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

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

July 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. The backbone is a set of consistent location identifiers for all geographic locations in India from 1990–2013, along with 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. A future version of this paper will present some illustrative data exercises that demonstrate the strengths of this data source.

*Thanks to Teevrat Garg, Francesca Jensenius, and Dan Keniston for sharing data 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 The World Bank, IGC, DFID or IZA. All errors are our own.

†World Bank, sasher@worldbank.org
‡Development Data Lab, toby@devdatalab.org
§Northwestern University, ryumatsuura@u.northwestern.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 information 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. Data are also aggregated to legislative constituencies.

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 or district level and does not repeatedly sample the same villages. NSS panels are often constructed at the district level, but these are often based on fewer than one hundred households per district, and no lower level of aggregation can be obtained. They thus do not well capture the massive variation in outcomes within districts.

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 typically occurs at higher levels of aggregation, pooling multiple towns and cities of different sizes into the same units.

Third, the SHRUG is multidimensional. It incorporates data on political outcomes (election results), politicians (criminal charges, assets, and liabilities), firm outcomes (four economic censuses), population demographics (three population censuses), remotely 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 non-exhaustively list three domains where the SHRUG has a comparative advantage:

1. Substantial variation in socioeconomic outcomes and in programs occurs at a very local level. The SHRUG makes it possible to study these programs at the level of that variation rather than at more aggregate levels where program variation is attenuated.
2. SHRUG permits time series analysis of socioeconomic characteristics at the level of India's 8000 towns. Other than for the largest cities, there is little prior data available at the town-level in India.
3. Few to none of India's prior data sources are aggregated at the legislative constituency level.[1] By aggregating microdata to the legislative constituency level, the SHRUG permits analysis of politics and development at the level at which representatives are elected.
4. Researchers running field experiments typically use a national population census as a sampling frame for a new experimental study in the field, but have limited additional data on locations before collecting their own baseline surveys. The collection of data in the SHRUG increases the scope of what is known at a local level. This allows, for instance, to test for divergent trends in field 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. (2019), Asher and Novosad (2017), and Asher and Novosad (2019). 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. Other authors have collected and merged administrative data for similar purposes (Burlig and Preonas, 2016; Chhibber and Jensenius, 2016; Lehne et al., 2018; Muralidharan et al., 2017); this often involves different researchers duplicating a very similar time-intensive matching process to that underlying the SHRUG. We propose the SHRUG as a backbone for administrative data in India which could substantially reduce the need for this kind of

[1] There are approximately ten legislative constituencies per district.

redundant work.

The SHRUG is based on a combination of census data collected by the Indian government, supplemented with several types of remote sensing data. The foundation of the SHRUG is a set of national 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 not previously been linked and aggregated into constant geographic units. 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.

At the time of writing, the SHRUG describes: (i) demographic and public goods data on every town and village in India from 1991 to 2011; (ii) employment and location of every firm in India from 1990 to 2013; (iii) legislative election results from 1980 to 2013; (iv) assets, liabilities, and criminal charges of all politicians in office and many additional candidates; (v) remotely sensed night lights from 1994 to 2013; (vi) remotely sensed forest cover from 2000 to 2014; (vii) the share of labor force in agriculture and small area estimates of consumption from the Socioeconomic and Caste Census of 2012; and (viii) administrative data from the implementation of India's national rural roads program. 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. A key contribution of the SHRUG is a set of time-invariant location identifiers, which are designed to be forward compatible with future censuses. This will substantially facilitate the posting of data in a geographic frame that is useful for other researchers.

A final important characteristic of the SHRUG is that it is released under the Open Data Commons Open Database License (ODbL), which is a copyleft license.[2] This requires that all researchers who use the SHRUG post their own linked data with SHRUG identifiers at the time when their research papers are accepted for publication. Release SHRUG under this license helps to set a norm of data availability and will motivate researchers to release their data in a format that is widely usable by others. Requiring data posting at the time of publication is analogous to the limited protection period offered by a patent; the present publication lags in Economics will still give researchers plenty of time on any further projects with that data. Data under written proprietary contracts that restrict sharing are excluded at this time.

