Identifying Urban Areas by Combining Data from the Ground and from Outer Space

An Application to India

Virgilio Galdo
Yue Li
Martin Rama
Abstract

This paper develops a tractable method to identify urban areas and applies it to India, where urbanization is messy. Google Earth images are assessed subjectively to determine whether a stratified large sample of Indian cities, towns and villages, as officially defined, are urban or rural in practice. Based on these assessments, a regression analysis combines two sources of information—data from georeferenced population censuses and data from satellite imagery—to identify the correlates of units in the sample being urban. The resulting model is used to predict whether the other units in the country are urban or rural in practice. Contrary to frequent claims, India is not substantially more urban than implied by census data. And the speed of urbanization is only marginally higher than official statistics suggest. But a considerable number of locations are misclassified in the midrange between villages and state capitals. The results confirm the value of combining subjective assessments with data from these different sources.

This paper is a product of the Office of the Chief Economist, South Asia Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at yli7@worldbank.org.
Identifying Urban Areas by Combining Data from the Ground and from Outer Space: An Application to India

Virgilio Galdo, Yue Li, and Martin Rama

Keywords: urban extent; urbanization rate; subjective assessment; Google Earth images; georeferenced data; satellite imagery

JEL Classification: O1, O18, R1

Virgilio Galdo, Yue Li, and Martin Rama are with the Office of the Chief Economist for South Asia at the World Bank. The corresponding author is Yue Li. Her contact information is yli7@worldbank.org.

Skillful research analysis was provided by Maria Florencia Pinto, Sutirtha Sinha Roy, and Wenqing Zhang. Comments and suggestions by Gilles Duranton and by participants at the 8th European Meeting of the Urban Economics Association are gratefully acknowledged.

This research was funded by the World Bank and by the Department for International Development of U.K. as part of the Sustainable Urban Development Multi-Donor Trust Fund.

The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the executive directors of the World Bank or the governments they represent.
1. **Introduction**

What is a city? The most candid answer may be “I know one when I see it.” The subjectivity of the city concept is understandable because of the continuum between urban and rural spaces. The many terms used to describe this in-between—suburb, exurb, edge city, and urban sprawl, among others—testifies to the absence of a clear divide. Just like beauty, a city is in the eyes of the beholder.

Official answers to this question are less straightforward, varying substantially across national governments and international organizations. A glimpse at this diversity is provided by World Urbanization Prospects, a comprehensive source of international urban data that builds primarily on official information. About 45 percent of the 232 countries reported in its 2014 revision used each country’s own administrative definition as the sole or primary criterion for distinguishing urban from rural. Some 27 percent used population size as the sole criterion, and another 25 percent adopted a combination of metrics. The remaining 3 percent of countries had no official definition (UN 2015). Moreover, many countries change their definitions of cities periodically.

Not surprisingly, this diversity of criteria results in a lack of consensus on which urban agglomerations should be considered cities. It is thus not clear how urban areas can be compared across countries and over time.

This seemingly trivial issue has important implications. The rural-urban transformation drives economic development and is often intertwined with structural transformation. Because cities create prosperity by bringing together households and firms and making everyone more productive, the rural-urban transformation is central to development economics, and cities are at the core of economic geography (Fujita, Krugman, and Venables 2001; Henderson 1974; Krugman 1991; Lewis 1954).

However, the difficulties in identifying urban areas have undermined empirical efforts to connect with the theoretical literature and even to establish the stylized facts with confidence. For example, it is widely accepted that a country’s urban population share is related to its income per capita. And it is also said that India’s urban population share is too low relative to its income level. However, some studies suggest that many areas of India are misclassified as rural by the official definition, and so the level of urbanization is underestimated (Denis and Marius-Gnanou 2011; World Bank 2013).

