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

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

November 2019

Abstract

The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG) is a new data source that describes socioeconomic development in India. This first version of the SHRUG contains demographic, socioeconomic, firm and political outcomes at a high geographic resolution for the period 1990–2018. It incorporates administrative data covering the universe of Indian households and non-farm economic establishments for every urban and rural location in India, as well as data from successive population censuses, satellite-derived measures of economic activity and forest cover, data on elections and politicians, and other administrative datasets available at the local level. The SHRUG is not only a static dataset but also a platform for future collaboration and data sharing between researchers working with large-scale data in India. The backbone is a set of consistent location identifiers for all geographic locations in India from 1990–2018, along with a methodology to extend this classification to units from future data sources. Researchers can 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 analyses to show that the SHRUG is consistent with other data sources and to demonstrate the value of high resolution data.

JEL Codes: C81/O12

*Thanks to Teevrat Garg, Francesca Jensenius, Dan Keniston, and Nishith Prakash for sharing data that contributed to this dataset. This paper describes 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. All errors are our own.

†Johns Hopkins SAIS, sasher2@jhu.edu
‡Development Data Lab, lunt@devdatalab.org
§Northwestern University, ryumatsuura@u.northwestern.edu
¶Dartmouth College, paul.novosad@dartmouth.edu

1 Introduction

The computerization of government administration and the proliferation of new data sources such as satellite imagery has provided researchers with unprecedented opportunities to understand economic growth, poverty alleviation and service delivery in low-income countries. Yet most research in developing countries uses either highly targeted small-scale surveys or traditional sample surveys available only at high levels of spatial aggregation. One reason for this is that it remains enormously costly and time-consuming to discover, obtain, clean and merge data sources that were not designed for social science research. When researchers do pay the fixed costs, it is often difficult for other research teams to make use of that data; despite recent efforts to make data availability a key part of the publication process, published data is rarely comprehensive or easily linked to other data sources. The high fixed costs of making use of new data is particularly harmful to researchers lacking the funding and time to make major investments in data assembly, such as many doctoral students and researchers in the very countries that development economics seeks to study.

This paper describes the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). This is a new dataset that provides multidimensional socioeconomic information on the universe of cities, towns and villages in India from 1990 to 2018, a location panel with over 500,000 constant boundary geographic units. Data are also aggregated to legislative constituencies, making this one of the first multidimensional datasets with economic information at the level at which politicians are elected. In addition to giving researchers access to a wide range of data sources, the SHRUG reduces the technical and incentive barriers that prevent researchers from sharing their own data in a form that is useful to others. This project thus aims to convert the high fixed costs of data assembly from private investments into public goods for the entire research and policymaking communities.

The SHRUG differs from conventional sample datasets used to study socioeconomic changes in developing countries along several dimensions. First, the SHRUG is a high geographic resolution census. By covering every one of India's 500,000 towns and villages, it allows researchers to characterize local development and take advantage of sources of variation in ways that are not possible when using lower resolution datasets. India's flagship National Sample Survey, in contrast,

is representative only at the state or district level and does not repeatedly sample the same villages or reveal the locations of its sampling units. Researchers construct NSS panels 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. These surveys do not well capture the massive geographic variation in outcomes *within* districts, which we document in detail later in this paper.

Second, because the SHRUG is a compilation of diverse data sources, it allows for analysis of a wide range of research questions with minimal data merging costs. It incorporates data on political outcomes, politicians, firm outcomes, population demographics, measures of local consumption, remotely sensed measures of forest cover and economic activity, and administrative data on government programs. We are aware of no prior dataset describing such a large range of measures and covering the universe of villages and towns in any developing country.

Third, the SHRUG is a platform built for growth. Consistent location identifiers across time enable researchers to easily link additional data sources from any period. These identifiers are designed to be forward-compatible as administrative boundaries continue to evolve. As future researchers link data to the SHRUG and share their own data with SHRUG identifiers, the core dataset will grow in dimension and become an even more useful public good. We describe several strategies to raise the likelihood that future researchers will be able and willing to make meaningful contributions.

There are at least four ways in which SHRUG may have a comparative advantage over existing datasets:

1. Substantial variation in programs and socioeconomic outcomes 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 and exogenous variation may not be available.
2. The SHRUG permits time series analysis of socioeconomic characteristics at the level of India's 8000 towns. Growing cities are central to recent socioeconomic change in India, but there is little prior data available at the town level, other than for the very largest cities.

3. Few to none of India's major 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 national population censuses as sampling frames for new field experiments, but have limited additional data on sample locations before collecting their own baseline surveys. The SHRUG increases the scope of what is known at a local level. Field experimentalists can begin to test for divergent trends in field locations even before conducting a baseline survey.

The data underlying the SHRUG have 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.[2] Other authors have collected and merged administrative data for similar purposes (for example, Burlig and Preonas (2016), Chhibber and Jensenius (2016), Lehne et al. (2018), and Muralidharan et al. (2017)). This has often involved different researchers duplicating a very similar time-intensive matching process to that underlying the SHRUG. Going forward, the SHRUG can serve as a backbone for administrative data in India, substantially reducing the need for this kind of redundant work and making researchers' additional investments in administrative data available to a far wider audience.

The SHRUG is based on a combination of census and administrative 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

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

[2] Adukia et al. (2019) and Asher and Novosad (2019) evaluate the impacts of India's flagship rural roads program on transportation, labor markets, firms, agriculture, living standards, and school enrollment, drawing on data from more than 10 different component parts of the SHRUG. Exogenous variation in road completion is available only at the village level; it is much more difficult to identify causal effects of road completion using conventionally aggregated data. Asher and Novosad (2017) shows that politician identity affects local economic growth; this study relies on economic outcomes linked to political boundaries, a key contribution of the SHRUG.

have not previously been linked and aggregated into constant geographic units. To link locations, we performed fuzzy merges on the basis of names and identifiers at various levels of aggregation, and supplemented these matches with 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 because of the inconsistency in listed names across periods and the substantial number of units that have merged of split; we estimate that over 5000 person-hours of work were involved in linking and cleaning all of these datasets.

