female population and several village/town amenities, but the SHRUG keys allow the SHRUG to be fully linked to the full data from the population censuses, which can be downloaded or purchased from the web site of the Census of India.[4] The constituency level data contains a much larger set of data from the PCA and the amenities tables; we have chosen approximately thirty of the fields to include in the present version.

2.3 Economic Census and Industry Definitions

The Economic Census of India is a complete enumeration of non-farm establishments, undertaken in 1990, 1998, 2005 and 2013. The frame for the survey is the house listing from the most recent population census. The Economic Census reports a range of establishment characteristics, including four-digit sector, source of finance, source of power, gender and social group of owner, and number of employees of each gender.

Sector is reported using the India-specific National Industry Classification system, which has changed several times over the sample period. The 1990 and 1998 Economic Census use NIC1987, the 2005 Economic Census uses NIC2004, and the 2013 Economic Census uses NIC2008. Using the concordances across these three groups, we constructed a set of 90 industry groups that are almost entirely constant across each period. We call these SHRUG Industry Codes, or Shrics. We include a correspondence between Shrics and industry codes in all the individual years so that users of the raw Economic Census data can reconstruct or modify our categories in any way.

[4]Code is supplied to consistently match the variables across different rounds of the village and town censuses.

To require perfect synchronization across all periods would have required an unreasonable amount of group pooling. For instance, if a small category of employment (such as digital journalism), moves between major sectors (i.e. from the computer sector (in 1987) to the media sector (in 2004)), then this would require us to treat computers and media as the same sector in all years. In other words, there is a tradeoff between consistency over time and distinguishing between very different sectors. To create a set of 90 industry groups, we removed entries from the concordance that affected a tiny number of jobs but, if kept, would cause large and distinct industry categories to be aggregated. To continue with the example above, digital journalism jobs are thus counted in the computer sector in 1998 but in the journalism sector in 2005. This does not create a substantial inconsistency across time, because there are very few jobs in digital journalism in 1998 or earlier.

We exclude jobs in the agricultural sector, which are not consistently recorded by the Economic Census, neither within round nor across rounds. We also exclude jobs in public administration and defense (NIC2008 Section O), which were not counted in the 2013 Economic Census. However, we have kept the majority of public sector establishments, including teachers, healthcare workers, railway employees, and employees of other state-owned enterprises.

The SHRUG contains total employment in each sector-year, in every town and village, as well as the key between Shrids and Economic Census identifiers. Researchers interested in additional fields from the Economic Census (such as worker gender, source of finance, or more granular sectoral information) or in working with the firm microdata can purchase the Economic Census from MoSPI and easily link it to the SHRUG via the keys.

2.4 Administrative Data

The first version of the SHRUG includes administrative data from the PMGSY, the Prime Minister's Road Building Program, under which over 100,000 roads were built or improved between 2000 and 2013. These data were scraped from the online program implementation portal (`http://omms.nic.in` at the time of writing). A wealth of data is available on each road, including the length, the construction material, the preconstruction state of the road,

the time of contract awarding, completion, and milestone dates, among others. PMGSY data are at the level of the road or the habitation; there are typically one to three habitations in each village. Data are thus many-to-one matched to Shrids. Data were matched on village names in the PMGSY habitation list, as well as in the list of villages connected by each road. 85% of villages in the PMGSY were matched to the SHRUG. More details are available in Asher and Novosad (2018).

2.5 Electoral Data

Electoral data are available for legislative constituencies. SHRUG v1.0 does not include electoral data for parliamentary constituencies. Constituency identifiers consist of a delimitation year (2007 for pre-2007, and 2008 for 2008 and later), a two digit census state code, and a four digit constituency identifier. The constituency identifiers are close but not exact matches to those used by the Electoral Commission of India (ECI), because the ECI itself has not used consistent numeric constituency identifiers over time. It is therefore advisable to match constituencies on both codes and names. Election data were scraped, cleaned, and shared with us by Francesca Jensenius (2017), and are now updated by the Trivedi Center for Political Data. We have included turnout, vote totals for each candidate, and party for all elections from 1980–2013. Party history and coalition information is available in the replication data posted with Asher and Novosad (2017).