This disconnect is unlikely to occur only in India. At the global level, the diversity of urban metrics could challenge regularities taken for granted. For example, it could well be that urbanization rates are more accurate in agrarian societies and in advanced economies but underestimated in between these two extremes. If so, the relationship between the urban population share and income per capita would have a different gradient relative to what is commonly accepted.

Recognition of this measurement problem has led to alternative approaches to defining the urban extent in ways that could ensure comparability across countries and over time. One approach is to rely on information from the ground, such as highly disaggregated data from population and economic censuses. An example of this more systematic approach is the paper by Rozenfeld and others (2011) on the Great Britain and the United States.

Another promising approach is the use of remote sensing technology to identify urban areas from outer space. A number of methods that rely on images taken by sensors aboard satellites have been proposed in this respect. Methods to identify cities from outer space can be broadly classified into three groups,
depending on their primary input data: (1) satellite images of land cover and of built-up characteristics; (2) satellite images of areas lit up at night; and (3) gridded population data created by using satellite images and other auxiliary information to redistribute aggregated data from population censuses.

To many, these approaches hold the promise of cross-country neutrality and comparability because they ensure global coverage, spatial granularity, and temporal consistency. But a consensus has yet to emerge on the most reliable method to identify urban areas. To date, then, the results have been heterogeneous. At the global level, the urban areas identified cover from less than 300,000 square kilometers to over 700,000 square kilometers, depending on the method (Potere and others 2009).

Some features of the input data and their interpretation may help explain the large variations in results when using these methods:

- Most analyses are conducted at the cell level, but a city is more than the sum of its cells. This remark goes back to the subjectivity of the concept of a city. The proper of an urban setting is the interlinking of cells in a broader spatial context. Assessing whether a cell is urban without referring to the surrounding cells is thus potentially misleading, and even more so when the urbanization process is messy (World Bank 2015).

- Analyses relying primarily on land cover classification algorithms can accurately identify built-up areas and infrastructure—the signature characteristics of urban areas. But cities include other land cover types such as parks and water bodies. The continuum between rural and urban areas poses a challenge to these algorithms as well (Mertes and others 2015; Pesaresi, Ehrlich, and others 2016; Schneider 2012).

- Analyses relying primarily on nightlight data capture the concentration of human activities in an unprecedented manner. However, problems such as blooming lead to biased estimates of the urban extent (Elvidge and others 2007; Elvidge and others 2009; Li and Zhou 2017). As access to electricity becomes more common and reliable, the pace of urbanization could be misinterpreted as well.

- When using gridded population data, one fundamental issue is the lack of reliable information on disaggregated administrative boundaries. Open-source boundaries are available mostly at the district or subdistrict level (GADM 2012). Population census data are at best georeferenced to these levels of administrative boundaries. And both population data and administrative boundaries tend to be out of date by several years.

To address these limitations, we develop a tractable method for identifying urban areas and illustrate its potential by applying it to India. The proposed method differs from the previous literature in three important ways:

- Instead of cells, we use cities, towns and villages, as officially defined, as our units of analysis. Because the available boundaries of administrative units tend to be unavailable, out of date, or erroneous, we rely on newly available georeferenced and digitized boundaries of administrative units at all levels, based on the India’s Administrative Atlas—2011 (ORGI 2011b). Although the analysis covers the entire universe of cities, towns, and villages, an in-depth assessment is conducted for a stratified sample of them. The sample design results in roughly three-quarters of the units being administratively urban and the rest administratively rural.
Subjective assessments of images provided by Google Earth are used to determine whether the cities, towns and villages in the sample are urban or rural in practice. Google Earth is an integrated collection of processed satellite, aerial, street view, and 3-D images. Using this combination of perspectives, a layperson can judge the characteristics of a specific location. Google Earth images have been used to crowd-source judgments on city safety, inequality, and gentrification (Hwang and Sampson 2014; Salesses, Schechtner, and Hildalgo 2013). They have also been used in the remote sensing literature to establish training sites and validation sites at the pixel level (Bontemps and others 2011; Schneider, Friedl, and Potere 2009; Yu and Gong 2012).