The current release of the SHRUG (version 1.4) 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 2018; (iv) assets, liabilities, and criminal charges of all politicians in office and many additional candidates from 2004 to 2017; (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. As new census, remote sensing and administrative datasets are released, the breadth of this panel will continue to grow.

1.1 SHRUG as Platform: Motivating Researcher Contributions to a Public Good

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. To date, the significant investments made by individual research teams in mobilizing these datasets have for the most part not resulted in substantial returns to scale, because the different data outputs have not been easily linkable.

There are both technical and institutional explanations for this socially inefficient equilibrium. The technical barriers to sharing data in a format useful to other researchers are substantial. The Indian government's limited digital resources for linking towns and villages across censuses make it difficult to work with administrative data across time periods, as does the limited availability of open shapefiles for villages, towns, and constituencies.

A key institutional barrier is that researchers who have invested in mobilizing administrative

data sources do not face strong incentives to publish their data in a highly usable format. The increasingly stringent data policies of economics journals have been effective in generating replication data, but that data is often posted without sufficient documentation or identifiers to be usable for other research projects. Replication files often describe only the researcher's final sample, which may be a small subset of the national data on which it is based. This is not the result of malign intent; rather, it can take weeks or months of effort to make replication data useful for future projects, and researchers may not see rewards for doing so.[3]

The SHRUG aims to mitigate both the technical and the institutional barriers to greater data sharing. To ease the technical barriers to data sharing, we have created a set of universal identifiers and key files that make it easy to plug in Indian datasets from any period between 1990 and the present. The codebook offers sample code to directly merge any population or economic census to the SHRUG in just a few lines of code.

The SHRUG has three characteristics that ease the institutional barriers to data sharing. First, our team at Development Data Lab plans to identify high value potential additions to the SHRUG and assist third party researchers in the steps required to transform a replication dataset into a public good. Second, researchers will be cited for their contributions to the SHRUG. When researchers download data from the SHRUG, they are required to cite the authors who created the component of the SHRUG that they are using. For instance, users of the criminal affidavit data in the SHRUG should cite Prakash et al. (2019). This citation structure gives researchers a substantial incentive to contribute, because clean data integrated into the SHRUG will receive far more users and generate more citations than data on journal web sites that is suitable for replication only.[4]

Finally, the SHRUG data are released to non-commercial users under a copyleft license (the Open Database License, or ODbL). This license commits researchers who link the SHRUG to non-proprietary data to also post that non-proprietary data in a complete form with SHRUG

[3]The low average quality of public replication data further creates a lemons problem: researchers who become accustomed to finding minimally useful data on journal web sites may be discouraged from discovering the gems that do exist.

[4]The present version of SHRUG already includes contributions from two groups of researchers other than ourselves. Since making the data public in September 2019, three additional groups of researchers are working on additional data sources that they plan to integrate with the SHRUG.

identifiers at the time that their research is published.[5] Given the long publication lags in economics, this license gives researchers ample lead time to work on additional projects with the data that they have collected, but commits them to sharing data in a reasonable time horizon. Like the limited protection time offered by a patent, this policy trades off the private incentive to researchers of develop unique data sources with the much greater public good of making those data sources available to the full network of researchers that can make use of them.[6]

This paper first describes the contents and construction of the SHRUG (Sections 2 and 3), as well the strengths and weaknesses of this dataset relative to traditional sample datasets like the NSS or Annual Survey of Industries (Section 4). Section 5 validates the SHRUG against external data sources and demonstrates the substantial local geographic variation in SHRUG that is missed by more aggregate data sources. Section 6 concludes with a discussion of data to be added to the SHRUG in the future, describes the copyleft license in greater detail, and proposes a framework for sharing of future data among members of the research community.

2 Data Contents

Table 1 summarizes the components of the SHRUG. The following subsections describe the different components in detail. Section 3 describes how they are linked together.

2.1 SHRUG Identifiers and SHRUG Keys

The main unit of observation in the SHRUG is a town or village, which is identified by a SHRUG identifier, or a *shrid*. A shrid describes a geographical unit that can be mapped consistently across all rounds of the Indian population and economic censuses from 1990 to 2013. In the majority of cases, a shrid describes a single village or town. When villages or towns have merged or separated in the sample period, we have aggregated them in the periods where they appear separately, such that the aggregation is represented by a single consistent shrid in all of the data. Some shrids are thus composed of multiple population census villages or towns, or a combination of villages and towns.

The use of a single consistent unit for each geographic location creates a critical simplification

[5] More information on the ODbL is available at https://opendatacommons.org/licenses/odbl/index.html.

[6] Data under formal proprietary contracts that restrict sharing are excluded from this commitment at this time.

for the researcher. Previously, tracking census units over time requires an inordinate amount of time studying duplications, merged and split villages, villages incorporated into towns, etc., all of which may not be consistently coded across different administrative datasets. In contrast, a location in the SHRUG is unique and consistent in all datasets and years. The SHRUG includes keys that link shrids to the original population and economic census codes and location names, making 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 will post keys to future censuses as they are released. Linking to these keys is easy and sample code is provided in the codebook.

2.2 Population Census and Amenities Tables

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

The Population Census also publishes village and town directories, which describe an increasingly large set of local amenities. The amenities include but are not limited to data on infrastructure (paved road, electrification, etc), the distance of villages to the nearest town, the presence of post offices and various medical facilities, the number of market days, and the sources of water. A different amenity list is used for census villages and census towns.[7] When villages and towns are pooled into larger units, the SHRUG reports both the village and town amenities for each unit.

The present SHRUG data package includes only a subset of the full list of village/town amenities that we have validated to be consistent 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.

[7]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.

