Development Research at High Geographic Resolution: An Analysis of Night Lights, Firms, and Poverty in India using the SHRUG Open Data Platform*

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

September 2020

Abstract

The SHRUG is an open data platform describing multidimensional socioeconomic development across 600,000 villages and towns in India. We present three illustrative analyses only possible with high-resolution data. We confirm that nighttime lights are highly significant proxies for population, employment, per-capita consumption, and electrification at very local levels. However, elasticities between night lights and these variables are far lower in time series than in cross section, and vary widely across context and level of aggregation. Next, we show that the distribution of manufacturing employment across villages follows a power law: the majority of rural Indians have considerably less access to manufacturing employment than is suggested by aggregate data. Third, we perform a poverty mapping exercise, identifying the targeting improvement from allocating programs at village rather than at district level. The SHRUG can serve as a model for open high-resolution data in developing countries.

JEL Codes: R11/C81/O12

*Thanks to Teevrat Garg, Francesca Jensenius, Dan Keniston, and Nishith Prakash for sharing data that contributed to this dataset. This project and underlying data platform were funded by Emergent Ventures. Early work on this project was 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. The SHRUG dataset can be downloaded at http://devdatalab.org/shrug.

†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

Development and economic growth, insofar as they affect the living standards of individuals, are highly localized phenomena. Employment opportunities, schools, public goods, and welfare schemes are relevant to individuals primarily if they are accessible within a short distance. However, due to the nature of data available to researchers, analysis of development has largely occurred at an aggregate geographic scale. In India, much policy and research has been based on the National Sample Surveys, which have tracked socioeconomic change at the district level since independence. But a single Indian district is home to around 2 million people, often with vast variation in living standards. This variation is demonstrated in Table 1, which shows the share of variation in various development outcomes at different levels of geographic variation. 55% of variation in mean village per capita consumption and over 90% of variation in rural non-farm employment occurs below the district level. The National Sample Surveys often base district statistics on fewer than 100 households per district, missing much of this microgeographic variation.

The computerization of government administration and the proliferation of new data sources such as satellite imagery has provided researchers with the opportunity to study development at higher geographic resolution than ever before. However, it has been 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 introduces the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG), a multidimensional socioeconomic data platform with high geographic resolution data on demography, non-farm employment structure, political outcomes, forest cover, night lights, and administrative program operation, among other dimensions, on the universe of cities, towns,

villages, and legislative constituencies in India from 1990 to 2018 (see Table 2). The SHRUG is an open dataset but also an open data platform that is structured to facilitate and encourage sharing of readily linkable data between researchers in the future.

In this paper, we present three analyses that shed light on the distribution of economic opportunity within highly localized regions, analyses that are only possible with high resolution data like the SHRUG.

We begin by examining the effectiveness of nighttime lights as a measure of local development. Nighttime lights, as measured from outer space, have been widely used in recent years to examine patterns of development in places where local economic data is not available. Nighttime lights are particularly used in analyses of programs which have geographic variation which is more precise than what can be observed in aggregate data; the ability to observe an economic outcome in a 1km x 1km cell is a key comparative advantage of night lights. Given the lack of high-precision economic data in most developing countries, the assumption that night lights have the same elasticities at both very high and very low levels of geographic aggregation with a range of outcome variables is largely untested.[1]

Our results confirm that night lights are a highly statistically significant log-linear proxy for a range of development outcomes—population, employment, per capita consumption, and electrification—at a very narrow geographical level, even in specifications with regional fixed effects, in the both the cross section and in the time series. However, there are two significant caveats. First, because night lights have independent correlations with each of these outcomes, it is difficult to tell from night lights alone which of these proxy variables are being measured. Researchers have used night lights as a proxy for GDP growth, cross-sectional GDP, urban extent, public expenditure, and electricity supply and demand, among other variables (Henderson et al., 2011; Baum-Snow et al., 2017; Bleakley and Lin, 2012; Min et al., 2013; Hodler and Raschky, 2014; Baskaran et al., 2015; Burlig and Preonas, 2016; Harari, 2020; Mahadevan, 2019). Our results suggest that all of these proxies are valid, but researchers have no guarantee that night lights are proxying the specific variable that

[1]Henderson et al. (2011) show that elasticities are similar across countries and across large regions within countries, but they do not measure elasticities at finer geographic levels. And yet the primary use of night lights in research is precisely to generate outcome variables at very high geographic resolution.

they would like to measure and not a different one.

Second, we find that night light elasticities vary substantially across the level of geographic aggregation and across contexts. Importantly, we find that night light elasticities are an order of magnitude smaller in the time series than they are in the cross section. Night light elasticities with most variables are also considerably higher in electrified villages than in non-electrified villages. This finding suggests that researchers should use extreme caution in applying regional luminosity-to-GDP elasticities from the literature to microgeographic contexts without additional information, because these elasticities may be wildly different. Applying cross-sectional elasticities to time series night light coefficients is likely to result in a substantial upward bias on GDP estimates.

Moving beyond night lights, we next examine the distribution of manufacturing and services employment across space. Economic activity is often highly agglomerated and firm size distributions are known to follow power laws (Ellison and Glaeser, 1997; Gabaix, 2016). The amount of concentration affects the extent to which regional averages reflect the access to opportunity of the majority of a region's residents. We show that the distribution of firms across space also follows a power law, such that some regions have orders of magnitudes more employment density than others. The extent of employment concentration is higher for manufacturing than for services firms, and the highest by far for manufacturing employment in rural areas. The mean Indian village has 15 manufacturing jobs per 1000 people, but the median has less than 5. Almost 80% of rural Indians live in villages with fewer manufacturing job opportunities than the rural mean. Regional employment opportunities, especially for women, may be of limited relevance if they are not geographically proximate. This analysis demonstrates that analysis of employment at aggregate geographic scales may fail to register the lack of labor market opportunities for most rural workers.

Finally, we conduct a poverty mapping exercise, considering the problem of a policymaker targeting a program to poor geographic regions (Ravallion, 1993; Baker and Grosh, 1994; Bigman et al., 2000; Bigman and Srinivasan, 2002; Elbers et al., 2007; Brown et al., 2019). We measure the efficiency gain that could arise from the ability to allocate a program reaching 25% of the population at increasingly finer geography. If the policy-maker can observe poverty only at the district level

(n=618), then allocating the program to the poorest districts would cover 37% of the rural poor. Allocating the same program at the village level would cover 44% of the rural poor. There is likely a tradeoff between precise geographic targeting and administrative cost; highly localized data makes it possible to estimate the parameters of this tradeoff.

The SHRUG has a number of strengths and weaknesses as an analytic platform. Its primary strengths are its geographic resolution and consistent location identifiers over a period of dramatic economic change. There are few other data sources in developing countries that allow tracking of public goods, economic activity, politician characteristics, election results, and many other outcomes at the village and town level for almost 600,000 consistently-bounded geographic units.[2] The key comparative disadvantages of the SHRUG are that (i) most of the underlying data sources are based on narrow surveys with a few dozen questions rather than hundreds; and (ii) the SHRUG describes locational aggregates, rather than individual or firm microdata.[3] Researchers interested in detailed characteristics of individual households and firms may thus be better served by traditional data sources.

There are a number of contexts in which the SHRUG may have a comparative advantage:

- Programs with highly local variation are difficult to evaluate with aggregate data. Examples of analysis of such programs using administrative data (some of which are now in the SHRUG) include Lehne et al. (2018), Chhibber and Jensenius (2016), Burlig and Preonas (2016), Asher and Novosad (2017), Adukia et al. (2019), Asher and Novosad (2020), and Muralidharan et al. (2017).
- Analysis at the city/town and legislative constituency level is challenging because neither of these units are identified in India's conventional data sources. SHRUG data are aggregated to both of these levels.
- 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

[2]There are approximately 605,000 towns and villages in India. We collapse these to about 590,000 unique locations that can be tracked over time.

[3]Note that it can be rapidly linked to firm-level data in the open Economic Census, which is openly available.

before conducting a baseline survey.

Most of the data underlying both the SHRUG and the papers using administrative data described above are public. The primary constraint to using data like these is the time input required to assemble, clean, and match them to other data sources. The returns to this activity are increasing in the number of other localized data sources that they can be matched to. We therefore hope that the existence of the SHRUG makes it worthwhile for researchers to assemble and share other administrative and microgeographic datasets describing additional dimensions of development in India.

We structured the SHRUG to maximize the ease and benefit to researchers of sharing new data. First, by providing a broad spectrum of data under a consistent set of identifiers, we hope to create a standard set of location identifiers for India. Such standardization can lower the cost of merging datasets for all researchers. Second, we have structured downloads from the SHRUG to credit individual contributors. When researchers download data from the SHRUG, they are asked to cite the authors who created each component of the SHRUG that they are using.[4] This citation structure is intended to give researchers an incentive to contribute, because clean data integrated into the SHRUG will receive more users and generate more citations than data on journal web sites that are suitable for replication only.[5]

Finally, SHRUG data are released to non-commercial users under a copyleft license (the Open Database License, or ODbL), which commits researchers who link the SHRUG to non-proprietary data to also post that non-proprietary data in a complete form with SHRUG identifiers at the time that their research is published.[6] 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 approach trades off the private incentive to researchers of developing unique data sources with the much greater public good of making those data sources available to the full

[4]For instance, users of the data in the SHRUG on the criminal charges facing Members of Legislative Assemblies should cite Prakash et al. (2019).

[5]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.

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

network of researchers that can make use of them.[7]

The current release of the SHRUG (version 1.5) describes: (i) demographic and public goods data on every town and village in India from 1991 to 2011; (ii) employment and location of every non-farm firm in India from 1990 to 2013; (iii) legislative election results from 1980 to 2018 (Jensenius, 2017); (iv) assets, liabilities, and criminal charges of all politicians in office and many additional candidates from 2004 to 2017 (Prakash et al., 2019); (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.

This paper first describes the construction of the SHRUG in some detail in Section 2, along with several validation exercises. Section 3 presents the analyses of night lights, agglomeration, and poverty. Section 4 discusses the strengths and weaknesses of the SHRUG as an analytical tool. Section 5 concludes with a discussion of the copyleft license and framework for improving institutions of data sharing between researchers.