This paper describes the construction and contents of the SHRUG (Sections 2 and 3), and the strengths and weaknesses of this dataset relative to traditional sample datasets like the NSS or Annual Survey of Industry (Section 4). Section 5 concludes with a discussion of data to be added to the SHRUG in the future, the copyleft license of the SHRUG, 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

[2]More information is available at https://opendatacommons.org/licenses/odbl/index.html

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 appear separately, 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.[3]

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.[4] We provide keys that link shrids to the original population and economic census codes, make it easy to link data to the SHRUG as long as that data can be matched to any census year from 1990 to 2013; we plan to post keys to future censuses as they are released.

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 publishes 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,

[3]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. Shrids consisting of multiple villages or towns are given the code of the unit with the largest population.

[4]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.

the 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.[5] When villages and towns are pooled into larger units, the SHRUG reports both the village and town amenities for each unit.

While the SHRUG data package includes only a subset of village/town amenities that are consistently measured over time, the SHRUG keys make it easy to merge the data with the complete amenities tables from the population censuses, which can be downloaded from the Census of India.[6]

2.3 Economic Census

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.

We exclude jobs in the agricultural sector, which are inconsistently recorded by the Economic Census (both within rounds and across rounds). We also exclude jobs in public administration and defense (NIC2008 Section O), which were not counted in the 2013 Economic Census. This ensures comparability across time. However, we have kept the majority of public sector establishments, which include public schools, medical clinics, and state-owned enterprises.

The SHRUG data package describes total employment in each year, in every town and village. Researchers interested in the Economic Census microdata (which describes additional fields such as worker gender, source of finance and electricity, and more granular sectoral information), can readily obtain the Economic Census from the Ministry of Statistics and Pro-

[5]Census towns, unlike statutory towns, are defined as units with population greater than 5000 with 75% of the employed workforce outside of the agricultural sector.

[6]The SHRUG codebook provides sample code for matching the SHRUG to multiple rounds of external census data.

gramme Implementation (MOSPI), which has made it free and open. The SHRUG keys allow the raw Economic Census data to be linked in one step, with sample code in the codebook.

2.4 Administrative Data

At the time of writing, 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. The raw PMGSY data are at the level of the road or the habitation; there are typically one to three habitations in each village. Data have been aggregated to shrids, but habitation-level data is available in the data package released with (Asher and Novosad, 2019). These data were matched to PMGSY on the basis of 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 (2019).

2.5 Political Data and Legislative Constituency Aggregates

Electoral data are available for legislative constituencies. SHRUG v1.1 does not include electoral data for parliamentary constituencies. As with towns and villages, we created a set of time-invariant constituency identifiers for the 3rd and 4th delimitations. This was necessary because the Election Commission of India (ECI) does not always use consistent numeric identifiers over time. The keys provided with the constituency-level data make it possible to link to the ECI data and thus to any other dataset that uses those identifiers.

The election data were scraped, cleaned, and shared with us by Jensenius (2017), and are now maintained 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). Users of the election data in SHRUG should cite Jensenius (2017).

We have also included data from affidavits submitted by politicians contesting office. These include the number of open criminal charges that they face, a severity measure for those charges (the maximum years imprisonment of all charges), the politicians' reported assets and liabilities, age, and education. The data cover the period 2004–2018; these affidavits do not exist for periods earlier than 2003. These data come from the Association for Democratic Reform and the Electoral Commission of India. The data were cleaned (for the full period)and re-entered by hand (for 2004-2007) by Prakash et al. (2019), who should be cited when these data are used. The ADR data have been harmonized at the constituency level with the electoral and socioeconomic constituency data; some but not all candidate identifiers have been matched with the electoral data.

2.6 Remote Sensing Data

SHRUG v1.1 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, from which means can be readily constructed. We also include a night light series that is calibrated to adjust for differences between satellite sensitivity and degradation of satellite sensors over time (Elvidge et al., 2014). These are available as annual aggregates from 1994–2013.

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. (2019) and Townshend et al. (2011). As with night lights, we match these to location boundaries and report total tree cover and number of pixels in each unit.

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 in the data should researchers wish to exclude them.

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.