2.3 Economic Census and the SHRUG Industry Codes

The Economic Census of India is a complete enumeration of non-farm economic establishments. Data is available for the third through sixth economic censuses, covering the years 1990, 1998, 2005 and 2013. The Economic Census includes all formal, informal, government, and private firms that are engaged in any sector other than crop production. 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.[8]

To track a consistent set of jobs over time, the SHRUG excludes jobs in the agricultural sector, which are inconsistently recorded by the Economic Census (both within rounds and across rounds). It also excludes (in all years) jobs in public administration and defense (NIC2008 Section O), which were not counted in the 2013 Economic Census. However, we have kept the vast majority of public sector establishments, which include public schools, medical clinics, and state-owned enterprises.

The SHRUG data package includes total employment in each census year in every town and village. The Economic Census microdata, which describes additional fields such as worker and owner gender, source of finance and power, and more granular sectoral information, can be easily linked to shrids using the SHRUG keys and the sample code included in the codebook. The Ministry of Statistics and Programme Implementation (MOSPI) has made these four rounds of the Economic Census (as well as the National Industry Classification Concordances) freely available online.

2.4 Program Administrative Data

The SHRUG includes administrative data from the PMGSY, the Prime Minister's Village Road Program, under which over 100,000 roads were built or improved between 2000 and 2017. 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

[8]The sector codes are based on the India-specific National Industry Classification system, which has changed several times over the sample period. We are developing a new industry classification that is consistent across the entire time series, which will be included in a future version of the SHRUG.

material, the pre-construction state of the road, and milestone dates (e.g. date of contract awarding, date of road completion), among others. The raw PMGSY data are at the level of the road or the habitation; there are typically between one and three habitations in each census 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).

The SHRUG also includes several aggregates from the Socioeconomic and Caste Census (SECC), a universal enumeration of household assets conducted in 2012. Asset censuses like this are commissioned by the federal government approximately every decade to determine individual eligibility for means-tested poverty relief programs. The SECC enumerates a list of household assets that can be rapidly assessed (including roof and wall material, number of rooms, and assets such as agricultural equipment, vehicles and mobile phones). The present version of the SHRUG includes two village-level aggregates from the SECC.

The first is a small area estimate of consumption, generated with the method of Elbers et al. (2003). Using the 2011-2012 IHDS-II (Indian Human Development Survey, 2011-12, available at `https://ihds.umd.edu`), we regressed total household consumption on a set of continuous and dummy variables that are equivalent to all asset and earnings information contained in the SECC. We then used the coefficients (shown in Table 5) to predict household-level consumption in the SECC microdata. We aggregate this to create a village-level measure of annual per capita consumption expenditure. [9]

Rural households directly report to the SECC whether their primary income source is from agriculture, small enterprise, wage work, or another source. We report the share of households in a village that draw their income from agriculture. Future versions of the SHRUG will include

[9]The coefficients from the model that generate these consumption numbers are estimated with statistical error. To allow researchers to account for these errors, we used a bootstrap approach, drawing households from IHDS with replacement and re-estimating village-level per capita consumption 1000 times. These 1000 draws are available as a SHRUG package for download and reflect the distribution of per capita consumption that arises from the first-stage estimation process. Researchers can use these 1000 draws in a second bootstrap process to account for the consumption estimation error. The Data Appendix in Asher and Novosad (2019) describes this process in additional detail.

additional asset variables from the SECC and small area estimates of urban consumption.[10]

2.5 Political Data and Legislative Constituency Aggregates

Electoral data are available for legislative constituencies. SHRUG 1.4 does not include electoral data for parliamentary constituencies, but these will be included in the future. 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 contributed by Jensenius (2017), who parsed them from official PDF files posted by the Election Commission of India. SHRUG includes these data in the format posted by the Trivedi Center for Political Data (TCPD), with the addition of unique and internally consistent SHRUG identifiers.[11] This includes turnout and vote totals for each candidate and party for all elections from 1980–2018. 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).

SHRUG also includes data from affidavits submitted by politicians contesting office, contributed by Prakash et al. (2019). 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 and are posted with SHRUG constituency identifiers, as well as the internally inconsistent ADR and Election Commission identifiers; some but not all candidate

[10] In the interim, users interested in more detailed asset data can use the SHRUG keys to match the SHRUG to the house listing of the publicly available 2011 Population Census, which provides village-level ownership data for a range of assets similar to the SECC.

[11] Data were downloaded in February 2019. We include the original TCPD codes to make it easy to link to future elections as they are posted by TCPD.

identifiers have been matched from the affidavit data to the electoral data.

2.6 Remote Sensing Data

SHRUG presently 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 a 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 flagged in the data should researchers wish to exclude them.

3 Data Construction

3.1 Matching the Population and Economic Censuses

The key challenge in creating time series administrative data in India is in dealing with changing unit boundaries. Faced with the challenge of villages being split, merged, and integrated with cities and towns, the decennial Census has opted to create new location identifiers in every decade since 1991. Further complicating the process of matching locations over time, district boundaries have changed substantially, with hundreds of new districts created between 1991 and 2011. The Census

provides digital keys to link villages and towns to prior censuses, but they are highly incomplete. The Census district handbooks contain detailed descriptions of boundary changes in narrative format only. All of these sources have errors and inconsistencies.

We used both the digital linking keys and the district handbooks to create the best possible correspondence of villages and towns across the 1991, 2001, and 2011 census.[12] We supplemented this with a custom fuzzy string matching program to match village and town names over time.[13] We conducted a hierarchical match from the largest to the smallest administrative units. We began 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 the majority of 2011 villages. 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. We validated the data using internal consistency checks and data from multiple sources, including geospatial village and town data assembled by other research groups. Table 2 summarizes the share of population from each population census that is matched at the village and town level to the SHRUG by state. Virtually all towns and villages were matched across the census periods: the match rates is 98% or higher for all but two small states in 1991.