2.6 Remote Sensing Data

SHRUG v1.0 includes data generated from two remote sensing sources. Night lights are widely used as a proxy for some form of electrification or economic activity when time series data on economic activity is otherwise unavailable (Henderson et al., 2011). Gridded night lights data from the National Oceanic and Atmospheric Administration (NOAA) were matched to village and town polygons, and aggregated into totals and means. These are available annually from 1992–2013.

Forest cover data comes from Vegetation Continuous Fields (VCF), a MODIS product

that measures tree cover at 250m resolution from 2000 to 2014. VCF is predicted from a machine learning algorithm based on broad spectrum satellite images and trained with human-categorized data, which can distinguish between crops, plantations and primary forest cover. For more information, see Asher et al. (2018) and Townshend et al. (2011). As with night lights, we match these to location boundaries and report mean tree and total tree cover for each location.

3 Data Construction

3.1 Matching the Population and Economic Censuses

The starting point for the SHRUG is the 2011 Population Census of India, which reports demographic data at the town and village level. We merged this at the town- and village-level to the 2001 and 1991 Population Censuses, and then to the 1990, 1998, 2005 and 2013 Economic Censuses using a range of information available in these datasets. The Population Censuses in some case (especially in 2011) provide the identifiers of units in earlier population censuses or data from earlier population cenuses, which are the easiest cases for matching.[5]

When the previous identifiers are not available, we conduct a hierarchical match from the largest to the smallest units. We begin with a match of districts across population censuses. The 1991–2001 was shared with us by Kumar and Somanathan (2015). We constructed the 2001–2011 district match based on the back-referenced village ids in the 2011 census, which provided a 2001 census village id for every 2011 village.[6] Within districts, we then matched subdistricts on the basis of names where possible, and then we matched villages within subdistricts, again on the basis of names.

To match names, we used a fuzzy matching algorithm that we developed called `masala_merge`, which executes a modified Levenshtein edit distance algorithm. This algorithm creates a distance between two words, which is the number of letter changes, insertions or deletions to move between two words. We modified this algorithm to apply smaller penalties to sets of

[5]For instance, the town censuses report sex ratio and population in earlier censuses.

[6]These back-references did not exist for the 1991–2001 census links.

phonemes with common spellings. For example, we imposed an edit distance of 0.2 between "X" and "KS" which are used interchangeable in many latin transliterations of Indian language words. We also penalized nasalizations, vowel changes and duplications at a lower rate than consonant changes. The algorithm is posted publicly on our web sites, with the complete list of edit costs. Table 2 summarizes the share of population from each population census that is matched to SHRUG by state.

To match the Economic Censuses to the Population Censuses, we used the location directories for 1998, 2005 and 2013, which were shared with us by the Ministry of Statistics, Planning and Information (MoSPI). We could not find a location directory for the 1990 Economic Census, but we were told by MoSPI that the identifiers used were the same as those in the 1991 Population Census for villages in many states. It was very straightforward to distinguish these states from the ones which created new codes, and we matched villages on the basis of identifiers in these places. For towns, we constructed an algorithm that matched towns only if: (i) they could be uniquely matched within districts to the 1991 population census on the number of wards; (ii) their within-district size rank was the same in the Economic and Population Censuses; and (iii) the number of people per economic census job was within an order of magnitude of the dataset mean, which was approximately 20. Because of the absence of the 1990 location directory, the match rate for the 1990 Economic Census is thus much lower than for the other censuses. Table 3 summarizes the share of employment in each Economic Census that is matched to the SHRUG.

Additional administrative datasets (such as the PMGSY road data) were matched using a similar approach. The PMGSY match is described in more detail in Asher and Novosad (2018).

3.2 Creating a Constituency-Level Panel

The constituency-level SHRUG was created by using geographic data on town, village, and constituency boundaries obtained from MLInfoMap.

Creating a constituency level panel of population and employment poses a number of

challenges. First, because of the fuzzy matching process, there are some villages which were matched to some Economic Censuses and not to others. Simply aggregating employment in matched villages to the constituency level would thus overstate employment gain in constituencies that have better village and town match rates over time.