This was a multistep process that used a combination of algorithmic and manual matching. The Population Censuses in some cases (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.[7] 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. A 1991–2001 district correspondence was shared with us by Kumar and Somanathan (2015). We constructed the 2001–2011 district match based on the back-referenced village identifiers in the 2011 census, which provided a 2001 census village identifier for every 2011 village.[8] 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. Where the district and subdistrict maps indicated substantial changes in district and subdistrict boundaries, we matched villages and towns within higher level aggregates of districts and subdistricts.

To match names, we used a fuzzy matching algorithm that we developed (`masala_merge`),

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

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

which implements a Levenshtein edit distance algorithm which was modified to work with Indian transliterations. The Levensthein algorithm creates a distance between two words, which is the number of letter changes, insertions or deletions required to transform one word into the other. We modified this algorithm to apply smaller penalties to sets of 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 with the complete list of edit costs is posted on the web.[9] 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 MOSPI. MOSPI was not able to provide a location directory for the 1990 Economic Census. The EC district codes were the same as those used in the 1991 Population Census, but the lower level codes were different in some states. We worked with MOSPI to identify the set of states that used the same identifiers in the 1991 Population Census and the 1990 Economic Census. It was straightforward in the data to distinguish these states from the ones which created new codes, and we matched villages and towns on the basis of identifiers in these places.

For towns that could not be reliably matched on the basis of the town codes, we obtained a number of additional matches in situations where three conditions all held: (i) towns could be uniquely matched within districts to the 1991 population census based on the number of wards;[10] (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

[9] See https://github.com/paulnov/masala-merge.

[10] For instance, if a district had two towns in the 1991 Population Census, with respectively four and seven wards, we matched them to the 1990 Economic Census towns with the same number of wards.

than for the other censuses. Table 3 summarizes the share of employment in each Economic Census that is matched to the SHRUG, by state.

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 (2019).

3.2 Creating a Constituency-Level Panel

We matched villages and towns to legislative constituencies using geographic data 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 match rates over time. Rather than treating these errors as noise, we used an error-correction process described here.

Note that in the 2011 Population Census, we have matched 100% of towns and villages to constituencies, while the match rates in 1991 and 2001 are very high but imperfect. For each constituency, we therefore know the 2011 population in towns and villages that were matched to the other censuses, and the 2011 population in towns and villages that were not matched. We can impute the prior population in unmatched locations, by assuming that the within-constituency 2001–2011 population growth rate was the same in towns and villages that we did *not* match in 2001 as in the towns and villages that we *did* match. We can repeat the process to obtain the full set of populations in 1991. Because the match rates in 1991 and 2001 are so high, any error in this imputation process is likely to be trivial. This process will cause the aggregated constituency population to be much closer to the truth than if we counted missing locations as having zero population.

We then repeat the process to aggregate the employment count in the economic censuses, and the public goods counts in the town and village directories. For each economic cen-

sus, we assume that the employment-to-population ratio for missing locations is the same as it is for non-missing locations within the same constituency. For public goods that are aggregated with means rather than sums (such as the mean number of hours of electricity), we generate an aggregate based on the population-weighted mean in non-missing locations. To avoid excess dependence on imputed values, we set fields to missing in constituencies where the imputation would be needed for more than 25% of individuals in the constituency. This means that different constituencies may be missing different fields depending on the underlying structure of the data.[11]

Importantly, note that this imputation process applies only to the constituency-level data; when economic and population census data are missing in the town and village-level data, they are reported as missing in the SHRUG.

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 some villages cannot be unambiguously assigned to a single constituency. We 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 ward level or lower, it is difficult to identify the population or other characteristics of these constituencies — we

[11]Imputed values for constituencies with high imputation rates are available from the authors, as is the share of imputed data in each constituency-field. These are not included in the online SHRUG package because the files are extremely large and have relatively narrow usefulness.

know only the aggregate population of the combined constituencies.[12] We therefore exclude constituencies that include any part of towns that cross constituency boundaries. For instance, the Delhi Legislative Assembly has 70 constituencies, but Delhi is aggregated to a single location identifier in the SHRUG because its internal boundaries have changed several times in the 20 years. As a result, it is impossible to map constituency boundaries to shrid boundaries. We therefore excluded all of these constituencies from the constituency-level socioeconomic data, though they are included in the election data.