To match the Economic Censuses to the Population Censuses, we used the location directories for 1998 and 2005, which were shared with us by the Ministry of Statistics (MOSPI). For 2013, we used the fact that the Economic Census location codes corresponded to the Population Census short codes, which were available with village and town names on the Population Census website. The final step in all these cases was a match using location names with the algorithm described above.

MOSPI was not able to provide a location directory for the 1990 Economic Census. The EC district

[12]The Census District Handbook is a 500+ page book describing all changes to boundaries in a given census district in each intercensal period. There is one book like this for each of India's ∼700 districts.

[13]The program is called masala-merge and is available at http://github.org/paulnov/masala-merge. It performs a Levenshtein string match, customized for common string substitutions used in Indian langages.

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[14]; (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. Table 3 summarizes the share of employment in each Economic Census that is matched to the SHRUG, by state. Because of the absence of the 1990 location directory, the match rate for the 1990 Economic Census is much lower than for the other censuses.

As noted above, a key innovation of the SHRUG is the creation of units with constant boundaries over time, which are not provided by the Census. When two locations are combined in any period, we combine them in all periods and aggregate their data appropriately. Similarly, if boundaries of units change, we create a new unit that is the smallest possible unit with unchanging boundaries. We call these units *shrids*. This process can involve combining villages, combining towns, or creating units that include villages and towns. In some cases, village and town boundaries have changed so dramatically that the aggregated constant boundary unit is quite large. In the case of New Delhi and Chandigarh, whose internal boundaries have changed frequently since 1991, the shrid is the entire city-state. In the case of Mumbai, a shrid is a district, which is the smallest non-changing unit. Creating time-consistent data within urban boundaries is a challenging and distinct project which we are also working on, but is beyond the scope of this iteration of the SHRUG.

Additional administrative datasets (such as the PMGSY road data and the Socioeconomic and Caste Census) were matched using a similar approach. These matches are described in more detail in Asher and Novosad (2019).

[14]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.

3.2 Creating a Constituency-Level Panel

We matched villages and towns to legislative constituencies using geographic data obtained from ML Infomap. 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 improving match rates over time. We corrected these errors with an imputation process which is 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 years' 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 it was 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 minimal. This process will cause the aggregated constituency population to be closer to the truth than if we counted missing locations as having zero population.

We then repeated the same process to aggregate the employment count in the Economic Censuses, and the public goods counts in the Town and Village Directories. For each Economic Census, we assumed that the employment-to-population ratio for missing locations is the same as it is for non-missing locations within the same constituency. For location amenities that are aggregated with means rather than sums (such as the mean number of hours of electricity), we generated 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 locations covering more than 25% of the population would be imputed. This means that different constituencies may be missing different fields depending on the underlying structure of the data.[15] Importantly, note that this imputation

[15]Imputed values for constituencies with high imputation rates are available from the authors, as is the share

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 they all use the same WGS84 projection and were obtained from the same firm. The misalignment is small—on the order of several hundred meters in the worst cases—but it is enough that some villages and towns cannot be unambiguously assigned to a single constituency. We dropped constituencies in which more than 25% of 2011 population 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 towns contain multiple constituencies. Because the Population Censuses do not report consistent identifiers at the ward level or below, it is difficult to identify the population or other characteristics of these constituencies — we know only the aggregate population of the combined constituencies.[16] We therefore exclude constituencies that include any part of towns that cross constituency boundaries. Constituencies in large urban areas are therefore missing population and economic data in the SHRUG. However, the election and remote sensing measures are still reported as the former are constituency-level data and the latter can be calculated directly from the spatial boundaries of constituencies.

The constituency SHRUG is therefore not representative — in particular, it excludes large cities. However, we are not aware of other research that measures or exploits socioeconomic data at the constituency level for large urban constituencies (other than the remote sensing measures described above), presumably due to the same boundary misalignment issues that we face here. Constructing this dataset using the ward maps for India's largest cities would be a valuable contribution that

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.

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

would enable better study of politics in India's growing cities.

Finally, India periodically redraws constituencies to account for population changes. The third delimitation came into effect in 1976 and the fourth in 2008, in the middle of the period covered by the SHRUG (Iyer and Reddy, 2013). This is not a problem for data construction, since constituencies are simply defined as polygons. We therefore match both sets of polygons to villages/towns and create separate complete constituency-level panels from 1990–2018 for the old and the new constituency delimitations. Researchers can thus make their own decisions regarding which polygons to use for which periods. We provide separate identifiers for the third and fourth delimitations; there is no correspondence between these as nearly all of the constituency boundaries were changed.

For the remote sensing data, we 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.

This release of the SHRUG does not include aggregates at the parliamentary constituency or the panchayat level because we have not yet created correspondences between towns/villages and these levels of aggregation. Given their obvious utility for research, we expect them to be added to a future version of the SHRUG.

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 long period for the universe of locations at a much higher geographic resolution than any other Indian panel dataset with broad scope.[17] This enables analysis of policies that vary at geographic units below the state or district level, such as politician identities and village-targeted programs.

Second, because of the comprehensive national geographic coverage of the data, the SHRUG

[17] Other specialized data sources have high geographic resolution but cover more narrow topics. For example, Prowess (maintained by the Center for Monitoring of the Indian Economy) describes the operations of large firms and contains headquarters addresses. Note that the SHRUG can be readily linked to Prowess at the village and town level, increasing the utility of both datasets. There are also several village-level projects that have created long time series' with broad and deep surveys, such as ARIS-REDS and ICRISAT. The data in these surveys is much more detailed than the data found in the SHRUG, but they cover fewer than three hundred villages, making them suitable for different kinds of projects.

will become more useful over 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 little 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. As noted above, many towns are missing economic data for 1990.

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 almost surely incorrect. This is inherent in the nature of the process of collecting data from hundreds of millions of respondents in a short time period. We offer some suggestions in the codebook on how to deal with these observations.