2 Data: A 25-year Multidimensional Panel of Locations

All variables in SHRUG are disaggregated to the level of the village (n=590,648), the town (7528), and the state legislative constituency (n=2800), covering the entire country. The different components are summarized briefly in Table 2 and in detail in the SHRUG metadata table.[8]

All of these data sources were initially generated with diverse geographic identifiers which were not straightforward to link. The Population and Economic Censuses are published with town and village codes with incomplete correspondences across different rounds. PMGSY data is published at the habitation level; most villages consists of between one and three habitations. Remote sensing data is published in grid cells, while political data is published at the constituency level; constituencies are approximately the size of subdistricts (a standard unit in the Population Censuses), but with

[7]Data under formal proprietary contracts that restrict sharing are excluded from this commitment at this time.
[8]The SHRUG metadata table can be found at devdatalab.org/shrug-metadata.

different and overlapping boundaries.

Any pair of these datasets can be reconciled and matched to each other, albeit with significant labor input. Sources of information used to match data include numeric identifiers, names, maps, data contents, and external data sources. In most cases, none of these information sources provide a perfect match. Identifiers are rarely consistent across datasets; sometimes they are consistent across some states but not others. Locations may be known by multiple names, and names may change over time; they virtually always have different spellings across and even within data sources. Maps are supplied with different projections and different degrees of error; boundaries which should be identical in different mapping sources often are not. Locations themselves change, splitting, merging, and realigning boundaries in complex ways.

The process of matching these datasets consists of developing algorithms to use as many of these information sources as possible, and then tuning those algorithms based on manual verification of samples of data. Individual verification of every match is however not feasible when observations number in the hundreds of thousands. Some units cannot be matched even with manual verification because there is insufficient information available. Some degree of incorrect matches are inevitable.

The core contribution of the SHRUG is a set of universal location identifers (one at the town and village level, and one at the constituency level) that span the entire sample period from 1990–2018. All of the datasets above can be rapidly linked to these identifiers in every period. When using the SHRUG, linking disparate data sources takes seconds rather than months. By releasing the data with these universal identifiers, our hope is to make them a standard for economic research in India, allowing future data sources to be linked to the data sources described here with ease.

This section describes the process of creating the universal identifiers that are at the heart of the SHRUG. The definitions of individual data fields are described in detail in the SHRUG codebook.[9]

To keep its size reasonable, the SHRUG 1.5 data package includes only a subset of the fields in the Economic and Population Censuses. We provide linking keys to each of these source datasets to make it easy for users to bring in additional fields from these open data sources.

[9]The SHRUG codebook can be found at http://www.devdatalab.org/shrug.

2.1 Building the Town and Village SHRUG

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 censuses.[10] We supplemented this with a custom fuzzy string matching program to match village and town names over time.[11] 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. Appendix Table A1 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

[10]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.

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

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 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;[12] (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. Appendix Table A2 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.

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 (2020).

Location splitting and merging over time results in an inordinately complex set of linking keys. To create a dataset that was unique on the same locations for all of the different underlying data sources, we aggregated location units until a set of boundaries could be found that was unique across all years. We gave each of these units a unique SHRUG identifier, or a *shrid*.

[12]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.

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.

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.

The use of a single consistent unit for each geographic location is a critical simplification for the researcher. Without a one-to-one relationship between locations across datasets, each additional bilateral merge requires an increasingly complex task of handling splits, joins, and unit duplications. When working with shrids, any number of datasets can be merged without increasing complexity.

The complete keys linking shrids to their original population and Economic Census codes and location names are posted with the SHRUG, making it easy to bring in additional data that is already linked to Population Census codes. As described in the codebook, we have designed a naming convention for shrids that will be forward-compatible with future censuses as well. New keys will be posted as future censuses are released.

2.2 Creating a Constituency-Level Panel

The boundaries of India's 543 parliamentary and some 4000 legislative constituencies do not align with the aggregate administrative boundaries used by India's data collection agencies. To create a constituency-level dataset with economic outcomes, we matched villages and towns to constituencies

using digital maps and collapsed the data to the legislative boundaries.[13] We created a set of time-invariant constituency identifiers for the 3rd and 4th delimitations.[14]

Creating a constituency-level panel from town and village microdata poses a number of challenges. First, because of the fuzzy matching process, there are some villages which were matched to some 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 we describe here.

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 impute the prior years' population in unmatched locations by assuming that the within-constituency 2001–2011 population growth rate is the same in towns and villages that we did *not* match in 2001 as it is 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 repeated 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 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

[13]SHRUG 1.5 does not include electoral data for parliamentary constituencies or panchayats, but these will be included in the future. Boundary data was obtained from ML Infomap.

[14]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 population would be imputed. This means that different constituencies may be missing different fields depending on the underlying structure of the data.

Appendix Figure A1 runs a simulation to calculate error rates associated with the imputation process. We randomly drop additional locations from the 2001 Population Census and then compare the calculated constituency population to what would be observed if no locations were dropped. When 25% of shrids are dropped (the upper limit of what we include in the data), the average constituency population is within 2% of the true value with a slightly negative bias of 0.015%.[15] It is important to note that this imputation process applies only to the constituency-level data; when Economic or Population Census data are missing in the town and village 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, we aim to integrate them 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

[15]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. While it would be desirable to use these values to econometrically adjust for the partial imputation that took place in calculating these values to adjust for the partial imputation of some constituency values, we are not aware of an out-of-the-box method that is well-matched to this partial imputation structure. The simulated error rates suggest that this source of error is small relative to other sources of error in the collection of administrative data in India, and thus not a major concern.

[16]The Population Census reports data at the ward level, but the wards change across rounds and do not necessarily

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 included for all villages because they are not aggregated from village and town data.

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 such a dataset using 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 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.

2.3 Remote Sensing Data

At present, SHRUG includes data from two remote sensing sources: nighttime luminosity (Henderson et al., 2011) and forest cover (Asher et al., 2019; Townshend et al., 2011). Nighttime luminosity data are from the National Oceanic and Atmospheric Administration (NOAA), and provide a luminosity value from 0–63 for each 1 km x 1 km grid cell. 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.

Gridded data from both of these sources were matched and aggregated to village, town, and

share boundaries with constituencies.

constituency boundaries. About 90% of towns and villages were georeferenced with polygons, permitting accurate measurement of night lights and forest cover. About 10% of locations, especially in the Northeast, were georeferenced in the mapping data only by points; we constructed Thiessen polygons to match these to the forest cover and night light rasters. Locations mapped with Thiessen polygons are flagged in the data.

2.4 Imputation of Consumption Expenditure

We created a measure of village- and town-level consumption expenditure using data 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).

We used this asset data to generate small area estimates of consumption expenditure, following 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 Appendix Tables A3 and A4) to predict household-level consumption in the SECC microdata. We aggregate this to create a village- and town-level measure of annual per capita consumption expenditure. Because the urban and rural asset lists are different, we use separate estimations depending on whether households were identified as rural or urban in the SECC.[17] Village- and town-level asset ownership shares are also available in the house listing for the 2011 Population Census and match the asset shares in the SHRUG closely.

In addition to average per capita consumption, we included the share of individuals in each town

[17]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 (2020) describes this process in additional detail.

and village who live in households with per capita consumption below the 2012 national poverty rate, defined in two ways: below $2, or below Rs. 27.2 per day in villages and Rs. 33.33 in cities and towns.[18]

We have validated these consumption estimates in three ways. First, we affirmed that the consumption distribution broadly matches the 2011–12 IHDS and 2012 NSS. The consumption distribution in each dataset is shown in Figure 1. 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).[19] The figure shows that the SHRUG has lower variance than the IHDS and NSS but otherwise a similar distribution. 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. Similarly, the structure of the asset regression is such that there is a minimum estimated household consumption quantity, which is the same for all households with none of the assets on the asset list. But in the NSS and IHDS there is no mechanical lower limit for household consumption. The SHRUG therefore has no households with either extremely high or extremely low consumption.

Second, we compared average consumption in SHRUG districts and IHDS districts. Figure 2 presents a binscatter showing that average consumption covaries very strongly across these two datasets in both rural and urban areas. Finally, we decomposed the difference in average consumption between the SHRUG and the IHDS. Appendix Tables A3 and A4 compare the mean of each component of the consumption index in the SHRUG and the IHDS. Columns 1 and 2 show the share of households with

[18] $2 threshold converted in PPP to Indian rupees in the following way. It should be noted that GDP per capita (current US$) of India in 2012 is 1,446.985, GDP per capita PPP of India in 2012 is 4,916.486, and exchange rate between Indian rupee and US$ in 2012 is 53.427 according to the World Bank. Thus, $2 PPP is equal to 31 rupees (2∗1,446.985/4,916.486∗53.427) or 31 rupees per day. 27.2/33.33 rural/urban classification is from the 2011 Suresh Tendulkar Committee report (Planning Commission of India, 2009).

[19] Each observation in this density plot is a village. For the sample surveys, we apply sampling weights when calculating village means (to take into account stratification on affluence and occupation), and when plotting densities (taking into account cross-village stratification on village size and oversampling in some states).

each characteristic in both datasets; Column 3 shows the difference. Column 4 shows the coefficient on consumption from the prediction regression in the IHDS. Column 5 shows the average difference in Indian Rupees between consumption estimates in the SHRUG and consumption as measured by 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 transformation of household assets into consumption values assumes a similar relationship between consumption and asset ownership in the IHDS and the SHRUG. But there is no mechanical link between the asset ownership shares in the two datasets; the similarity in the estimated consumption distributions is therefore informative and gives us confidence that the consumption measures in the SHRUG are good proxies for these direct survey measures.

3 Analysis: The High-Resolution Geography of Development

3.1 What do Night Lights Proxy?

The use of satellite imagery as a data source on local development has been a major innovation in the last 20 years. The most popular form of satellite imagery in economics has been the use of night lights as a proxy for cross-national GDP (Henderson et al., 2011). Night lights have been particularly valuable for the study of low-income countries because they offer a high geographic resolution proxy of development where no other data may be available.