The constituency SHRUG is therefore not representative — in particular, it excludes large cities. However, we are not aware of any other research that measures or exploits socioeconomic data at the constituency level for large urban constituencies, because of the same boundary misalignment issue that we face here. 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 4th 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. We provide separate identifiers for the 3rd and the 4th delimitations; there is no correspondence between these as nearly all of the constituency boundaries were changed.

For the remote sensing data, we simply generated total night light and tree cover variables for each constituency-year using the geographic boundaries of the both the 3rd and the 4th delimitation constituencies.

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

4 Strengths and Weaknesses of our Approach

The SHRUG has two main advantages relative to other data sources. First, it describes a wide set of socioeconomic outcomes over a two decade period for the universe of locations at a much higher geographic resolution than any other Indian panel dataset with the same information.[13] This 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 administrative or remote sensing data source that is added to the SHRUG can be fully 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 SHRUG has three main limitations. First, not all villages and towns are matched in all periods. If a researcher's goal was to estimate the number of firms in India, for example, then aggregating from NSS samples is arguably a better approach, because there are no missing locations. Economic Census data in the SHRUG has a slight rural bias because rural boundaries are more easily tracked over time than urban boundaries.

Second, the SHRUG is only as good as the collection process for the administrative data that underlies it. The NSS enumerators spend far more time with each firm owner than the Economic Census enumerators, and have more quality checks and cross validations in their survey process. Some of the outliers in the Economic Census (and thus the SHRUG) are

[13]Other specialized data sources have high geographic resolution. For example, Prowess describes the operations of large firms at high spatial resolution, but does not have data on individuals. Note that the SHRUG can be easily linked to Prowess on a location basis, increasing the utility of both datasets.

almost surely incorrect; we offer some suggestions in the codebook on how to deal with these observations.

Finally, the breadth of any survey in the SHRUG is smaller than in the conventional sample surveys. Because the administrative censuses are implemented for every household and firm in India, they are necessarily based on shorter surveys. This disadvantage is balanced by the high geographic precision and wide breadth of data available for towns and villages in the other modules of the SHRUG. An NSS consumption survey is much more detailed than the SECC; but the short asset survey in SHRUG can be supplemented with data on night lights, forest cover, administrative programs, village public goods, and local firms.

Whether the strengths or the limitations dominate will depend on the particular research question. Research questions that rely on high-resolution geographic variation, or that require socioeconomic outcomes in units with political boundaries will be well suited for analysis with the SHRUG.

5 Conclusions: A Model for Collaborative Data Sharing

Most researcher-initiated data collection projects 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, 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 efforts 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.[14] Our aspiration is that the SHRUG can serve as a common geographic frame for

[14]Some examples include the NREGS public works and wage support program, the RGGVY rural

all these datasets, standardizing their location identifiers and increasing researchers' positive externalities on each other.

It is to this end that the SHRUG is released under a copyleft license that commits users to share any data that they link to the SHRUG under a similar license when their papers are accepted for publication. Researchers often face a tradeoff between sharing data (which enables more socially valuable research) and keeping data restricted (which ensures that they will not scooped on future projects with that data). Some balance between these objectives is needed; some private returns to developing new data sources are desirable as motivating factors.

A similar tradeoff motivates governments to provide patents to private individuals, allowing them to temporarily earn rents on their inventions but eventually ceding their designs to the public domain. We suggest a similar solution. The time lag to publication in economics all but ensurees that researchers who develop new data sources in tandem with the SHRUG will have a huge time lead on developing projects with that data.

It is naturally necessary to provide an exception for data that is obtained under a proprietary license that prohibits sharing, or data with personally identifiable information. The latter is unlikely to be important at the level of village aggregates; even proprietary microdata can often be released at more aggregated levels.

As a final point, because SHRUG amalgamates multiple sources of data, we ask users to cite all of the research papers that underlie those data sources. When data is downloaded from the SHRUG platform, we automatically generate a list of relevant citations. Doing so will further increase the returns to other researchers to invest in developing new data sources and making them easily usable by others.

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 researcher to have to reinvent the wheel, and limiting the scope of each research project to the amount of data that its research team is willing to clean.

References

Adukia, Anjali, Sam Asher, and Paul Novosad, "Educational Investment Responses to Economic Opportunity: Evidence from Indian Road Construction," *American Economic Journal: Applied Economics (forthcoming)*, 2019.