We can resolve this problem by noticing that in the 2011 Population Census, we have matched 100% of towns and villages to constituencies. For each constituency, we therefore know the population in towns and villages that were matched to a given Economic Census and the population that were not matched. We then scale up employment in each constituency, by assuming that the employment to population share in missing locations is the same as the employment to population share in non-missing locations. We perform the same rescaling based on 2011 population totals to estimate constituency population in 1991 and 2001, but in these cases the imputation is almost inconsequential because the match rates for the 1991 and 2001 censuses are so high. We report the share of jobs that are imputed in the constituency-level SHRUG, so researchers can choose to exclude constituencies with high imputation shares. We drop all constituencies where more than 25% of employment is imputed.

Another challenge that arises is that the available polygon shapefiles for constituencies and towns/villages are not perfectly aligned, even though all of them use the same WGS84 projection. The misalignment is small—on the order of several hundred meters in the worst cases—but it is enough that the smallest villages cannot be unambiguously assigned to a single constituency. We then drop constituencies where more than 25% of their population in 2011 is in villages or towns that cannot be decisively assigned. We have explored several alternate sources of data and spoken with several other experts on Indian spatial data, and to our knowledge there are currently no higher accuracy shapefiles than these, so this amount of error is unavoidable. There are several ongoing projects to assign villages to constituencies by digitizing electoral rolls; as these data become available, they will be integrated into future versions of the SHRUG.

A third challenge is that some census towns contain multiple constituencies. Because the

population censuses do not report consistent identifiers at the subtown level, it is difficult to identify the population of these constituencies — we know only the aggregate population of the combined constituencies.[7] We therefore exclude constituencies that include any part of towns that cross constituency boundaries. The constituency SHRUG is therefore not fully representative, in that it excludes big cities. However, we are not aware of any other research that measures or exploits socioeconomic data at the constituency level for a large number of urban constituencies, because of the boundary misalignment issue. Constructing this data using the ward maps for India's largest cities would be a valuable contribution that would enable better study of politics in India's growing cities.

Finally, India began the process of redrawing constituencies in 2002 following the 2001 census, with the new delimitation taking place in 2008 (Iyer and Reddy, 2013). This is not a problem for data construction, since constituencies are simply defined as polygons. We therefore create separate complete SHRUG panels from 1990–2013 under both the old and the new constituency delimitation. Researchers can thus make their own decisions regarding which polygons to use for which periods.

For the remote sensing data, we simply generated mean and total night light and tree cover variables for each constituency-year, under both the old and the new constituencies.

4 Strength and Weaknesses of our Approach

The SHRUG has two main advantage relative to other data sources. First, it describes socioeconomic outcomes over a two decade period for the universe of locations at a much higher geographic resolution than any other Indian panel dataset. The enables analysis of factors and policies that vary at geographic units below the state or district level, such as politician identities, or village-targeted programs.

Second, because of the Census nature of the data, the SHRUG will continue to improve as a research tool with time. Each new census or administrative or remote sensing data source

[7]The population censuses do report data at the ward level, but the wards change over time and do not necessarily overlap with constituencies.

that is added to the SHRUG will be seamlessly integrated with all the other data sources, expanding the scope of potential analysis. This is a tremendously valuable feature that is not found in sample datasets. If two research teams each conduct new sample surveys (for example, a household finance survey and a consumption survey), those datasets can rarely be used together, because there is virtually zero overlap in the set of sample villages. In contrast, if two research teams work to integrate new sources of administrative or remote sensing data into the framework of the SHRUG, both of those data sources can immediately become useful to all other researchers who are working with the SHRUG.

The main limitations of the SHRUG are that (i) not all villages and towns are matched in all periods; and (ii) the set of fields is more limited than those in conventional research datasets. If a researcher's goal is to collect national statistics, aggregating from the NSS samples is likely to generate less biased national statistics, in part because the SHRUG has a lower match rate for cities than for rural areas, because their boundaries change more frequently. Second, because the administrative censuses at the heart of the SHRUG are implemented for every household and firm in India, they are necessarily based on short surveys, and do not contain the kind of detailed household or firm information that can be found in the NSS.

Researchers should therefore choose between SHRUG and NSS/ASI with care to the particular research question. Questions that rely on high geographic resolution variation, or that require socioeconomic outcomes in units with political boundaries are those who will find the SHRUG most valuable.