The research has not been entirely consistent on what exactly night lights are proxying. Night lights have been assumed in various studies to represent GDP growth (Henderson et al., 2011), cross-sectional GDP (Bleakley and Lin, 2012), urban extent (Harari, 2020; Baum-Snow et al., 2017), public expenditure (Hodler and Raschky, 2014), rural electrification (either supply or demand) (Min et al., 2013; Baskaran et al., 2015; Burlig and Preonas, 2016; Mahadevan, 2019). Many of these studies work from an underlying assumption of a constant elasticity between night lights and the outcome of interest in the cross-section and in the time series, as well as across different geographic scales of analysis. A common approach is to show that night lights are correlated with outcomes at an aggregate scale (e.g.

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

across countries), and then to use night lights as a proxy for the same variable at a microgeographic scale, where data to test the relationship between night lights and the outcome do not exist.

In this section, we show that: (i) night lights are indeed strongly and significantly correlated with many of their presumed proxies even at the highest geographic resolution; (ii) because night lights are independently correlated with a wide range of correlates of development, it is difficult from light data alone to determine exactly which proxy is being measured; (iii) night light elasticities vary widely across outcome and across context; and (iv) high geographic resolution night light elasticities are an order of magnitude smaller in the time series than in the cross section.

We combine data on village-level luminosity with population, employment in manufacturing and services (or combined), electricity supply, and per capita consumption. The night lights variable in each case is the log of average luminosity in a location polygon, where luminosity is a value from 0–63 observed in each 1km x 1km grid cell. We focus on a strictly rural sample because our measure of electricity is not comparable across villages and towns in the cross-section.[21]

3.1.1 Night Light and Development in Cross Section

Table 3 describes the relationship between each outcome variable and luminosity, controlling only for the area of the geographic unit. The estimating equation, based on Henderson et al. (2011), is:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(light_i) + \beta_2 \ln(numcells_i) + \epsilon_i, \quad (1)$$

where Y_i is the outcome variable, $light_i$ is the average pixel luminosity in the geographic polygon, $numcells_i$ is the village area, and ϵ_i is an error term.[22] The unit of analysis i is a district (Column 1), subdistrict (Column 2), or village (Columns 3 and 4). Columns 3 and 4 respectively include

[21] In villages in 1991 and 2001 we use an indicator variable for whether the village had any power supply. In 2011 we use the number of hours of electricity available per day. The village directories report hours available in winter and in summer; we take the mean of these measures. All of these power variables are normalized in this analysis to make their spatial variation maximally comparable; however, it is difficult to determine whether power has increased between 2001 and 2011 because of the change in the measure used. The electrification measures used in the town directories (i.e. for urban locations) are the number of electricity connections for different types of users. With these measures, it is not possible to compare the difference in electrification between a village and a town. All of these variables are from the census village directories.

[22] The area control is the number of 1km x 1km cells in the geographic unit. We include this to omit mechanical effects of geographic size of administrative units. Columns 2–4 are clustered at the district level.

district and subdistrict fixed effects to limit variation in the analysis to microgeographic variation *within* these larger units. All variables are measured in logs, so the coefficients can be interpreted as elasticities.[23] Each entry in the table shows a coefficient β_1 from a separate regression.

Night lights are extremely highly correlated with local population, employment in manufacturing and services, electrification, and per capita consumption, at all levels of geographic aggregation. Even when variation is limited to be within each of India's 5000 subdistricts (approximately 100 hundred villages each, Column 4), night lights are extremely strong predictors of all six variables.[24] The validity of night lights at high geographic resolution for all these outcomes has been presumed to be true by a large number of papers using night lights as local proxies of development, but has seen little testing in a developing country context. Further, the log-log relationship between the outcomes and night lights is broadly linear across the distribution for all of the variables, as shown in Figure 3, even at the lowest levels of development.

However, the elasticities are far from constant across geographic scales, and vary widely across variables. The cross-district and cross-subdistrict elasticity of population with respect to night lights is 1.3; the within-district and within-subdistrict elasticity is less than a third of that at 0.4. For employment and power supply, we also find that elasticities at aggregate levels are more than twice as high as the local elasticities. For per capita consumption, the elasticites are much less variable but still vary by a factor of two.

The ordering of the elasticities across variables is consistent across all four specifications. The outcome with the highest elasticity with respect to night lights is employment, followed by population, then power supply, and finally per capita consumption. In all specifications, the employment elasticity is more than eight times higher than the consumption elasticity. At the village level, a 50% increase in night lights in associated with a 26% increase in non-farm employment, a 12% increase

[23] In all log specifications, we add one to the employment and power variables before taking logs (reflecting respectively, one job and one hour of daily electricity), so that zeroes are not dropped. Results are virtually identical if we use the inverse hyperbolic sine transformation or add a different quantity (e.g. 0.1 hours of electricity). Population and consumption are never zero, and none of the quantities are zero at the aggregate levels of subdistrict and district.

[24] These results are robust to using both level and standardized measures of population, employment, consumption and power. Measuring population and employment in levels results in a much worse fit (lower correlations, regression coefficients, and r-squared) because of the presence of outliers.

in population, an 8% increase in power, and a 3% increase in per capita consumption.

At high geographic resolution, all of these elasticities remain statistically significant even after controlling for the other potential proxy variables, with the partial elasticities ranging between 0.05 (for consumption) and 0.71 (for non-farm employment) (Appendix Table A5). A key result is that night lights cannot be interpreted as a specific proxy for only one of these outcomes in isolation; they reflect a combination of development outcomes. In any given context where night lights are used as a proxy for development, it is difficult to determine whether they are describing population, employment, power, or expenditure. They could be performing different roles in different places, making interpretation difficult.

Night light elasticities are not constant across place. Figure 4 shows that for all outcomes other than consumption, night light elasticities are 50–100% higher in electrified villages; for consumption, the elasticity is similar in electrified and non-electrified villages.[25]

3.1.2 Night Light and Development in the Time Series

Many of the papers that use night lights as an outcome variable focus on *changes* in light rather than cross-sectional differences in light across space.[26] It is therefore important to understand whether elasticities in the time series are similar to elasticities in the cross section. This relationship might not hold if the elasticities above are driven in part by correlations with additional unobserved variables. It also might not hold if the covariates have different variance in cross-section and time series; for instance, in a period where there are very few changes in electricity supply, time series night light elasticities with respect to electricity supply will be relatively smaller compared with the other elasticities. In this section, we estimate night light elasticities in a time series setup with village fixed effects.

To align the years across data sources, we extrapolate the values from the 1991, 2001, and 2011 Population Censuses (population and electricity) to the Economic Census years of 1990, 1998, 2005, and 2013.[27] Our only measure of per capita consumption is in 2012, so it is not possible to include

[25]Specifically, the sample is split according to whether a village reports non-zero hours of electricity.

[26]For instance, many programs are evaluated with models with location fixed effects, which effectively compare night light changes in treated places to non-treated places.

[27]We assume a constant rate of population growth between census years and extrapolate the 1991–2001 growth rate for 1990 and the 2001–2011 growth rate for 2013. We use a linear extrapolation/interpolation for the standardized

changes in consumption in this analysis.

Table 4 shows the partial correlation between each outcome variable and luminosity, at various levels of geographic aggregation. Each coefficient in the table is the coefficient β_1 from a single regression that takes the form:

$$\ln(Y_{i,t}) = \beta_0 + \beta_1 \ln(light_{i,t}) + \nu_i + \psi_t + \epsilon_{i,t}. \quad (2)$$

$Y_{i,t}$ and $light_{i,t}$ are defined as above. This specification includes fixed effects for year (ψ_t) and location (ν_i), such that all variation picked up by β_1 is within the analysis unit over time. Columns 1 and 2 show the aggregate district and subdistrict level regressions, while Columns 3 and 4 show village-level regressions with subdistrict*year and district*year fixed effects respectively. These fixed effects isolate the relationship between changes in light and development outcomes from regional changes.

Elasticities vary substantially at different geographic scales. Non-farm employment elasticities are substantially higher at higher levels of aggregation than at more geographically fine units. At the village level, once district-year fixed effects are taken out and variation is restricted to changes across villages within districts, the employment elasticities fall to 0.03 and lower. The population elasticities are very small in all specifications and are less than 0.01 at the village level. Changes in power supply are in fact negatively correlated with changes in night lights at aggregate geographic scales, but have an elasticity of 0.03–0.04 at the village level.

In spite of substantially smaller elasticities, the relationship between night lights and all covariates remains highly significant in the expected direction at the village level, with the exception of manufacturing employment in the within-district regression. Across all specifications, night lights have higher elasticities with respect to employment in the services sector than in the manufacturing sector, reversing the pattern observed in the cross-section. Services employment in rural India is concentrated in education, health and retail; the higher elasticity of lights with respect to services may suggest that night lights are picking up changes in consumption and expenditure patterns

electrification measure.

(because higher demand results in a larger retail sector), rather than concentrations of production.[28]

3.1.3 Discussion: What Can We Learn from Night Lights?

The results in this section provide clarification and nuance for analyses that use night lights as a proxy for development. Night lights are predominantly used to estimate measures of development at high geographic resolution where other development proxies are unavailable. As such, they are often used exactly in contexts where the assumption that lights will be associated with various development measures at a fine geographic scale is difficult to test.

The results shown here support the idea that night lights do indeed describe highly local variation in development in contexts where little data is available. We have shown that night lights have strong cross-sectional associations with all of the variables that they are presumed to proxy in the literature, even at a highly local scale and in a very rural environment. These correlations are to some extent independent: population, employment, electricity, and per capita consumption are all strongly correlated with lights, even after controlling for each other. The relationships are strongly log-linear in the cross section, holding equally for both rich and poor places, as well as well-lit and poorly-lit places. This finding is therefore supportive of the dominant methodology in the research of interpreting night lights as a valid high geographic resolution proxy for a range of development outcomes.

However, transforming the coefficients from night lights into specific development measures is a fraught exercise. Night lights are independently correlated with all of the covariates described here. If a policy or program is found to be correlated with increased luminosity, then it will be difficult for researchers to determine whether that correlation is proxying for population, employment, electrification, per capita consumption, or something else.

Importantly, our results have implications for the application of luminosity-to-GDP elasticities from the literature to new contexts. Many studies of night lights estimate time series regressions with place and year fixed effects. Our results suggest that time series elasticities in these contexts may be an order of magnitude smaller than the widely reported cross-sectional elasticities. This would

[28] In Appendix Table A6, we show that these results are robust to limiting the analysis to the census years with the highest quality data (2001–2013) and to weighting regressions by population.

imply that program effects measured with night lights may in fact be much smaller than presumed.