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," *American Economic Review (forthcoming)*, 2019.

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

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

Chhibber, Pradeep and Francesca R Jensenius, "Privileging oneâs own? Voting patterns and politicized spending in India," 2016. Working Paper.

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.*

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.

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.

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, Town, Constituency, District
Economic Census	1990, 1998, 2005, 2013	Employment and sector of all non-ag firms	Village, Town, Constituency
SECC	2012	Consumption estimates and agricultural labor share	Village
Election Results	1980–2018	Candidate name / party / votes	Constituency/Candidate
Politician Assets/Crime	2003–2018	Criminal charges, assets, liabilities, education	Constituency/Candidate
Night Lights	1994–2013	Proxy for electrification and economic activity	Village, Town, Constituency
Forest Cover	2000–2014	% Tree cover from Vegetation Continuous Fields	Village, Town, Constituency
Road Implementation	2000–2013	Administrative data from PMGSY	Village

Table 2

Population share matched to the SHRUG, by state

States	PC91	PC01	PC11
India	826117.89 / 833122.68 (99%)	1028120.50 / 1028349.73 (100%)	1209944.68 / 1210741.69 (100%)
Andaman Nicobar Islands	280.66 / 280.66 (100%)	356.15 / 356.15 (100%)	380.55 / 380.58 (100%)
Andhra Pradesh	65140.70 / 66455.27 (98%)	76210.01 / 76210.01 (100%)	84580.78 / 84580.78 (100%)
Arunachal Pradesh	621.18 / 637.04 (98%)	1097.97 / 1097.97 (100%)	1383.17 / 1383.73 (100%)
Assam	22278.90 / 22311.78 (100%)	26640.04 / 26655.53 (100%)	30999.61 / 31205.58 (99%)
Bihar	86119.25 / 86374.47 (100%)	82825.55 / 82825.55 (100%)	104099.45 / 104099.46 (100%)
Chandigarh	642.01 / 642.01 (100%)	900.63 / 900.63 (100%)	1055.45 / 1055.45 (100%)
Chhattisgarh		20827.74 / 20833.80 (100%)	25544.25 / 25545.20 (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	1155.51 / 1169.79 (99%)	1347.67 / 1347.67 (100%)	1458.55 / 1458.55 (100%)
Gujarat	41284.77 / 41309.58 (100%)	50671.02 / 50671.02 (100%)	60439.69 / 60439.69 (100%)
Haryana	16285.72 / 16459.98 (99%)	21139.38 / 21144.56 (100%)	25193.50 / 25351.46 (99%)
Himachal Pradesh	5165.07 / 5170.53 (100%)	6077.90 / 6077.90 (100%)	6864.45 / 6864.60 (100%)
Jammu Kashmir		10142.76 / 10143.70 (100%)	12539.86 / 12541.30 (100%)
Jharkhand		26945.83 / 26945.83 (100%)	32983.76 / 32988.13 (100%)
Karnataka	44663.16 / 44977.20 (99%)	52785.20 / 52850.56 (100%)	61032.42 / 61095.30 (100%)
Kerala	28631.18 / 29098.52 (98%)	31841.37 / 31841.37 (100%)	33406.06 / 33406.06 (100%)
Lakshadweep	51.71 / 51.71 (100%)	60.65 / 60.65 (100%)	64.47 / 64.47 (100%)
Madhya Pradesh	62281.73 / 63026.21 (99%)	60345.27 / 60348.03 (100%)	72626.81 / 72626.81 (100%)
Maharashtra	78363.48 / 78936.42 (99%)	96878.63 / 96878.63 (100%)	112323.51 / 112374.34 (100%)
Manipur	1806.38 / 1837.15 (98%)	2166.79 / 2166.79 (100%)	2851.43 / 2855.79 (100%)
Meghalaya	1764.66 / 1774.74 (99%)	2288.95 / 2318.82 (99%)	2961.91 / 2966.89 (100%)