5 Illustrative Analyses with the SHRUG

5.1 Data Preparation

The SHRUG includes a multidimensional data quality score for each observation. This reflects our confidence in the quality of the match across datasets, and the quality of the data recorded by administrators. Note that errors in data are not unique to our context, but are a feature of every sample survey as well. The advantage of our context is that we can

generate some measures of confidence in individual observations.

We flag an observation if it was matched only with a high level of fuzziness, if it joins multiple villages or towns with non-overlapping polygons, and if it is a 0.1% outlier in its employment-to-population ratio in any period, or in its size to light output from the satellite data.

For cross-village and cross-town analysis with the SHRUG, we recommend taking a conservative approach and excluding all observations which have quality flags. However, these observations may be worth studying in more detail for projects specifically focused on the regions where quality is uncertain.

In all of our analysis, we drop all observations with quality flags. We also drop villages with population less than 100 in 2011, as their rates of change are large relative to their importance. We also drop observations from island states including Lakshadweep and the Andaman and Nicobar Islands.

6 Conclusions: A Model for Collaborative Data Sharing

Most data collection projects initiated by researchers continue to have relatively narrow scope. A local survey is conducted for the purpose of an experimental or policy study, one or several research papers are written up, and the data is re-used only for replication, or in rare cases, for long-term followup.

The era of administrative data makes possible a framework for research where projects have many more positive externalities on other researchers. Because administrative data is often comprehensive at the state or the national level, one researcher's effort at collecting and rationalizing an administrative dataset may yield dividends to many other researchers. Many researchers in India are already making use of administrative data, but in the absence of a common platform to link these datasets to each other, there is both considerable duplication of work and many potential complementarities across projects are not being realized.[8]

[8]Some examples include the NREGS public works and wage support program, the RGGVY rural electrification program, and the ongoing Total Sanitation Campaign, all of which are the subject of multiple research papers. And yet none of these programs have easily accessible data frames, causing each new

One goal for this work is for it to provide such a common platform to help these positive externalities emerge.

India is particularly well-suited to such an approach because of the high digitization and transparency of government programs. This approach may nevertheless serve as a model for other countries, and will perhaps motivate the effort of bringing existing government censuses into a common spatial frame.

researcher to have to reinvent the wheel, and limiting the scope of what any one team can study.

References

Adukia, Anjali, Sam Asher, and Paul Novosad, "Educational Investment Responses to Economic Opportunity: Evidence from Indian Road Construction," 2017. Working paper.

Asher, Sam and Paul Novosad, "Politics and Local Economic Growth: Evidence from India," *American Economic Journal: Applied Economics*, 2017, *9* (1), 229–273.

____ and ____ , "Rural Roads and Local Economic Development," 2018. Working paper.

____ , Teevrat Garg, and Paul Novosad, "The Ecological Footprint of Transportation Infrastructure," 2018.

Burlig, Fiona and Louis Preonas, "Out of the Darkness and Into the Light? Development Effects of Rural Electrification," 2016. Working Paper.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "A Bright Idea for Mesuring Economic Growth," *American Economic Review*, 2011, *101* (3), 194–199.

Iyer, Lakshmi and Maya Reddy, "Redrawing the Lines: Did Political Incumbents Influence Electoral Redistricting in the Worldâs Largest Democracy?," 2013. Harvard Business School Working Paper 14-051.

Jensenius, Francesca, *Social Justice through Inclusion* 2017.

Kumar, Hemanshu and Rohini Somanathan, "State and district boundary changes in India: 1961-2001," 2015. Working Paper.

Lehne, Jonathan, Jacob Shapiro, and Oliver Vanden Eynde, "Building Connections: Political Corruption and Road Construction in India," *Journal of Development Economics*, 2018, *131*, 62–78.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar, "General equilibrium effects of (improving) public employment programs: Experimental evidence from india," 2017.

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.