A regression analysis combining multiple sources of information is used to identify the correlates of units in the sample being urban in practice. We georeference population census data, which come from the ground, to all administrative boundaries in India. We also incorporate data from open-source satellite imagery, which come from outer space, on both built-up areas and nighttime lights. Bootstrapping and cross-validation are applied to choose the model specifications with higher prediction accuracy. These model specifications and the estimated coefficients are then used to predict whether each of the other units in the country is urban or rural in practice.

This exercise allows a more accurate assessment of the rate and speed of urbanization in India than was available before. Contrary to frequent claims, we do not find that India is substantially more urban at present than stated by census data. And the speed of urbanization over the last decade is only marginally higher than official statistics suggest. At the same time, however, we find a considerable misclassification of locations in the midrange between villages and state capitals.

The analysis also has implications beyond India. We find that the accuracy of the assessment increases substantially when combining data from the ground and from outer space. In the process, it appears that some of the most frequently used indicators of urban status can yield misleading results. Among the indicators from the ground, population density is a poor predictor of urban status. Among those from outer space, nighttime light intensity leads to a drastic overestimation of the speed of urbanization.

2. A representative sample of units

The official boundaries of cities, towns and villages are built on historical data and capture the best knowledge governments have of the spatial distribution of economic activities and people. Admittedly, these boundaries may be too small to account for large metropolitan areas, which typically include nearby suburban and exurban areas. But they define the constituency to which each local government provides services and is accountable. And they are the basic sample unit used by national statistical agencies to conduct censuses and surveys. Therefore, important information about cities, towns and villages can be captured at this level of spatial disaggregation.

The universe of administrative units

In this study, we take advantage of newly digitized boundaries of administrative units in India, now available all the way down to the town or village level. Although not official, these boundaries are based on India’s Administrative Atlas—2011 (ORGI 2011b). They were generated as part of a broader research project, the Spatial Database for South Asia (Li and others 2015). Construction of the boundaries required
scanning and georeferencing 6,598 physical maps: 35 maps of states and union territories, 640 district maps, and 5,923 subdistrict maps.

The subdistrict maps present the administrative boundaries of towns and villages in their jurisdiction. The location, size, and shape of these towns and villages were digitized in the form of 649,818 boundary polygons. Attribute such as location codes, names, and administrative types were added to the polygons. This study takes advantage of these attributions. Information on the administrative types of the cities, towns and villages allows a refined stratification strategy for sampling.

Data on administrative boundaries are adjusted for this study in three ways. First, in the original database multiple polygons were retained for towns that extended over more than one subdistrict. In this study, those multiple polygons are merged into a single administrative boundary for each town. This results in 649,649 units whose boundaries have a one-to-one correspondence with administrative units.

Second, some villages were presented as points on the source maps and their boundaries were unavailable, totally or partially—they are mostly in the mountainous or forest areas of four states (Arunachal Pradesh, Chhattisgarh, Himachal Pradesh, and Meghalaya). These villages are excluded from the analysis. And, third, population information was not available in the 2011 Census of India for some villages, which raises concerns about the quality of the corresponding official maps. These villages are excluded as well.

As a result of these mergers and exclusions, 564,052 administrative units with well-identified boundary polygons are retained as the universe for the analysis. However, for the purpose of extracting a sample of cities, towns and villages, the set of administrative units is further restricted to those belonging to the 21 largest states and union territories. These large states and union territories are covered by the high-quality household surveys whose data are needed for a proper stratification. Focusing exclusively on them further reduces the number of administrative units considered to 561,624.