Mizoram	689.54 / 689.76 (100%)	888.57 / 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	31515.51 / 31587.64 (100%)	36799.75 / 36804.66 (100%)	41945.54 / 41969.76 (100%)
Puducherry	771.56 / 807.78 (96%)	974.35 / 974.35 (100%)	1247.95 / 1247.95 (100%)
Punjab	19053.16 / 19053.16 (100%)	24334.90 / 24359.00 (100%)	27650.20 / 27743.34 (100%)
Rajasthan	43354.10 / 43879.50 (99%)	56502.28 / 56507.19 (100%)	68548.43 / 68548.44 (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%)	72117.59 / 72147.03 (100%)
Tripura	2430.67 / 2757.20 (88%)	3198.93 / 3199.20 (100%)	3666.08 / 3673.92 (100%)
Uttarakhand		166186.02 / 166197.92 (100%)	199763.41 / 199812.34 (100%)
Uttar Pradesh	138452.58 / 138837.84 (100%)	8479.34 / 8489.35 (100%)	10071.41 / 10086.29 (100%)
West Bengal	66929.92 / 67887.31 (99%)	80079.76 / 80088.56 (100%)	91085.92 / 91167.27 (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	43266.88 / 62211.08 (70%)	62851.43 / 70891.77 (89%)	79038.38 / 85388.85 (93%)	107639.65 / 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	4080.46 / 5263.04 (78%)	5742.84 / 6243.11 (92%)	8568.18 / 8991.79 (95%)	10492.67 / 11563.89 (91%)
Arunachal Pradesh	13.00 / 61.86 (21%)	48.80 / 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%)	3606.55 / 3665.87 (98%)
Bihar	2467.22 / 2915.64 (85%)	1715.85 / 2028.94 (85%)	2031.13 / 2096.17 (97%)	2929.19 / 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%)	1154.25 / 1377.39 (84%)	1800.44 / 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%)	81.42 / 81.42 (100%)
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%)	6143.60 / 6246.70 (98%)
Haryana	939.56 / 1190.77 (79%)	1052.97 / 1408.53 (75%)	1742.25 / 1950.83 (89%)	2811.10 / 2845.80 (99%)
Himachal Pradesh	324.97 / 357.05 (91%)	446.01 / 461.38 (97%)	543.54 / 552.25 (98%)	894.05 / 938.60 (95%)
Jammu Kashmir		100.83 / 430.17 (23%)	546.40 / 645.96 (85%)	1043.19 / 1065.65 (98%)
Jharkhand		866.09 / 947.85 (91%)	991.34 / 1030.31 (96%)	1377.32 / 1386.44 (99%)
Karnataka	3571.51 / 6339.23 (56%)	4069.62 / 4228.16 (96%)	5035.00 / 5165.28 (97%)	5790.34 / 5829.52 (99%)
Kerala	2223.42 / 2961.80 (75%)	585.07 / 3249.12 (18%)	2931.26 / 4309.21 (68%)	5649.97 / 5701.44 (99%)
Lakshadweep		5.87 / 12.18 (48%)	8.37 / 8.37 (100%)	9.92 / 10.24 (97%)
Madhya Pradesh	2867.56 / 3190.24 (90%)	3142.60 / 3325.93 (94%)	3274.40 / 3531.72 (93%)	4086.12 / 4241.05 (96%)
Maharashtra	7187.69 / 7577.37 (95%)	8134.96 / 8381.88 (97%)	9036.32 / 9526.52 (95%)	11797.80 / 11947.80 (99%)
Manipur	9.93 / 133.45 (7%)	109.61 / 167.68 (65%)	147.97 / 204.65 (72%)	353.88 / 385.92 (92%)
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%)	3003.82 / 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	1745.15 / 2203.52 (79%)	2687.16 / 2885.55 (93%)	3288.03 / 3569.26 (92%)	4897.19 / 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%)	148.40 / 218.62 (68%)	258.37 / 324.29 (80%)	379.29 / 382.24 (99%)
Uttarakhand		354.44 / 448.05 (79%)	7249.33 / 7328.97 (99%)	11377.23 / 11422.24 (100%)
Uttar Pradesh	5406.84 / 7505.02 (72%)	6045.04 / 6283.58 (96%)	564.74 / 619.01 (91%)	800.46 / 980.15 (82%)
West Bengal	3908.86 / 6539.10 (60%)	7588.40 / 7976.98 (95%)	8958.33 / 9277.06 (97%)	10988.06 / 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.