Table 1
SHRUG Summary

Panel A. Data in the SHRUG

Dataset	Years	Description	Units of observation
Population Census	1991, 2001, 2011	Demographic data, social groups, village & town public goods	Village, Constituency, District
Economic Census	1990, 1998, 2005, 2013	Employment and sector of all non-ag firms	Village, Constituency, District
Election Results	1980-2013	Candidate name / party / votes	Constituency
ADR Summary	2004-2013	Winner's assets, liabilities, criminal charges	Constituency
Night Lights	1992-2014	Proxy for electrification and economic activity	Village, Constituency, District
Forest Cover	2000-2014	% Tree cover from Vegetation Continuous Fields	Village, Constituency, District
Road Implementation	2000-2013	Administrative data from PMGSY	Village

Panel B. Identifiers for matching, but no data

Dataset	Years	Description	Units of observation
Candidate Affidavits (ADR)	2004-2013	Assets, liabilities, criminal charges	Candidate, Constituency

Notes: Economic Census employment includes informal and service sectors among all non-agricultural jobs.

Table 2

State-level population for all states

States	PC91	PC01	PC11
India	823727.52 / 833127.72 (99%)	1023125.55 / 1023350.18 (100%)	1206229.26 / 1207131.57 (100%)
Andaman Nicobar Islands	280.66 / 280.66 (100%)	356.15 / 356.15 (100%)	380.55 / 380.58 (100%)
Andhra Pradesh	64979.35 / 66455.27 (98%)	75525.48 / 75527.30 (100%)	83704.11 / 83704.11 (100%)
Arunachal Pradesh	621.18 / 637.04 (98%)	1097.97 / 1097.97 (100%)	1383.17 / 1383.73 (100%)
Assam	22231.89 / 22311.78 (100%)	26516.38 / 26531.87 (100%)	30920.17 / 31126.14 (99%)
Bihar	86005.49 / 86374.47 (100%)	82713.36 / 82713.37 (100%)	104074.33 / 104074.34 (100%)
Chandigarh	642.01 / 642.01 (100%)	900.63 / 900.63 (100%)	1046.43 / 1046.43 (100%)
Chhattisgarh		20705.04 / 20711.10 (100%)	25430.23 / 25431.17 (100%)
Dadra Nagar Haveli	138.48 / 138.48 (100%)	220.49 / 220.49 (100%)	343.71 / 343.71 (100%)
Daman & Diu	101.59 / 101.59 (100%)	158.20 / 158.20 (100%)	243.25 / 243.25 (100%)
Goa	1139.41 / 1169.79 (97%)	1336.66 / 1336.66 (100%)	1427.57 / 1427.57 (100%)
Gujarat	40872.70 / 41309.58 (99%)	49512.94 / 49512.94 (100%)	59742.89 / 59742.89 (100%)
Haryana	16278.88 / 16459.98 (99%)	21052.17 / 21057.34 (100%)	25127.14 / 25285.10 (99%)
Himachal Pradesh	5146.16 / 5170.53 (100%)	6077.90 / 6077.90 (100%)	6851.99 / 6852.15 (100%)
Jammu Kashmir		9831.41 / 9832.35 (100%)	12341.06 / 12346.58 (100%)
Jharkhand		26889.57 / 26889.57 (100%)	32937.65 / 32942.15 (100%)
Karnataka	44189.29 / 44977.20 (98%)	52413.44 / 52472.17 (100%)	60848.72 / 60911.60 (100%)
Kerala	28246.68 / 29098.52 (97%)	30999.67 / 30999.67 (100%)	33079.96 / 33079.96 (100%)
Lakshadweep	51.71 / 51.71 (100%)	60.65 / 60.65 (100%)	64.47 / 64.47 (100%)
Madhya Pradesh	62057.47 / 63026.21 (98%)	60075.04 / 60077.79 (100%)	72460.40 / 72460.40 (100%)
Maharashtra	78313.54 / 78936.42 (99%)	96878.63 / 96878.63 (100%)	112311.34 / 112362.17 (100%)
Manipur	1806.38 / 1837.15 (98%)	2159.86 / 2159.86 (100%)	2837.34 / 2841.70 (100%)
Meghalaya	1764.66 / 1774.74 (99%)	2288.95 / 2318.82 (99%)	2961.91 / 2966.89 (100%)
Mizoram	689.54 / 689.76 (100%)	888.36 / 888.57 (100%)	1094.51 / 1097.21 (100%)
Nagaland	1207.14 / 1209.55 (100%)	1989.66 / 1990.04 (100%)	1978.50 / 1978.50 (100%)
NCT of Delhi	9420.64 / 9420.64 (100%)	13850.51 / 13850.51 (100%)	16787.94 / 16787.94 (100%)
Odisha	31431.20 / 31587.64 (100%)	36690.56 / 36695.47 (100%)	41745.87 / 41774.55 (100%)
Puducherry	723.80 / 807.78 (90%)	914.37 / 914.37 (100%)	1234.59 / 1234.59 (100%)
Punjab	18985.14 / 19053.16 (100%)	24192.64 / 24216.74 (100%)	27497.85 / 27590.98 (100%)
Rajasthan	43307.33 / 43879.50 (99%)	56434.21 / 56439.12 (100%)	68461.18 / 68461.19 (100%)
Sikkim	405.02 / 405.02 (100%)	540.85 / 540.85 (100%)	610.57 / 610.58 (100%)
Tamil Nadu	55111.89 / 55834.15 (99%)	62367.39 / 62405.68 (100%)	72069.93 / 72099.37 (100%)
Tripura	2456.15 / 2757.20 (89%)	3198.93 / 3199.20 (100%)	3666.08 / 3673.92 (100%)
Uttarakhand		165898.13 / 165910.03 (100%)	199519.85 / 199568.78 (100%)
Uttar Pradesh	138350.71 / 138842.87 (100%)	8440.64 / 8450.65 (100%)	10000.39 / 10015.27 (100%)
West Bengal	66771.42 / 67887.31 (98%)	79948.71 / 79957.51 (100%)	91043.62 / 91221.62 (100%)