The total population of these 561,624 polygons in 2011 was 1,174 million, or 3.0 percent less than reported by the 2011 Census for the country as a whole (ORGI 2011). Almost a third of this population gap—about 11.3 million people—is accounted for by the small states and union territories excluded from the analysis. Some of these small states and union territories are richer than the Indian average and some are poorer, which suggests that their exclusion from the analysis may not bias results much. The rest of the gap is associated with the villages excluded because their boundaries or population figures were unavailable. These villages are most likely rural in practice, and their population size is small—about 200 people at the median—which implies that they too are unlikely to be an important source of bias.

**A two-stage stratification process**

A two-stage disproportionate stratification approach is used to select the sample of administrative units to be inspected visually (figure 1). In the first stage, the universe is divided into eight strata by administrative category, following the definitions of the 2011 Census of India. Five of these categories—state capitals, other municipal corporations, municipalities, industrial towns, and other towns—are statutorily designated as urban. All state capitals are governed as municipal corporations; we use the term other municipal corporations to refer to the rest of the cities in this administrative category. Two of the categories—census towns and outgrowths—are statutorily rural but recognized by the 2011 Census as
urban. The eighth category is villages, or areas that are rural according to both the Constitution and the 2011 Census of India (ORGI 2011a).

A disproportionate sampling approach is used in the first stage. The 561,624 administrative units comprise 3,852 statutory urban units, 4,645 units that are urban in terms of the census, and 553,127 villages. If a proportionate sampling approach had been adopted, the share of statutory and census urban units in the sample would be a mere 2 percent. But villages are likely to have rural characteristics despite the limitations of administrative definitions, while misclassification is more likely for statutory and census urban units. It therefore makes sense to overrepresent the latter.

**Figure 1. Two-stage stratification sampling**

For each administrative category stratum, we choose a sample size sufficiently large to ensure the accuracy of subsequent analyses. There is a well-known trade-off in this respect. A small sample could fail to generate enough information, whereas a large sample would be costly to implement. The literature on survey design has developed a general method to handle this trade-off (Knottnerus 2003).

Building on the literature, we rely on the Central Limit Theorem and determine the sample size of each stratum using

\[
 n_i^* = \frac{z^2p_i(1-p_i)}{e^4} \left(1 + \frac{z^2p_i(1-p_i)}{N_i e_i^2} \right) \tag{1}
\]

where \( n_i^* \) is the sample size for administrative category \( i \); \( N_i \) is the total number of units belonging to the category; and \( p_i \) is the prior probability of the unit being urban in practice. In the absence of other information, the prior probability \( p_i \) should be set equal to 0.5. But administrative categories provide information on the likelihood that a unit is urban or rural in practice. On average, the values of \( p_i \) should be correlated with the order of these administrative categories in the rural-urban gradation. Our priors build on this expected correlation (table 1). The visual analysis of a random sample of 250 administrative units supports these priors.
Table 1. The sample and its distribution across administrative categories

| Administrative Category                        | N     | Margin of error | Prior estimate on the probability of being urban | n* |
|-----------------------------------------------|-------|-----------------|-------------------------------------------------|----|
| Capitals and Municipal Corporations           | 147   | 0.05            | 0.97                                             | 34 |
| Municipalities                                | 905   | 0.05            | 0.90                                             | 120|
| Industrial Towns                              | 31    | 0.05            | 0.90                                             | 25 |
| Other Towns                                   | 2,769 | 0.05            | 0.80                                             | 226|
| Census Towns                                  | 3,710 | 0.05            | 0.71                                             | 292|
| Outgrowths                                    | 935   | 0.05            | 0.60                                             | 264|
| Villages                                      | 553,127 | 0.03         | 0.10                                             | 315|
| **Total**                                     | 561,624 |                | **Mean probability of being urban** | 0.11 |
| **Relative standard error**                   |       |                 | **Relative standard error**                      | 0.08 |

In equation (1) the term $e_i$ is the margin of error allowed for administrative category $i$, and $z$ indicates the boundaries of the chosen confidence interval. For a conventional 95 percent confidence, the value of $z$ is 1.96, and $e_i$ is equivalent to 1.96 times the standard error of $p_i$. We assume that the standard error of the prior probability of being urban is much smaller for villages than for the seven other administrative categories. Therefore, $e_i$ is set at 3 percent for villages and at 5 percent for all other strata.