Further, night light elasticities with respect to various development measures are highly variable across different levels of aggregation and in different places. For instance, we found that elasticities can be twice as large in places that are electrified. Our analysis suggests that the application of rule of thumb conversions between night lights and GDP at highly disaggregated geographic levels in developing countries should be used with extreme caution.

3.2 What is the Distribution of Jobs across Villages and Towns?

Economic opportunities are strongly contingent on local labor and goods markets. Given high transportation costs in developing countries, the relevant labor market for an individual may be geographically very small. Analysis of regional opportunities based on large geographic aggregates may therefore not reflect the opportunities available to most individuals in a region, especially if economic opportunities are not evenly distributed across space.

In Figure 5, we rank each town and village according to the number of nonfarm jobs per 1000 people (henceforth job density), and plot this measure in logs against the town and village rank.[29] Rank percentiles are adjusted for population, such that the 20th percentile village has higher job density than villages representing the most job-dense 20% of the rural population. If jobs were equally distributed across space, the job density would be constant and these lines would have a slope of zero. Instead, the distribution of job density follows a power law as evidenced by the linearity of the function across most of its support. In other words, highly ranked places have orders of magnitude higher job density than lower ranked places.

One way to characterize this relationship is with the coefficient from a regression of location job density on location density rank; this is the slope of the curve in Figure 5. We estimate the slope of this function within percentiles 10 and 90 to mitigate the effect of outliers, though we find similar estimates if we include all data points.

For urban manufacturing jobs, the coefficient is -0.019. Moving from the 20th percentile town to the 80th percentile town changes the manufacturing job density from 48 jobs per 1000 people

[29] On average there are 22 manufacturing jobs and 61 service jobs per 1000 people in India.

to 16 jobs per 1000 people. Urban service sector jobs are more evenly distributed, with a regression coefficient of -0.012. Jobs in rural areas are much more concentrated, especially in manufacturing; the coefficients for rural manufacturing and services jobs are respectively -0.04 and -0.027. The 20th percentile village has 17 manufacturing jobs per 1000 people (just above the national rural average), while the 80th percentile village has less than 1. Table 5 summarizes the concentration of job density, showing the difference between mean and median job density, as well as the slope of the regression function described above with and without district fixed effects.

The uneven distribution of jobs, especially rural manufacturing, is consequential for local labor markets. The mean village has 15 manufacturing jobs per 1000 people, but the median has only a third of that. 79% of rural Indians live in villages with fewer manufacturing sector opportunities than what is implied by mean values. These results are not driven by variation across districts; as shown in Table 5, the concentration measures are very similar when calculated with district fixed effects.

The power law distribution of rural manufacturing is important in many domains and implies that analysis at higher geographic aggregations may be misleading. District means may not reflect the experiences of the overwhelming majority of individuals.

3.3 How Localized is Poverty?

Development programs are often geographic in scope, targeting regions which are under-serviced, poor, or populated by marginalized groups. In India, policies are often targeted to the district and subdistrict level; for instance, India's District Primary Education Program (DPEP) built schools, hired teachers, and improved infrastructure in districts with below median literacy rates (Khanna, 2017). India's national and state governments have long maintained lists of so-called "backward" districts and blocks which affect eligibility for a range of programs (Kumar, 2020).[30]

Targeting programs at high levels of geographic aggregation may be desirable for implementation reasons or simply because higher resolution data is not available to policymakers. The cost of allocating programs at aggregate levels is that needy individuals will be left out if they live in better off regions.

[30]There are on average 10 blocks per district (compared with 1000 villages per district), making them comparable to subdistricts, the primary intermediate unit in the Indian Census and used here.

Both policy planning and evaluation are difficult in the absence of high geographic resolution data, such that the extent of mistargeting driven by geographic aggregation is often not even known. If poverty is primarily a phenomenon of broad regions, then district-level targeting may be almost as efficient as village-level targeting. If variation in poverty is mostly local, then district-level targeting could be much worse. India's primary sample survey with broad geographic coverage, the National Sample Survey, is minimally helpful in this regard, because it is representative at the district level and does not sample enough individuals within primary sampling units to describe the geographic distribution of development at geographically finer levels.

In this subsection, we examine the effects of using different geographic aggregations on the efficiency of a hypothetical program targeting individuals who live below the poverty line. Our analysis follows an established literature examining the social benefit of targeting anti-poverty programs at various geographic scales (Ravallion, 1993; Baker and Grosh, 1994; Bigman et al., 2000; Bigman and Srinivasan, 2002; Elbers et al., 2007; Brown et al., 2019).

For each level of aggregation (district, subdistrict, and village), we rank the locations by poverty rate. Ranks are adjusted for population such that a rank X implies that $X\%$ of the national population lives in a location with a higher poverty rate. Figure 6 shows the average poverty rate at each rank, based on the geographical granularity of the ranking unit. The poverty rate in the 5th percentile district is 43%; it is 47% in the 5th percentile subdistrict, and 57% in the 5th percentile village.

If the hypothetical program was targeted to the bottom 25% of the population as ranked by the regional poverty rate, then it would reach 34.6% of poor people if targeted by state, 40.4% if targeted by district, 42.4% if targeted by subdistrict, and 49.1% if targeted by village. Relative to district targeting, we find that targeting at the village level could allow an additional 21% more individuals below the poverty line to receive program benefits, a greater improvement than that of moving from state to district targeting.

Even though subdistricts are 10 times smaller than districts, the targeting gain in moving from district to subdistrict is small, but the gain in moving to village-level targeting is much higher. This effect size is not something that could be measured from district-level data because the geographic

aggregate at which poverty is most concentrated is not known a priori.

A caveat to all of these measures is that targeting exercises based on small area estimates necessarily rely on assumptions of a uniform relationship between assets and consumption that may not hold (Tarozzi and Deaton, 2009). Standard errors on targeting estimates are therefore likely to be underestimated.

4 Comparative Advantages and Disadvantages of the SHRUG

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 a wider set of high geographic resolution locations than any other Indian panel dataset with broad scope.[31] This enables analysis of outcomes at the level of policy and program variation, which is often at the village, town or constituency level.

Second, because the SHRUG's geographic coverage is comprehensive, it is particularly valuable when used in tandem with other datasets. Each new administrative or remote sensing data source that is linked to the SHRUG can be fully integrated with all the other data sources in the SHRUG, expanding the scope of potential analysis. In contrast, sample surveys run by different research teams do not have the same complementarities, because there is rarely enough sample overlap for the same locations to appear in both datasets. Two research teams integrating new administrative or remote sensing data into the framework of the SHRUG will immediately be able to benefit from each others' work due to the linked identifiers and comprehensive geographic coverage. For this reason, the utility of the SHRUG will increase over time as we continue to extend its breadth.

Like every data source, the SHRUG has several limitations. First, 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 non-farm firm in India, they are necessarily based on much shorter surveys than detailed sample surveys like the NSS and ASI. This disadvantage is traded

[31] 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. The SHRUG can in principle be linked to Prowess at the village and town level using georeferences, but we have not yet created a key. There are also several village-level time series datasets with broad and deep surveys, chiefly ARIS-REDS and ICRISAT. The data in these surveys is much more detailed than the data found in the SHRUG, but they cover at most a few hundred villages.

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.

Second, the SHRUG consists of locational aggregates, rather than individual or firm microdata. It is thus not suitable for studying variation across individuals or firms *within* locations. Note that we have provided keys to link SHRUG to the firm-level Economic Censuses, which describe every non-farm establishment in India in four cross sections from 1990 to 2013.

Third, not all villages and towns are matched in all periods. If a researcher's goal was to estimate the number of non-agricultural firms in India, for example, then aggregating from NSS or Annual Survey of Industries (ASI) 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, and many towns are missing economic data for 1990. Similarly, as noted above, we were not able to obtain census-based data for predominantly urban constituencies.

Fourth, 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, most of the data in SHRUG is available only for the years when large-scale administrative data were collected. Consumption data is available only in 2012, and none of the four Economic Census years are exact matches for the Population Census years. The remote sensing data, however, is available annually.

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. The SHRUG is particularly well-suited for research requiring high-resolution geographic variation, rich location information, or socioeconomic outcomes in units with political boundaries.

5 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 yield many more positive externalities for 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.[32] Our aim in 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.

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

[32]Some examples of Indian administrative programs that have been studied with administrative data include the NREGS public works and wage support program, the RGGVY rural electrification program, and the ongoing Total Sanitation Campaign, all of which are the subject of multiple research papers. And yet none of these programs (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.

projects, and researchers may not see rewards for doing so.[33]

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-linked data sources will have a large lead on any other projects working with their data.[34] Our project aims to raise the returns of data assembly, resulting in new public goods for the research and policymaking communities.

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 model a set of norms and protocols to facilitate a similar equilibrium around the sharing of data for social science research.

We conclude by reiterating our request to users of the SHRUG to cite all of the research papers that underlie the different components of the SHRUG. When data is downloaded from the SHRUG platform, a list of 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.

5.1 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

[33]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 searching for the gems that do exist.

[34]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.

archived on the Harvard Dataverse, so that researchers can always replicate an exact prior download.[35]

Researchers wishing to make their data easily linkable to SHRUG need only to post their data with unique SHRUG identifiers. We aim to maintain a listing of 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 help to make these contributions as accessible as possible to outside researchers.[36]

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.

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

[36]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*, 2020, *110* (3).

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

Baker, Judy L and Margaret E Grosh, "Poverty Reduction through Geographic Targeting: How well does it work?," *World Development*, 1994, *22* (7).

Baskaran, Thushyanthan, Brian Min, and Yogesh Uppal, "Election Cycles and Electricity Provision: Evidence from a Quasi-experiment with Indian Special Elections," *Journal of Public Economics*, 2015, *126*, 64–73.

Baum-Snow, Nathaniel, Loren Brandt, J. Vernon Henderson, Matthew A. Turner, and Qinghua Zhang, "Roads, Railways and Decentralization of Chinese Cities," *Review of Economics and Statistics*, 2017, *99* (3).