Table 2 presents the state-level population included in the SHRUG panel in the numerator divided by the state-level population in the PC datasets in the denominator for all the states and union territories in India. It also presents the share of state-level population in the SHRUG panel to state-level population in the PC datasets in the parentheses. Note that the population numbers are reported in 1,000.

Table 3

State-level employment for all states

States	EC90	EC98	EC05	EC13
India	43270.64 / 62211.08 (70%)	62876.29 / 70891.77 (89%)	79098.00 / 85388.85 (93%)	107400.96 / 110513.80 (97%)
Andaman Nicobar Islands	12.27 / 31.14 (39%)	48.32 / 48.32 (100%)	17.00 / 39.05 (44%)	61.09 / 61.21 (100%)
Andhra Pradesh	4081.34 / 5263.04 (78%)	5743.06 / 6243.11 (92%)	8568.37 / 8991.79 (95%)	10492.03 / 11563.89 (91%)
Arunachal Pradesh	13.00 / 61.86 (21%)	48.50 / 54.68 (89%)	64.96 / 81.30 (80%)	89.80 / 108.38 (83%)
Assam	994.49 / 1265.52 (79%)	1626.39 / 1914.82 (85%)	1731.44 / 2037.68 (85%)	3604.46 / 3665.87 (98%)
Bihar	2467.26 / 2915.64 (85%)	1715.33 / 2028.94 (85%)	2031.15 / 2096.17 (97%)	2929.18 / 3116.34 (94%)
Chandigarh	137.46 / 137.46 (100%)	148.16 / 148.16 (100%)	185.33 / 185.33 (100%)	244.27 / 244.27 (100%)
Chhattisgarh		1003.77 / 1154.32 (87%)	1154.27 / 1377.39 (84%)	1795.89 / 1834.96 (98%)
Dadra Nagar Haveli	13.23 / 13.23 (100%)	27.36 / 31.04 (88%)	64.61 / 64.61 (100%)	94.31 / 94.31 (100%)
Daman & Diu	18.55 / 18.55 (100%)	29.80 / 29.86 (100%)	59.84 / 59.84 (100%)	18.20 / 81.42 (22%)
Goa	87.27 / 169.84 (51%)	153.98 / 191.81 (80%)	187.36 / 208.13 (90%)	284.58 / 284.92 (100%)
Gujarat	2287.73 / 2831.85 (81%)	3676.17 / 3779.33 (97%)	3957.48 / 4412.87 (90%)	6144.44 / 6246.70 (98%)
Haryana	939.68 / 1190.77 (79%)	1052.98 / 1408.53 (75%)	1742.47 / 1950.83 (89%)	2772.45 / 2845.80 (97%)
Himachal Pradesh	327.27 / 357.05 (92%)	450.05 / 461.38 (98%)	543.54 / 552.25 (98%)	916.07 / 938.60 (98%)
Jammu Kashmir		100.83 / 430.17 (23%)	546.40 / 645.96 (85%)	1043.16 / 1065.65 (98%)
Jharkhand		866.09 / 947.85 (91%)	991.34 / 1030.31 (96%)	1377.48 / 1386.44 (99%)
Karnataka	3571.70 / 6339.23 (56%)	4069.62 / 4228.16 (96%)	5035.00 / 5165.28 (97%)	5787.71 / 5829.52 (99%)
Kerala	2223.42 / 2961.80 (75%)	585.07 / 3249.12 (18%)	2931.26 / 4309.21 (68%)	5646.54 / 5701.44 (99%)