Finally, the length of most of the surveys underlying the SHRUG is smaller than in many other data sources. Because the administrative censuses are implemented for every household and firm in India, they are necessarily based on much shorter surveys than detailed sample surveys like the National Sample Survey and the Annual Survey of Industries. This disadvantage is traded off against the high geographic precision and wide breadth of data available for towns and villages across all of the modules of the SHRUG. An NSS consumption survey is much more detailed than the SECC; but the short asset survey in SHRUG can be analyzed in conjunction with data on night lights, forest cover, administrative programs, public goods and local firms, across several hundred thousand

villages with high resolution geography.

Administrative data by no means obviates the need for high quality sample field surveys. Whether the strengths or the limitations of the SHRUG dominate will depend on the particular research question. Research that relies on high-resolution geographic variation or rich location information, or that requires socioeconomic outcomes in units with political boundaries, will be particularly well-suited for analysis with the SHRUG.

5 Illustrative Analyses with the SHRUG

This section presents several empirical results that demonstrate the usefulness of the SHRUG and validate it against other data sources.

A central reason to work with high spatial resolution socioeconomic data is that state and district level datasets inherently miss a large proportion of the variation in economic outcomes across space. In Table 4 we decompose the variance in a series of town/village outcomes into state, district, and subdistrict components. Across all of these variables, a very large share of the variation is actually *within* districts and even within subdistricts.[18] For example, more than half of variation in average village household consumption occurs below the district level and 45% is below the subdistrict level. More strikingly, more than 80% of variation in rural per capita night lights is within subdistricts, and 80% of variation in urban light is within districts. We report similar decompositions for employment per capita, female labor share, population density and forest cover; forest cover is the only one of these fields where district identifiers explain more than half of the variation. And yet the vast majority of prior work on geographic patterns in Indian socioeconomic outcomes has focused only on cross-district variation; policies are often targeted at the district level as well. The primary reason for this has been an absence of easily accessible data below the district level; the SHRUG aims to remedy this absence.

A novel characteristic of the SHRUG is that it does not take a stand over the definition of an urban or rural space. There are many feasible and conflicting definitions of urban and rural spaces; researchers using the SHRUG can define these categories in flexible ways. Figure 1 shows the distributions of population density (per square kilometer) in the 2011 Population Census, disaggregated by

[18]Because most subdistricts have only one census town, we decompose town variation only to the district level.

urban and rural census classification. There is considerable overlap in the distributions, meaning that many rural areas are actually denser than many locations classified as urban. The 90th percentile of density for rural areas (1350 people per sq km) corresponds to the 20th percentile for urban areas. The SHRUG would allow researchers to use a population density definition of urban space, rather than being constrained to use the definition used by survey enumerators.[19]

The SHRUG is the first dataset to provide consumption per capita estimates for every village in India. We validate this consumption data in three ways. First, in Figure 2 we compare the distribution of village-level consumption per capita in the SHRUG, the IHDS, and the NSS. For the sake of comparability, the locations used are population census villages in the SHRUG and the closest analog in the other two datasets (first stage unit in the NSS and primary sampling unit in the IHDS). The figure shows that the SHRUG has a nearly identical mean to the IHDS, while the NSS is slightly lower than IHDS and SHRUG. The SHRUG consumption measure has lower variance than the other two datasets. These differences are mechanically related to the construction process of small area estimates through several channels. First, the SHRUG uses predicted consumption, and is thus lacking the error term found in the other datasets. Second, the observation counts per location in the NSS and IHDS are far lower than in the SHRUG, so outlier households have a smaller effect on the variance in village-level consumption in the SHRUG. Third, the SECC asset measures aim to identify poor households and do not include luxury goods. Households in the top percentiles of the consumption distribution are thus effectively topcoded in the SHRUG, which explains the absence of villages with very high per capita consumption.

In a second validation exercise, we match SHRUG districts to IHDS districts based on population census district identifiers, and compare their average consumption. Figure 3 presents a binscatter of the results, showing that average rural consumption covaries very strongly across these two datasets. As noted above, the relationship is weaker at the top of the consumption distribution where many SHRUG households are effectively topcoded.

Finally, we can decompose the difference in average consumption between the small area estimates

[19]The qualification is that the SHRUG is constrained to its underlying data. For instance, the village amenities table is generated by the Census only for locations classified as villages. However, we have aggregated as many SHRUG fields as possible across towns and villages, including some fields from these amenity tables, such as the number of primary schools, which are reported in both the village and town tables.

in the SHRUG and the IHDS measures. Table 5 breaks down the difference in availability of the consumption predictors used in the small area estimates between SHRUG and IHDS. Columns 1 and 2 show the share of households with each characteristic in both datasets and Column 3 shows the difference. Column 4 shows the the coefficient on consumption from the prediction regression in the IHDS. Column 5 shows the average consumption difference in Indian Rupees between the SHRUG and the IHDS that is driven by this component of the regression.[20] There are some large differences in the roof and wall material variables, suggesting that the two datasets classified materials slightly differently; but these differences cancel each other out and thus create little spread between SHRUG and IHDS. The sum of expected differences in column 5 suggests a reduction in consumption by approximately 1,000 INR from the IHDS.

Note that there is no mechanical association between consumption in SHRUG and consumption in IHDS or NSS. We used IHDS only to provide coefficients translating SECC assets to INR consumption values. The similarity in these distributions, except as noted above, gives us confidence that the consumption measures in the SHRUG are good proxies for these direct survey measures.

6 Conclusions: A Model for Collaborative Data Sharing

Most researcher-initiated data collection projects have a relatively narrow scope. A local survey is conducted for the purpose of an experimental or observational study, one or several research papers are written, and the data is reused only for replication or in rare cases, for long-term followup.

The existence of comprehensive administrative data makes possible a paradigm where investments in data 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 are already making use of administrative data in India and in other developing countries. In the absence of a common platform to link these datasets to each other, there is considerable duplication of effort and many potential complementarities and externalities across projects are not being realized.[21] Our aim in

[20] The average exchange rate in 2012 was 53 INR to USD.