Based on these assumptions, the total sample size needs to be 1,277 or higher: 405 units statutorily urban; 557 units urban in terms of the census; and 315 villages. For the full sample we achieve a relative standard error of 8 percent, which is below the 10 percent threshold required for a survey design to be considered good (UN 2005).

In the second stage, each of the eight strata is divided into three substrata defined by average consumption per capita. Consumption data are from the household consumer expenditure module of the 68th round of National Sample Survey of India (NSSO 2012). The data quality is considerably better for the 21 largest states and union territories than for the smaller ones. The monthly household expenditure is divided by household size to estimate consumption per capita. To account for the spatial variation in prices, the result is deflated by a year-specific price index that differs across states and union territories and between rural and urban areas. The 21 states and union territories are ranked on the basis of their average consumption per capita, and they are divided into three groups of equal size: richer, intermediate, and poorer.

The sample for each administrative category is split evenly across these three groups of states and union territories. For example, of the 264 units retained in the outgrowth stratum of the sample, 88 would be from richer states and union territories, another 88 units from the intermediate group, and the remaining 88 from the poorer group.

The sampling process is completed by conducting random draws without replacement for each of the 24 substrata. Our final sample with 1,277 administrative units is relatively evenly distributed across India’s
territory (map 1). These selected units cover five of the six regions of the country—northern, central, eastern, western, and southern—relatively well. Only for the northeastern region is the coverage of the sample poor. This is due to the shortcomings of village boundaries in the states of Arunachal Pradesh and Meghalaya and the low quality of the household surveys in the states of Manipur, Mizoram, Tripura, Meghalaya, and Assam.

Map 1. The spatial distribution of sample units

3. The urban status of units in the sample

Cities have diverse characteristics both across and within countries. India has a continuum of urban and rural spaces, ranging from a glamorous metropolis such as Mumbai, capital of Maharashtra, to a tiny shabby town such as Bagula in West Bengal, to any among hundreds of thousands of villages. An objective method based on one or two metrics with fixed thresholds for distinguishing between urban and rural areas may not adequately capture this diversity. A richer combination of criteria would help describe the multifaceted nature of a city’s environment, but the joint interpretation of these criteria would invariably involve an element of judgment. Building on this intuition, we rely on subjective assessments to determine whether each administrative unit in the sample is urban or rural in practice.
A subjective but structured process

Subjective assessments have been used in the economic literature to complement more objective approaches. For example, in measuring individual well-being a subjective assessment has been adopted on the grounds that everybody has his or her own views on what a good life looks like, and what makes a good life may touch on dimensions for which no reliable indicator is available (Frey and Stutzer 2002; Veenhoven 2004). Deciding whether a specific place is a city poses challenges similar to those posed by measuring well-being.

For a subjective assessment to be thorough, researchers should ideally visit each city, town, and village in the sample and “see” how they look like. However, the time and cost implications of visiting 1,277 locations scattered across a large country are dissuasive. Building on the literature, we adopt a second-best approach, which is to rely on high-quality, open-source images by Google Earth.

Google Earth is an integrated collection of processed satellite, aerial, street view, and 3-D images. A zoom-in option is available, and ground views and different time views are also available whenever possible. This combination of perspectives enables a layperson to judge the characteristics of a specific location. No special training in remote sensing is needed.

We overlaid the digitized boundary of each sample unit on top of Google Earth images and conducted a subjective assessment of the images falling within the corresponding boundary. For the vast majority of units, we rely on images dated as having been taken in 2010–12 to be consistent with the population data from the 2011 Census of India. However, when the images for these three years are of poor quality, we complement them with images from other years.