Bigman, David and PV Srinivasan, "Geographical Targeting of Poverty Alleviation Programs: Methodology and Applications in Rural India," *Journal of Policy Modeling*, 2002, *24* (3).

___ , Stefan Dercon, Dominique Guillaume, and Michel Lambotte, "Community Targeting for Poverty Reduction in Burkina Faso," *The World Bank Economic Review*, 2000, *14* (1).

Bleakley, Hoyt and Jeffrey Lin, "Portage and Path Dependence," *The Quarterly Journal of Economics*, 2012, *127* (2).

Brown, Caitlin, Martin Ravallion, and Dominique van de Walle, "Most of Africa's Nutritionally Deprived Women and Children are Not Found in Poor Households," *Review of Economics and Statistics*, 2019, *101* (4).

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.

___ , Tomoki Fujii, Peter Lanjouw, Berk Ozler, and Wesley Yin, "Poverty Alleviation through Geographic Targeting," *Journal of Development Economics*, 2007, *83* (1).

Ellison, Glenn and Edward L Glaeser, "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 1997, *105* (5), 889–927.

Gabaix, Xavier, "Power Laws in Economics: An Introduction," *Journal of Economic Perspectives*, 2016, *30* (1).

Harari, Nina, "Cities in Bad Shape: Urban Geometry in India," *American Economic Review*, 2020, *110* (8).

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

Hodler, Roland and Paul A Raschky, "Regional Favoritism," *The Quarterly Journal of Economics*, 2014, *129* (2).

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

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

Khanna, Gaurav, "Large-scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India," 2017.

Kumar, Hemanshu and Rohini Somanathan, "State and District Boundary Changes in India: 1961-2001," 2015. Working Paper.

Kumar, Naveen, "Public School Quality and Student Outcomes: Evidence from Model Schools in India," 2020. 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.

Mahadevan, Meera, "The Price of Power: Costs of Political Corruption in Indian Electricity," 2019. Working paper.

Min, Brian, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou, "Detection of Rural Electrification in Africa using DMSP-OLS Night Lights Imagery," *International Journal of Remote Sensing*, 2013, *34* (22).

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

Planning Commission of India, "Report of the Expert Group on Methodology for Estimation of Poverty (Tendulkar Committee Report)," Technical Report, New Delhi, India 2009.

Prakash, Nishith, Marc Rockmore, and Yogesh Uppal, "Do Criminally Accused Politicians Affect Economic Outcomes? Evidence from India," *Journal of Development Economics*, 2019.

Ravallion, Martin, "Poverty alleviation through regional targeting: a case study of Indonesia," in A Braverman, K Hoff, and J Stiglitz, eds., *The Economics of Rural Organization: Theory, Practice and Policy*, World Bank and Oxford University Press, 1993.

Tarozzi, Alessandro and Angus Deaton, "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas," *The review of economics and statistics*, 2009, *91* (4).

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
Geographic Variance Decomposition of SHRUG Variables

Panel A. Share of Variance in Town Outcomes Explained by State and District

	State	District
Consumption per capita	0.392	0.640
Nonfarm employment per Capita	0.102	0.254
Night lights per capita	0.056	0.203
Population density	0.162	0.304
Average forest cover	0.465	0.736

Panel B. Share of Variance in Village Outcomes Explained by State, District, Subdistrict

	State	District	Subdistrict
Consumption per capita	0.321	0.450	0.526
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
Average forest cover	0.472	0.640	0.729

Table 1 presents the spatial decomposition of the variance of a selection of variables in the SHRUG. Stated values are the R^2 from 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. 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 2
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, including electrification, roads, post offices, markets, health centers, schools, land use, water source, etc.	Village, Town, Constituency, District
Economic Census	1990, 1998, 2005, 2013	Employment and sector of all non-ag firms, including manufacturing and services, formal and informal, government and private firms	Village, Town, Constituency
SECC	2012	Consumption estimates, poverty, agricultural labor share	Village, Town
Election Results	1980–2018	Candidate name / party / votes	Constituency/Candidate
Politician Assets/Crime	2003–2018	Criminal charges, assets, liabilities, education for all politicians contesting MLA seats	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, including length, material, cost, milestones, etc.	Village

Table 3
Cross-Sectional Correlates of Night Lights

	District	Subdistrict	Village	Village
	(1)	(2)	(3)	(4)
Log Population	1.377***	1.257***	0.433***	0.407***
	(0.124)	(0.034)	(0.003)	(0.003)
Log Non-Farm Employment	1.755***	1.740***	0.711***	0.697***
	(0.156)	(0.049)	(0.005)	(0.005)
Log Manufacturing Employment	2.291***	2.090***	0.714***	0.687***
	(0.174)	(0.055)	(0.005)	(0.006)
Log Services Employment	1.625***	1.669***	0.683***	0.660***
	(0.152)	(0.047)	(0.004)	(0.005)
Log Hours Electricity	0.604***	0.553***	0.264***	0.215***
	(0.092)	(0.031)	(0.003)	(0.004)
Log Consumption	0.167***	0.114***	0.105***	0.086***
	(0.027)	(0.008)	(0.005)	(0.004)
N	632	5756	430982	430788
Fixed Effects	None	None	District	Subdistrict

$^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$

Table 3 shows the cross-sectional relationship between a set of rural development outcomes and night-time luminosity at various levels of aggregation. Each coefficient in the table is from a separate estimation of Equation 1—a regression of the log outcome variable on log luminosity and the log polygon size (the number of 1km x 1km cells in the geographic unit). Column 1 aggregates data to the district level, Column 2 to the subdistrict level, and Columns 3 and 4 to the village level. Columns 2-4 are clustered at the district level. Standard errors on consumption are calculated using 1000 bootstraps as described in Section 2.4. Source: SHRUG.

Table 4
Time Series Correlates of Night Lights

	District	Subdistrict	Village	Village
	(1)	(2)	(3)	(4)
Log Population	0.013	0.026***	0.006***	0.004***
	(0.011)	(0.008)	(0.001)	(0.001)
Log Non-Farm Employment	0.555***	0.346***	0.020***	0.021***
	(0.130)	(0.082)	(0.003)	(0.002)
Log Manufacturing Employment	0.398***	0.202***	0.006	0.009***
	(0.113)	(0.075)	(0.004)	(0.003)
Log Services Employment	0.513***	0.329***	0.019***	0.020***
	(0.124)	(0.077)	(0.003)	(0.002)
Electricity	-0.147***	-0.106***	0.038***	0.031***
	(0.047)	(0.034)	(0.004)	(0.003)
N	2358	22060	1479176	1478715
Fixed Effects	District, Year	Subdistrict, Year	Village, District * Year	Village, Subdistrict * Year

$^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$

Table 4 shows the time series relationship between a set of rural development outcomes and night-time luminosity at various levels of aggregation, controlling for location fixed effects. Each coefficient in the table is from a separate estimation of Equation 2—a regression of the log outcome variable on log luminosity, with location and time fixed effects. Column 1 aggregates data to the district level, Column 2 to the subdistrict level, and Columns 3 and 4 to the village level. All regressions are clustered at the district level. Source: SHRUG.

Table 5
Concentration of Non-Farm Employment in Villages and Towns

	Mean Employment per 1000 ppl	**Median Employment per 1000 ppl**	**Slope (National)**	**Slope (Within District)**
Towns				
Manufacturing	39.1	26.6	-0.019	-0.012
Services	115.8	104.0	-0.012	-0.009
Villages				
Manufacturing	15.3	4.8	-0.040	-0.035
Services	40.2	26.0	-0.027	-0.021

Table 5 shows several statistics describing the distribution of non-farm employment in villages and towns, disaggregated by manufacturing and services. Columns 1 and 2 are mean and median employment per 1000 people, across the distribution of towns and villages. The mean is population-weighted, and the median refers to the town or village with the median individual, so that 50% of the national population lives in places with above-median employment per capita. Column 3 shows the slope from a regression of employment per 1000 people on the town or village rank, where each location's rank is a function of employment per 1000 people in the same category. Column 4 shows the slope from the same regression, but with district fixed effects. Source: SHRUG and Economic Census.

Figure 1
SHRUG, IHDS, and NSS Consumption Per Capita

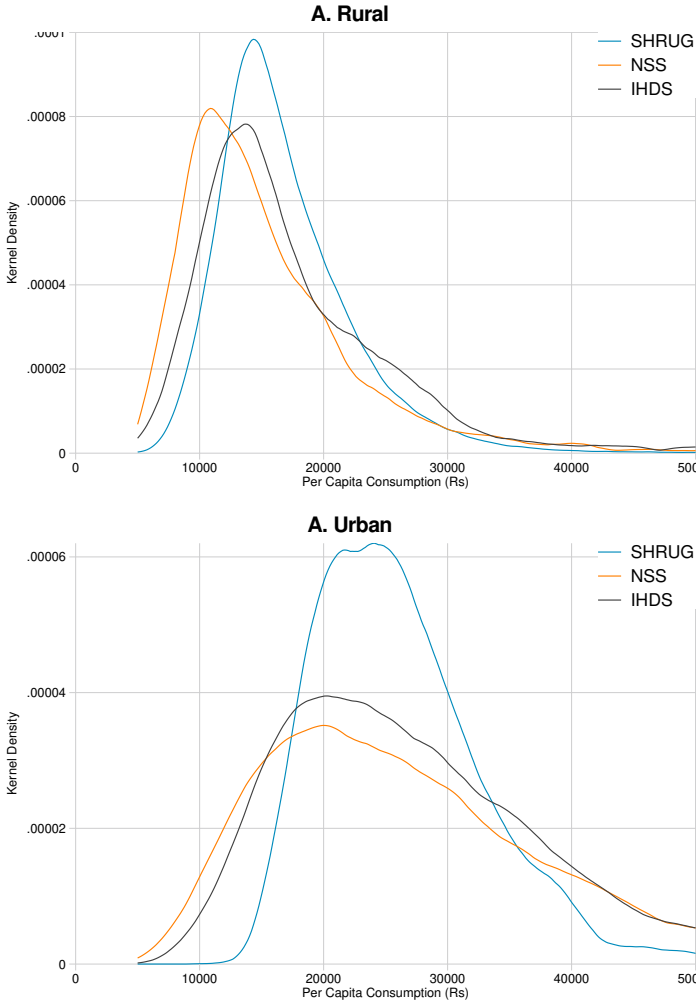


Figure 1 shows the distribution of location-level consumption per capita from the NSS, the IHDS, and SHRUG. Consumption in each location is a mean across households, calculated with sampling weights. Location units are PSUs in IHDS, FSUs in NSS, and towns or villages in SHRUG.