[21] Some examples from India 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

building the SHRUG is to create a common geographic frame for all these datasets, standardizing their location identifiers and lowering the cost to researchers of creating positive externalities for each other.

Researchers often face a trade-off between sharing data, which enables more socially valuable research, and keeping data restricted, which ensures that they will not be scooped on future projects with that data. Some balance between these objectives is needed; private returns to developing new data sources are desirable as motivating factors.

In creating the SHRUG, we aim to both lower the transaction costs of sharing data and to change the institutional incentives around data sharing. Researchers who create data sources that are compatible with the SHRUG will have their work cited, because the data will be more valuable when linked with the range of fields in the SHRUG. To add weight to this incentive, the SHRUG is released under a copyleft license that commits users to share any data that they link to the SHRUG when their papers are accepted for publication. The time lag to publication in economics all but ensures that researchers who develop new non-proprietary SHRUG-link data sources will have a large lead on any other projects working with their data.[22]

The open source software universe has demonstrated that an equilibrium can exist where highly-skilled individuals freely share the fruits of their labor, creating valuable externalities but also receiving enough private benefits to make these contributions worth their while. The work underlying the SHRUG aims to build a set of norms, protocols, and institutions which would facilitate a similar equilibrium around the sharing of data for social science research.

As a final point, because SHRUG amalgamates multiple sources of data, we reiterate our request to users to cite all of the research papers that underlie those data sources. When data is downloaded from the SHRUG platform, a list of the citations underlying the download is automatically generated. Citing these appropriately will further increase the returns to other researchers to investing in developing new data sources and making them easily usable by others, creating positive externalities for all.

papers. And yet none of these programs (or research projects) have easily accessible or linkable 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.

[22]Even in the case of proprietary microdata, it is often possible to share village/town aggregates that strike a balance between respecting restrictions on data sharing and benefiting the wider research community.

6.1 Addendum: Updates and Contributions

The SHRUG will be regularly updated as new data is brought online and inevitable errors are found and corrected. The latest version will be maintained at the SHRUG web site, and all prior versions will be archived on the Harvard Dataverse, so that researchers can always replicate an exact prior download.[23]

Researchers wishing to make their data easily linkable to SHRUG need only post their data with unique SHRUG identifiers. We will aim to maintain a listing of all external datasets that can be linked directly to the SHRUG in this way. Following the data organization standards and protocols described on the SHRUG web site will assist in making these contributions as accessible as possible to outside researchers.[24]

Research teams that assemble national datasets at the shrid or constituency level that are of wide general interest and are extremely clean and consistent with the SHRUG format can request to have their data maintained and released directly via our web site, which will maximize that data's availability to others. We have made every effort to ensure that users downloading data from our site will appropriately reference all contributors to the specific datasets being downloaded, ensuring proper attribution.

[23]The SHRUG web site is http://devdatalab.org. The URL for the archived versions at the Harvard Dataverse is https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPESAK.

[24]Contribution protocols are described in more detail at http://devdatalab.org/shrug_contribute.

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 one's own? Voting patterns and politicized spending in India," 2016. Working Paper.

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

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

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "A Bright Idea for Measuring 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 amenities	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
Rural Road Construction	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 (numerator), the state-level population in the population census datasets (denominator), and the share of state-level population captured by the SHRUG, for all states and union territories in India. Population numbers are reported in thousands.

Table 3

Employment share matched to the SHRUG, by state

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 (numerator), the state-level employment in the economic census datasets (denominator), and the share of state-level employment captured by the SHRUG, for all states and union territories in India. Employment numbers are reported in thousands.

Table 4

Geographic variance decomposition of SHRUG variables

Panel A. Urban

	State	District
Nonfarm employment per Capita	0.102	0.254
Night lights per capita	0.056	0.203
Population density	0.162	0.304
Female labor share (EC13)	0.394	0.609
Average forest cover	0.465	0.736

Panel B. Rural

	State	District	Subdistrict
Consumption per capita	0.329	0.462	0.538
Nonfarm employment per Capita	0.021	0.044	0.075
Night lights per capita	0.072	0.141	0.181
Population density	0.202	0.290	0.338
Female labor share (EC13)	0.069	0.133	0.196
Average forest cover	0.472	0.640	0.729

Table 4 presents the spatial decomposition of the variance of a selection of variables in the SHRUG. Stated values are the R^2 of regressions of each variable on a set of, respectively, state, district, and subdistrict level fixed effects. Employment per capita is calculated by dividing 2013 Economic Census (EC13) nonfarm employment by the total population from the 2011 Population Census. Female labor share (EC13) is the total number of female EC13 jobs divided by total jobs. Night lights per capita is calculated by dividing calibrated total light from 2013 by the 2011 Census population. Average forest cover is a measure derived from MODIS VCF. Population density is the 2001 total population divided by 2001 Census land area.