Lakshadweep		5.87 / 12.18 (48%)	8.37 / 8.37 (100%)	9.92 / 10.24 (97%)
Madhya Pradesh	2867.69 / 3190.24 (90%)	3160.40 / 3325.93 (95%)	3334.37 / 3531.72 (94%)	3972.34 / 4241.05 (94%)
Maharashtra	7187.69 / 7577.37 (95%)	8134.96 / 8381.88 (97%)	9036.32 / 9526.52 (95%)	11797.20 / 11947.80 (99%)
Manipur	9.93 / 133.45 (7%)	109.61 / 167.68 (65%)	147.97 / 204.65 (72%)	323.30 / 385.92 (84%)
Meghalaya	30.52 / 126.71 (24%)	133.20 / 144.36 (92%)	179.10 / 194.70 (92%)	269.67 / 277.45 (97%)
Mizoram	46.78 / 49.23 (95%)	46.98 / 52.25 (90%)	68.40 / 70.18 (97%)	93.97 / 101.05 (93%)
Nagaland	3.67 / 98.66 (4%)	92.67 / 95.23 (97%)	114.70 / 115.90 (99%)	157.44 / 159.77 (99%)
NCT of Delhi	1860.30 / 1860.30 (100%)	3331.36 / 3331.36 (100%)	3387.83 / 3387.83 (100%)	3003.69 / 3003.82 (100%)
Odisha	738.33 / 2205.11 (33%)	1842.30 / 2738.37 (67%)	3312.57 / 3355.95 (99%)	3891.08 / 4051.32 (96%)
Puducherry	84.80 / 104.51 (81%)	143.85 / 155.09 (93%)	101.85 / 165.52 (62%)	211.31 / 213.67 (99%)
Punjab	1210.66 / 1555.16 (78%)	1844.14 / 1914.10 (96%)	2366.73 / 2399.82 (99%)	3125.31 / 3139.81 (100%)
Rajasthan	1745.24 / 2203.52 (79%)	2687.16 / 2885.55 (93%)	3288.03 / 3569.26 (92%)	4892.09 / 5165.42 (95%)
Sikkim	18.00 / 35.24 (51%)	15.69 / 33.56 (47%)	6.39 / 48.67 (13%)	84.61 / 84.65 (100%)
Tamil Nadu	976.67 / 5266.63 (19%)	5842.72 / 6377.40 (92%)	6903.60 / 8052.45 (86%)	8718.60 / 8812.22 (99%)
Tripura	0.00 / 203.84 (0%)	153.32 / 218.62 (70%)	257.57 / 324.29 (79%)	376.84 / 382.24 (99%)
Uttarakhand		354.44 / 448.05 (79%)	7249.33 / 7328.97 (99%)	11377.18 / 11422.24 (100%)
Uttar Pradesh	5406.84 / 7505.02 (72%)	6045.04 / 6283.58 (96%)	564.74 / 619.01 (91%)	801.38 / 980.15 (82%)
West Bengal	3908.85 / 6539.10 (60%)	7587.08 / 7976.98 (95%)	8958.33 / 9277.06 (97%)	10993.35 / 11065.24 (99%)

Table 3 presents the state-level employment included in the SHRUG panel in the numerator divided by the state-level employment in the town- or village-level collapsed EC datasets in the denominator for all the states and union territories in India. It also presents the share of state-level employment in the SHRUG panel to state-level employment in the EC datasets in the parentheses. Note that the employment numbers are reported in 1,000.

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.co.uk