Figure 2
District-level SHRUG Consumption vs. IHDS Consumption

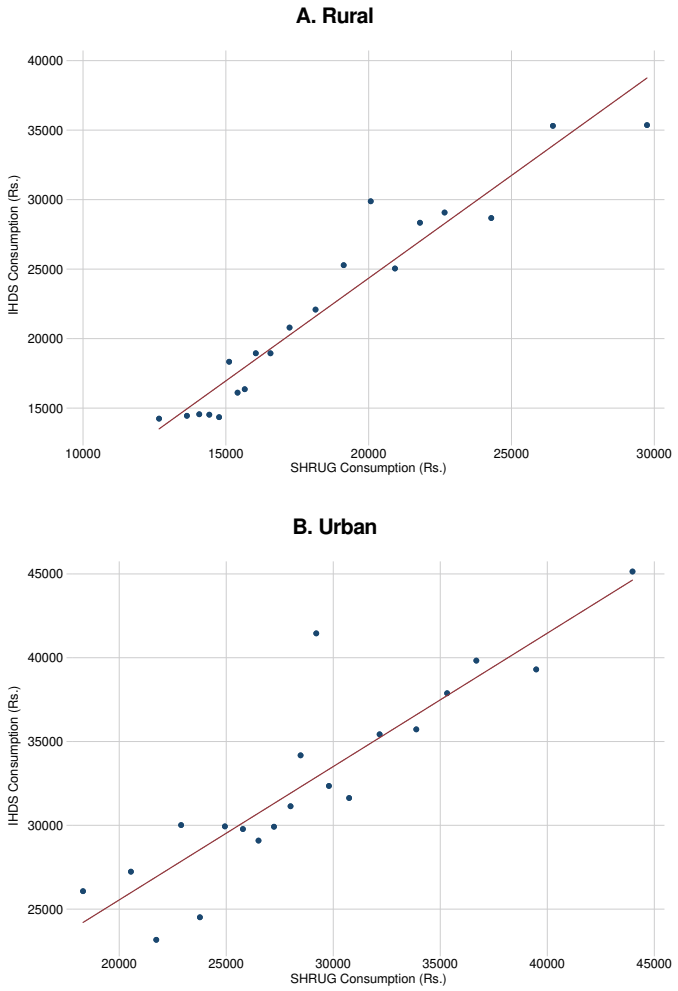


Figure 2 is a binned scatterplot of district-level average per-capita consumption in the IHDS and the SHRUG, weighted by the count of individuals in each survey. The SHRUG sample has been restricted to the 331 districts where IHDS matches the Indian Population Census.

Figure 3
Cross-sectional Night Lights vs. Development Proxies

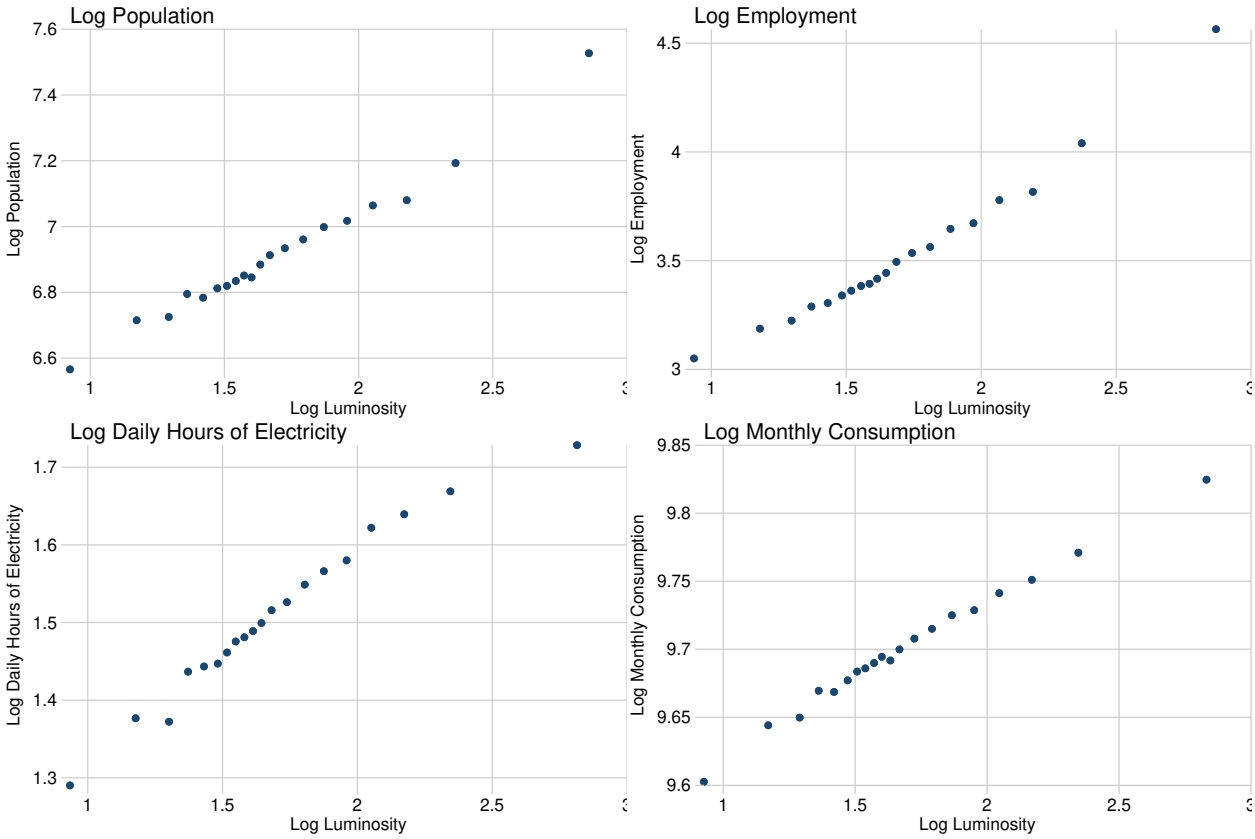


Figure 3 shows the cross-sectional relationship between log night lights and log population (top left), log employment (top right), log hours of electricity (bottom left), and log monthly consumption (bottom right). The graphs are generated by binscatter, which shows the mean Y variable in equally-dense bins across the distribution of log night lights. All graphs are run at the village level and include district fixed effects. Source: SHRUG.

Figure 4

Cross-sectional Relationship between Night Lights and Development Proxies as a Function of Electrification

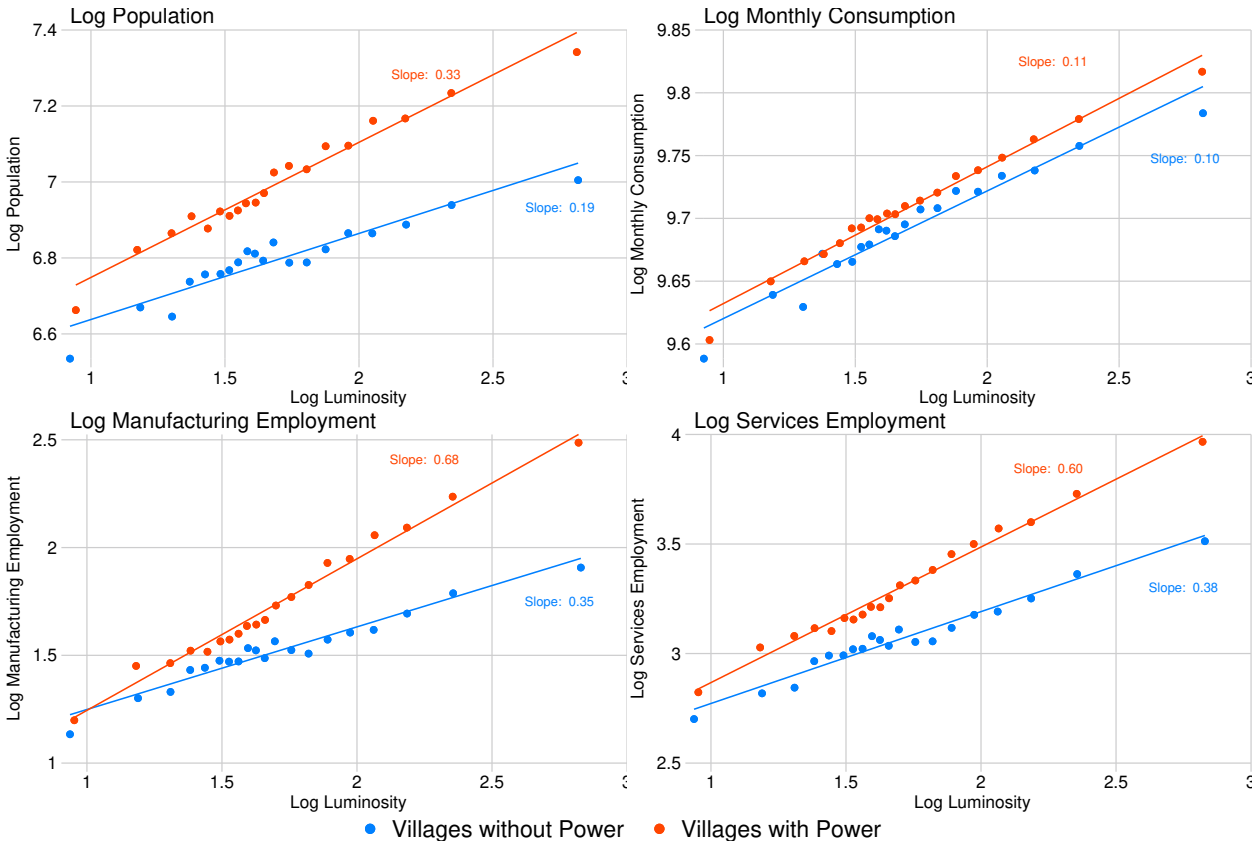


Figure 4 shows the village-level cross-sectional relationship between four development proxies (population, consumption, manufacturing and services employment), with separate series' plotted for villages that do and do not report access to electricity. The graphs are generated by binscatter, which shows the mean Y variable in equally-dense bins across the distribution of log night lights. All graphs are run at the village level and include district fixed effects. Source: SHRUG.