Table 5
Asset Decomposition of Small Area Consumption Estimates

	(1) IHDS	(2) SECC	(3) Difference	(4) Coefficient	(5) Delta
Income 5000-10,000 Rs	0.12	0.15	0.03	-115.28	-3.99
Income Above 10,000 Rs	0.06	0.08	0.02	11441.27	216.24
Home Ownership	0.99	0.96	-0.03	-7641.37	206.32
Kisan Credit Card	0.07	0.04	-0.03	-1169.17	36.01
Land Ownership	0.61	0.52	-0.09	1989.43	-173.88
Number of Rooms in Home	2.60	2.23	-0.37	-1190.14	437.97
Both Mobile and Landline	0.03	0.02	-0.02	16209.44	-270.70
Landline Phone	0.01	0.01	0.00	25436.31	78.85
Mobile Phone	0.68	0.62	-0.06	8129.20	-461.74
Refrigerator	0.11	0.08	-0.04	7099.50	-274.75
Brick Roof	0.05	0.07	0.02	-2721.83	-63.69
Concrete Roof	0.12	0.19	0.07	5689.37	375.50
GI Roof	0.16	0.14	-0.01	1727.17	-19.17
Grass Roof	0.23	0.19	-0.04	1518.06	-55.41
Plastic Roof	0.00	0.02	0.02	17800.27	267.00
Slate Roof	0.05	0.04	-0.01	4567.51	-63.49
Stone Roof	0.08	0.04	-0.04	3578.95	-147.45
Tile Roof	0.12	0.30	0.17	428.18	74.89
Four Wheeled Vehicle	0.02	0.02	0.00	31466.08	-22.03
Two Wheeled Vehicle	0.17	0.15	-0.02	5460.41	-122.86
Brick Walls	0.27	0.35	0.08	9123.88	740.86
Concrete Walls	0.24	0.03	-0.21	12508.78	-2685.64
GI Walls	0.01	0.01	-0.01	12169.84	-74.24
Grass Walls	0.07	0.14	0.07	6588.86	448.70
Mud Walls	0.33	0.34	0.01	4186.25	47.30
Plastic Walls	0.00	0.01	0.00	35857.69	179.29
Stone Walls	0.05	0.10	0.05	4972.48	243.15
Wooden Walls	0.01	0.02	0.00	7869.09	38.56

Table 5 presents a comparison of IHDS and SECC covariates that were used to generateper capita consumption small area estimates in the SHRUG. The IHDS and SECC columns indicate the value for each covariate in the SECC and IHDS surveys taken at the village level; because the SECC is a census, no weights were required, while the IHDS required the use of sampling weights. Column 3 presents the difference between the two. Column 4 shows the coefficient for each covariate when regressing per capita consumption on the set of covariates in the IHDS. Column 5 multiplies column 4 by column 3, representing the expected difference in per capita consumption between IHDS and SHRUG that is explained by that covariate.

Figure 1
SHRUG Urban and Rural Population

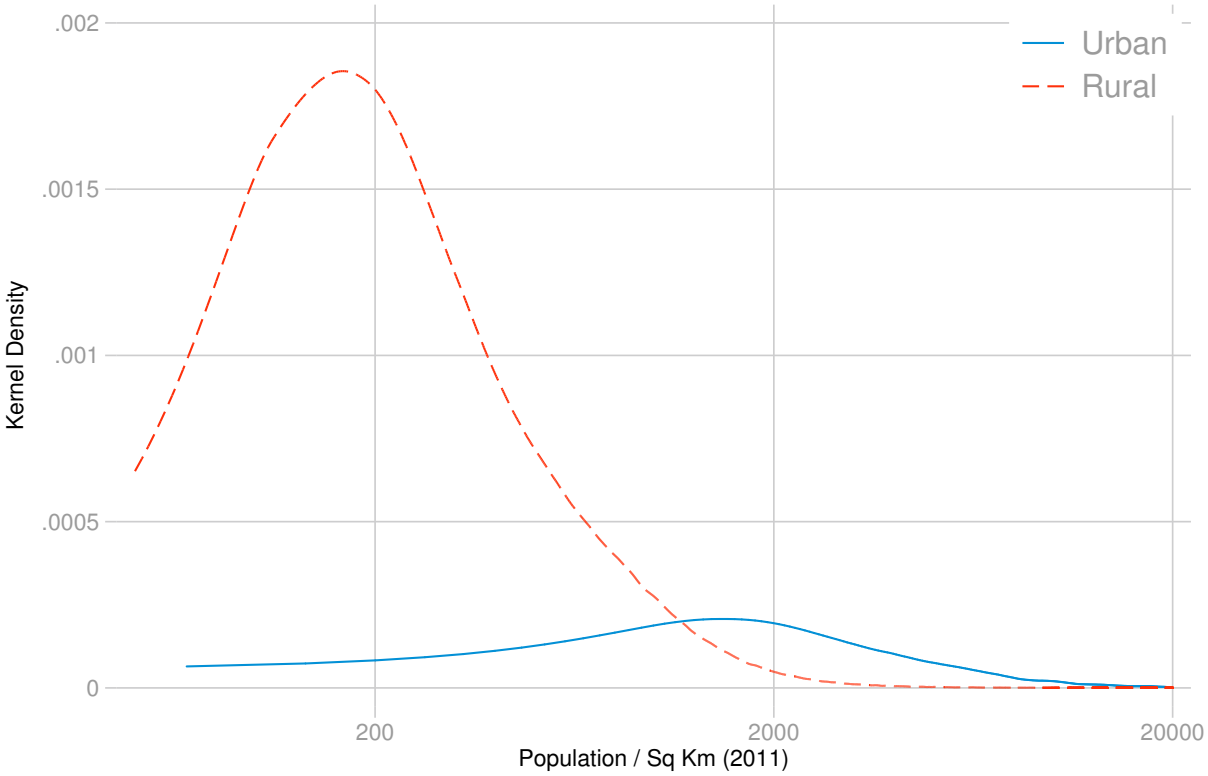

Figure 1 shows the distributions of population density (per sq km) in the 2011 Population Census, disaggregated by urban and rural census classification. For the purposes of this figure, shrids that contain both rural and urban areas in the 2011 Population Census are classified as urban.