Figure 5
Distribution of Job Density across Space

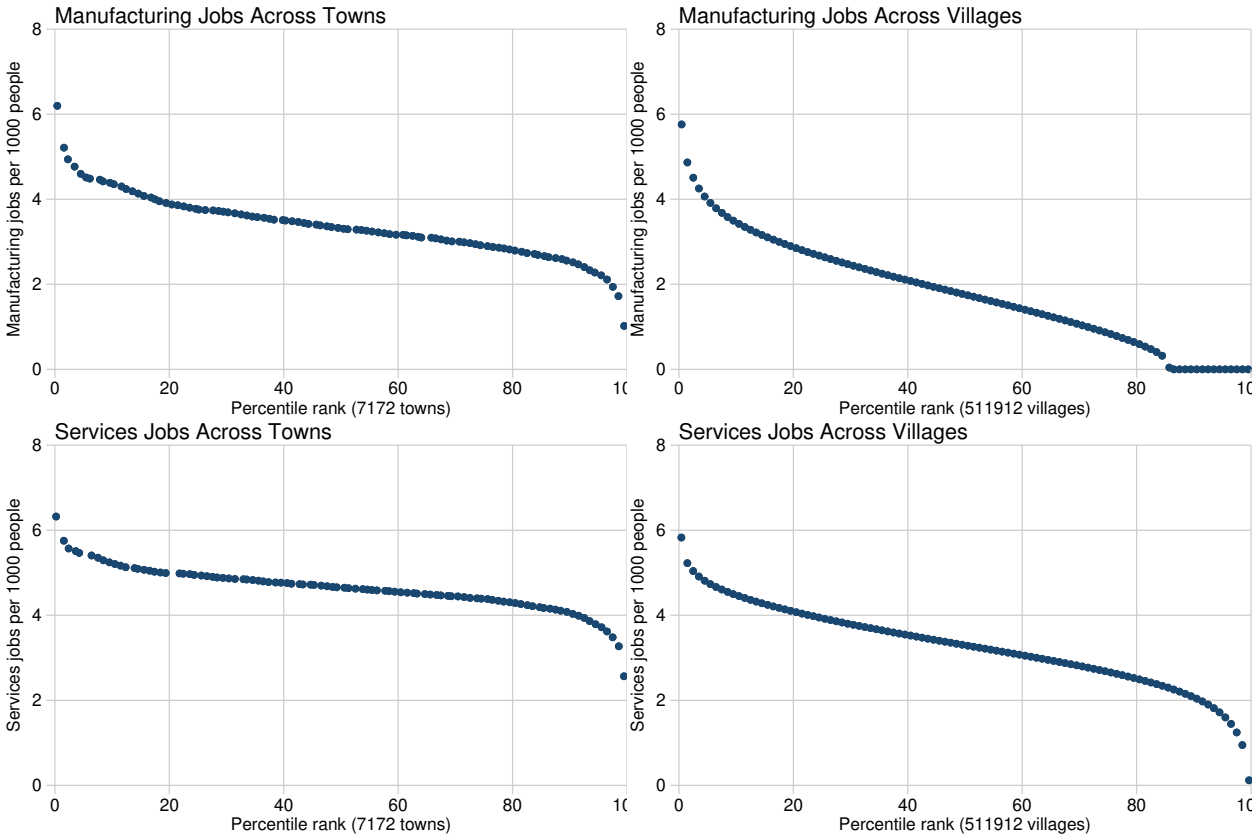


Figure 5 describes the distribution of non-farm jobs per 1000 people across villages and towns. The Y axes shows the log number of manufacturing- or service-sector jobs per 1000 people. The X axis shows the town or village rank on the same measure. The rank is population-weighted such that a village with rank 50 has more manufacturing or services jobs per 1000 people than villages representing 50% of the Indian population. Each point represents the mean of the measure in one percentile bin, or approximately either 70 towns or 5000 villages. Source: SHRUG and Economic Census (2013).

Figure 6

Concentration of Poverty at Different Levels of Aggregation

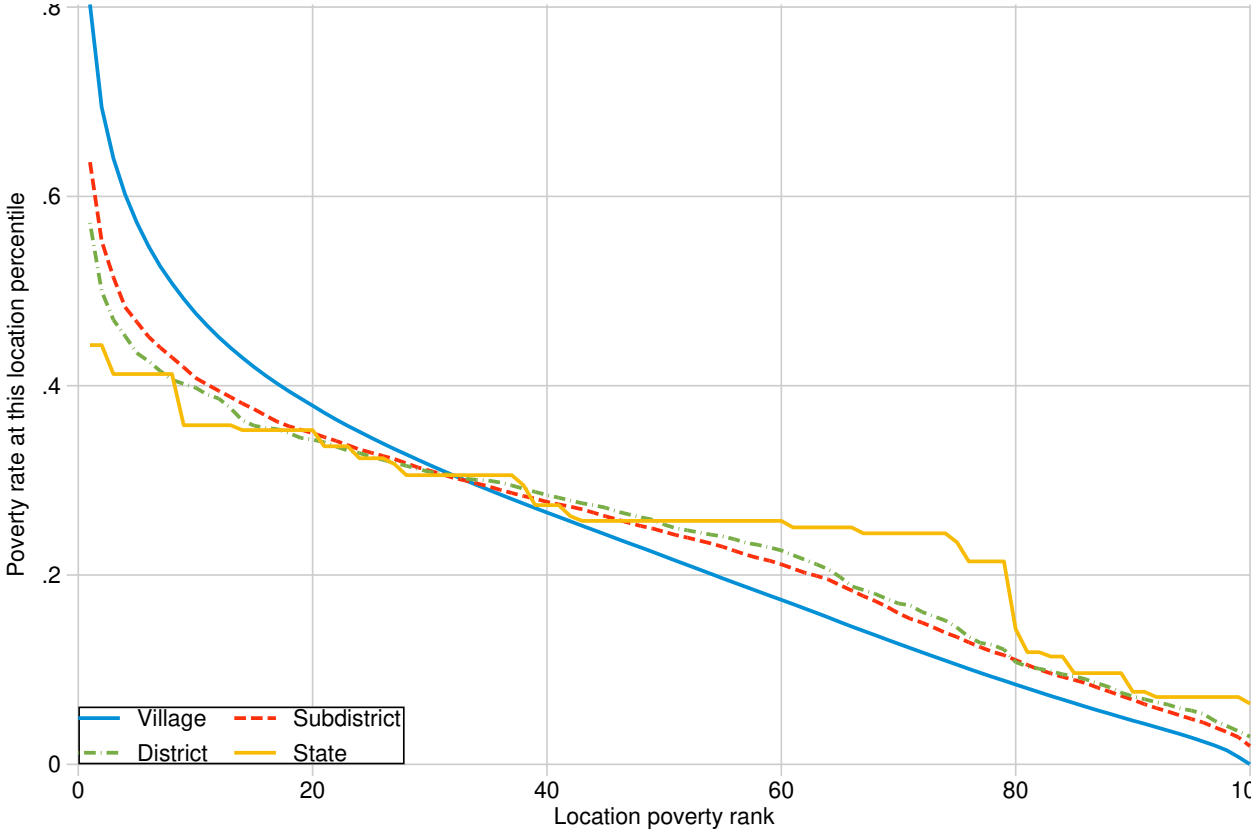


Figure 6 plots locations' average poverty rates against their poverty rate ranking. The four data series' represent four different level of aggregation: state (n=35), district (n=600), subdistrict (n=5600), and village/town (n=590,000). The graph shows, for instance, that the poverty rate in the 5th percentile district is 37%, while the poverty rate in the 5th percentile village is 50%. Source: SHRUG.

A Appendix A: Additional Tables and Figures

Table A1

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 A1 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. Chhattisgarh, Jharkhand, and Uttarakhand were created in 2000 and are thus left blank in earlier years.

Table A2

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 A2 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. Chhattisgarh, Jharkhand, and Uttarakhand were created in 2000 and are thus left blank in earlier years.

Table A3

Asset Decomposition of Small Area Consumption Estimates: Rural Households

	(1) IHDS	(2) SECC	(3) Difference	(4) Coefficient	(5) Delta
Income 5000-10,000 Rs	0.12	0.18	0.06	10076.33	590.07
Income Above 10,000 Rs	0.06	0.09	0.03	38933.33	1100.99
Home Ownership	0.99	0.95	-0.04	-1334.48	55.42
Kisan Credit Card	0.07	0.04	-0.03	12441.10	-388.38
Land Ownership	0.61	0.44	-0.17	9657.24	-1613.72
Number of Rooms in Home	2.60	2.15	-0.45	3428.70	-1549.51
Both Mobile and Landline	0.03	0.03	-0.00	31479.48	-86.19
Landline Phone	0.01	0.01	0.00	24639.32	15.93
Mobile Phone	0.68	0.69	0.01	23997.18	339.78
Refrigerator	0.11	0.12	0.01	29476.68	363.90
Brick Roof	0.05	0.07	0.02	-9604.72	-235.39
Concrete Roof	0.12	0.22	0.10	1431.76	149.18
GI Roof	0.16	0.14	-0.02	-3359.10	65.99
Grass Roof	0.23	0.16	-0.07	-2919.61	212.58
Plastic Roof	0.00	0.02	0.02	6473.79	110.06
Slate Roof	0.05	0.04	-0.01	2316.43	-34.52
Stone Roof	0.08	0.05	-0.03	11637.33	-341.08
Tile Roof	0.12	0.28	0.16	-6508.29	-1068.59
Four Wheeled Vehicle	0.02	0.03	0.01	85685.73	532.55
Two Wheeled Vehicle	0.17	0.18	0.01	34253.34	501.96
Brick Walls	0.27	0.43	0.16	23029.78	3703.36
Concrete Walls	0.24	0.04	-0.20	22316.29	-4512.30
GI Walls	0.01	0.01	-0.00	14184.44	-27.89
Grass Walls	0.07	0.11	0.04	12808.09	531.28
Mud Walls	0.33	0.27	-0.06	13371.95	-772.25
Plastic Walls	0.00	0.01	0.01	19748.41	128.43
Stone Walls	0.05	0.11	0.06	17065.06	1066.39
Wooden Walls	0.01	0.01	0.00	9216.71	15.00

Table A3 presents a comparison of IHDS and SECC covariates that were used to generate per capita consumption small area estimates in the rural 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. The omitted category for roof and wall materials was "other."