Figure 2
SHRUG, IHDS, and NSS Consumption Per Capita

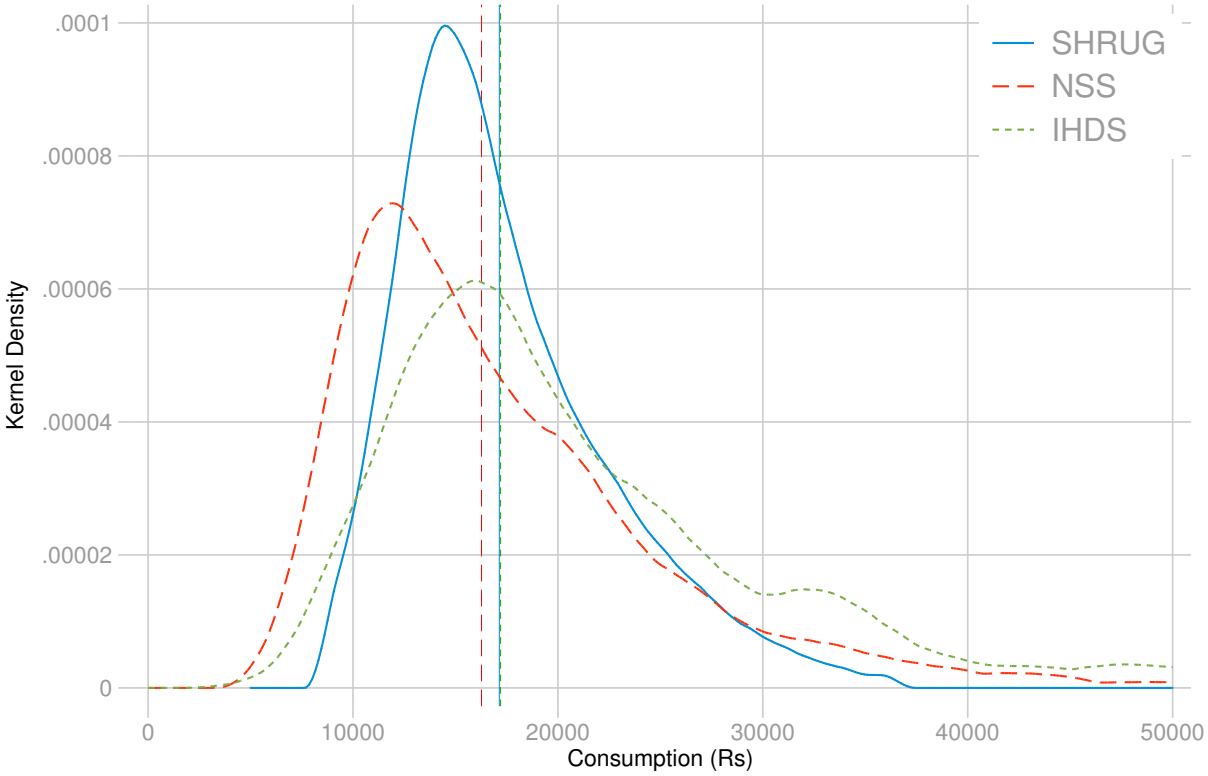

Figure 2 shows the distribution of location-level consumption per capita from the NSS, the IHDS, and SHRUG. The sample has been restricted to districts where the IHDS, NSS, and SHRUG could be matched. Consumption is measured by the weighted mean at the village level in each of the three datasets.

Figure 3

District-level SHRUG Consumption vs. IHDS Consumption

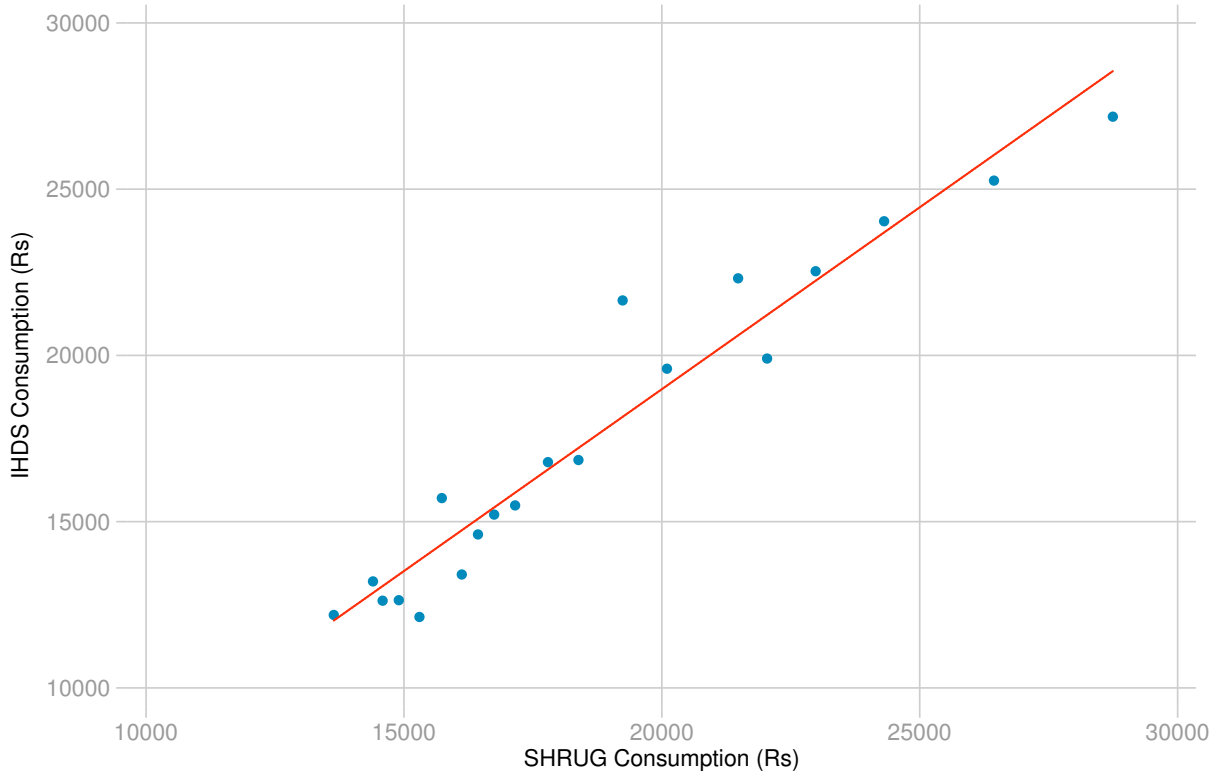

Figure 3 is a binned scatterplot of district-level average per-capita consumption in the IHDS and the SHRUG, weighted by the count of individuals in each survey. The SHRUG sample has been restricted to districts where the IHDS and SHRUG could be matched.