Table A4

Asset Decomposition of Small Area Consumption Estimates: Urban Households

	(1) IHDS	(2) SECC	(3) Difference	(4) Coefficient	(5) Delta
Air Conditioning	0.04	0.08	0.04	17466.61	715.74
Computer	0.13	0.14	0.01	35989.55	475.83
Indoor Toilet	0.67	0.79	0.12	6061.17	725.41
Home Ownership	0.84	0.77	-0.07	.	.
Separated Kitchen	0.71	0.72	0.01	-1356.62	-8.42
Number of Rooms in Home	2.76	2.57	-0.19	4714.67	-873.69
Both Mobile and Landline	0.11	0.07	-0.04	35447.19	-1501.55
Landline Phone	0.01	0.01	0.00	15741.89	66.50
Mobile Phone	0.80	0.80	-0.00	30750.63	-25.67
Refrigerator	0.46	0.46	0.00	23585.75	24.26
Brick Roof	0.02	0.07	0.05	-8717.23	-469.53
Concrete Roof	0.30	0.54	0.24	-1861.32	-450.93
GI Roof	0.11	0.14	0.03	-8522.64	-260.14
Grass Roof	0.06	0.04	-0.02	-10871.80	194.61
Plastic Roof	0.00	0.01	0.01	-6085.06	-64.70
Slate Roof	0.05	0.03	-0.02	-12192.98	275.30
Stone Roof	0.06	0.05	-0.01	144.10	-1.03
Tile Roof	0.07	0.10	0.03	-8480.40	-258.71
Four Wheeled Vehicle	0.07	0.07	0.00	67581.44	234.30
Two Wheeled Vehicle	0.37	0.34	-0.03	35366.95	-1212.34
Brick Walls	0.31	0.66	0.35	19329.30	6740.54
Concrete Walls	0.52	0.13	-0.39	23026.67	-8990.93
GI Walls	0.02	0.01	-0.01	26490.49	-208.85
Grass Walls	0.01	0.03	0.02	6927.12	152.49
Mud Walls	0.07	0.08	0.01	15267.08	172.38
Plastic Walls	0.00	0.00	0.00	33435.91	63.89
Stone Walls	0.06	0.07	0.01	13178.63	94.28
Wooden Walls	0.01	0.01	-0.00	6579.85	-24.11
Washing Machine	0.16	0.22	0.06	19691.06	1171.91

Table A4 presents a comparison of IHDS and SECC covariates that were used to generate per capita consumption small area estimates in the urban 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. The omitted category for roof and wall materials was "other."

Table A5
Cross-Sectional Partial Correlations of Night Lights

	District	Subdistrict	Village	Village
	(1)	(2)	(3)	(4)
Log Population	0.236***	1.257***	0.431***	0.406***
	(0.052)	(0.034)	(0.003)	(0.003)
Log Non-Farm Employment	0.104*	1.740***	0.705***	0.693***
	(0.058)	(0.049)	(0.005)	(0.005)
Log Hours Electricity	0.130	0.553***	0.262***	0.213***
	(0.083)	(0.031)	(0.003)	(0.004)
Log Consumption	0.053**	0.083***	0.097***	0.080***
	(0.026)	(0.008)	(0.001)	(0.001)
N	632	5756	425049	424855
Fixed Effects	None	None	District	Subdistrict

$^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$

Table A5 shows the partial correlation between rural luminosity and population, employment, electricity, and consumption. Each estimate is from a separate cross-sectional regression of Equation 1, with added controls for all the other variables in the table. Each entry also controls the number of 1km x 1km cells in the geographic unit. Column 1 aggregates data to the district level, Column 2 to the subdistrict level, and Columns 3 and 4 to the village level. Columns 2-4 are clustered at the district level. The table is comparable to Table 3, but has additional controls for the other outcome variables in the table. Standard errors on consumption are calculated using 1000 bootstraps as described in Section 2.4. Source: SHRUG.

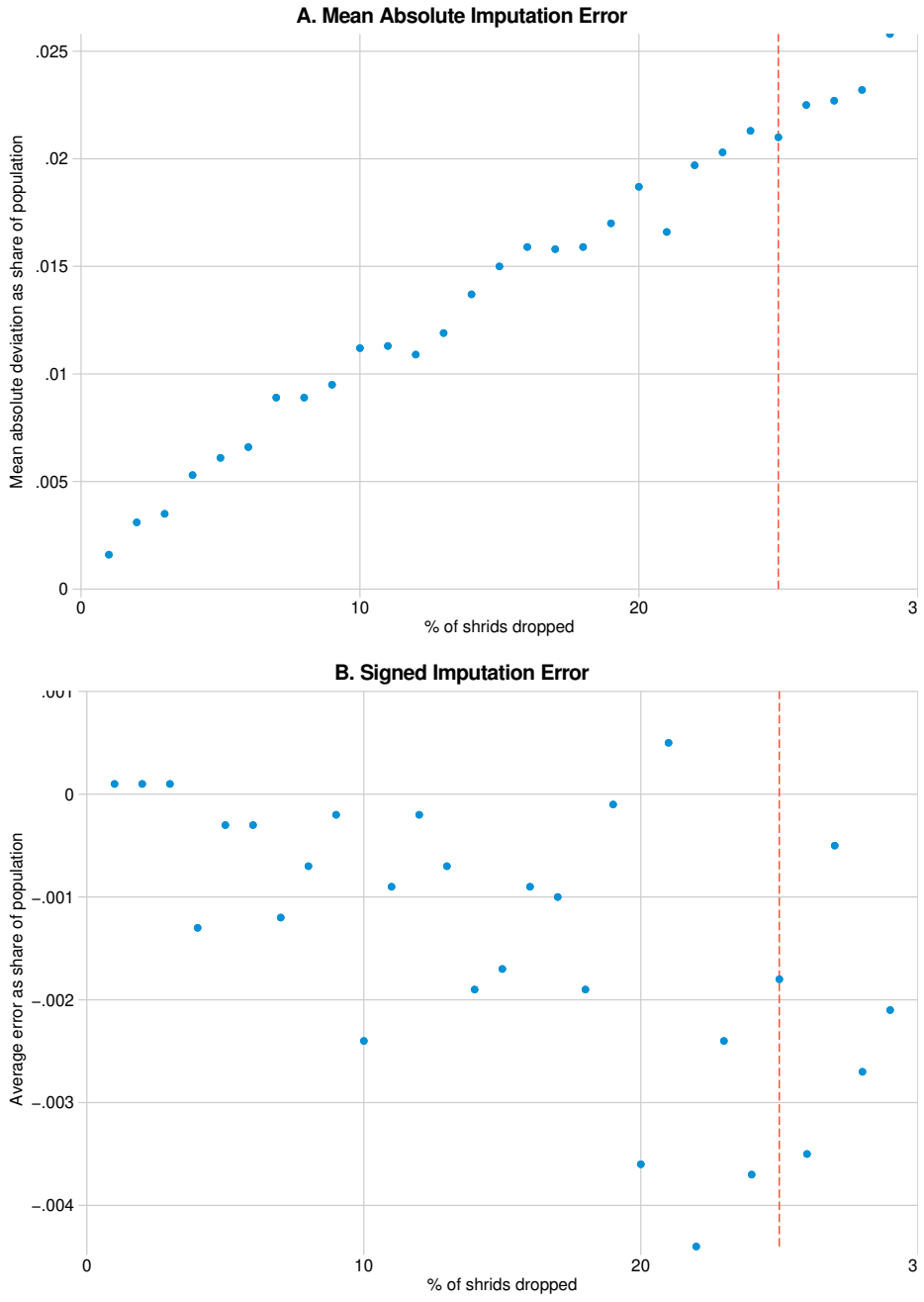
Table A6
Time Series Correlates of Night Lights: Robustness Checks

	2001–2013		Pop-weighted	
	(1)	(2)	(3)	(4)
Log Population	0.006***	0.003***	-0.004**	0.000
	(0.002)	(0.001)	(0.001)	(0.001)
Log Non-Farm Employment	0.026***	0.025***	0.047***	0.049***
	(0.009)	(0.005)	(0.004)	(0.003)
Log Manufacturing Employment	-0.019*	-0.006	0.001	0.008**
	(0.011)	(0.007)	(0.005)	(0.004)
Log Services Employment	0.033***	0.027***	0.055***	0.057***
	(0.006)	(0.005)	(0.004)	(0.003)
Electricity	0.066***	0.049***	0.035***	0.026***
	(0.013)	(0.009)	(0.004)	(0.003)
N	982594	982024	1766561	1765407
Geographic Aggregation	Village	Village	Village	Village
Fixed Effects	Village, District * Year	Village, Subdistrict * Year	Village, District * Year	Village, Subdistrict * Year

$^{*}p<0.10,^{**}p<0.05,^{***}p<0.01$

Table A6 shows the time series relationship between a set of rural development outcomes and night-time luminosity at the village level. Columns 1 and 2 restrict the data sample to 2001–2013. Columns 3 and 4 weight regressions by village population. Otherwise, estimates are similar to Columns 3 and 4 in Table 4. All variables are measured in logs. All regressions include village fixed effects and are clustered at the district level. Source: SHRUG.

Figure A1
Error Rates from Imputation Simulation



Panel A of Figure A1 shows the mean absolute deviation of simulated estimates of constitutency population as compared with actual constituency population, under scenarios where we set a different share of the population to missing. We run a simulation where before calculating constituency population, we drop village and town observations representing X% of constituency population, and then use our imputation methodology. The graph shows, for instance, that when we drop 20% of the data before imputation, our average constituency has a total population error of approximately 1.8%. Panel B shows the signed error rather than the absolute error, indicating a very small downward bias in population estimation from our method. Source: SHRUG.