The sample of cities, towns, and villages was randomly split into three subsets, which were in turn allocated to three research analysts. All research analysts relied on the same decision tree to independently classify the units allocated to them as urban or rural.

The common decision tree considers characteristics observable from Google Earth images: distribution of land cover types, relationship between buildings, characteristics of buildings, prevalence of transportation networks, and availability of amenities (schools, universities, hospitals, cultural sites, etc.). The selection of characteristics for the decision tree builds on the economic geography and urban economics literature:

- The simple monocentric model of land use predicts that the density of construction declines as one moves away from a city center to more rural areas (Alonso 1964; Mills 1967; Muth 1969). Extensions of this model dealing with polycentric cities and scattered development have not fundamentally contradicted this prediction (Bueckner 1987; Duranton and Puga 2015). We thus expect both the share of built-up area and the density of buildings to be higher in urban areas.

- Lumpy infrastructure investments facilitating access to markets are critically important in giving rise to the concentration of economic activities (Duranton and Puga 2004; Fujita, Krugman and Venables 2001; Krugman 1991; Scotchmer 2002). We therefore expect both inter- and intracity transport networks to be more visible in urban areas.

- Services and amenities are crucial in explaining where firms prefer to operate and households prefer to live. The self-selection of households across locations accounts for 40–50 percent of productivity differences across cities (Ahlfeldt and others 2015; Combes and others 2010; Straszheim 1987).
greater presence of amenities such as schools, hospitals, and government services can thus be expected in urban areas.

The decision tree is implemented in three steps (figure 2). First, the research analyst focuses on land cover. If the built-up area appears to exceed 30 percent the unit is likely to be urban, and it is likely to be rural if the built-up area is less than 10 percent. Between these two thresholds, the assessment is inconclusive. Next, the research analyst adjusts this preliminary judgment based on the relationship among buildings. Compactness or clustering of buildings reinforces the belief or increases the likelihood that the unit is urban, whereas a scattered pattern of buildings suggests that the unit is rural. Finally, the researcher zooms in and pulls out ground views (if available) to assess the availability of amenities, high-quality buildings, and transportation networks. The availability of some of these structures confirms that the unit is urban in practice.

**Figure 2. The decision tree for the subjective assessments is implemented in three steps**

Subjective assessments were cross-checked to ensure the robustness of the results. A randomly selected subset of 10 units was reallocated from the original research analyst to a different one to verify that their assessments were the same. When one research analyst could not reach an unambiguous conclusion on a specific unit, another analyst was called in to consult and make a joint decision. For 46 units, an unambiguous classification seemed out of reach; the most plausible choice was made, but the units were
flagged. As a robustness check, the subsequent analysis was conducted both including and excluding the flagged units.

**The resulting classification of the sample**

Subjective assessments confirm the administrative classification of most cities, towns, and villages as urban or rural, but there are also discrepancies. For example, Jigani of Karnataka is administratively rural but meets most of the criteria to be considered urban in practice. The built-up area covers a substantial share of the unit’s surface, the buildings are compact, and a network of roads and signs of amenities are clearly visible. Conversely, Bahadurganj of Bihar is administratively urban, but its cropland areas are vast, built-up area is relatively small, buildings are scattered, amenities are few, and the road network is limited (map 2).

**Map 2. Examples of discrepancies with the administrative classification**

- **Jigani (Karnataka)** is reassessed as urban

- **Bahadurganj (Bihar)** is reassessed as rural
According to the subjective assessments, 49 percent of the 1,277 units in the sample are urban in practice, and 51 percent are rural (table 2). Therefore, the sampling strategy does ensure a balanced representation of rural and urban units. At the same time, the exercise reveals substantial differences between the administrative classification of units in the sample and their subjective assessment as urban or rural. Discrepancies are greater for units deemed urban in terms of the census, many of which are reassessed as rural. A nonnegligible fraction of statutorily urban units is also reclassified as rural. But there are no substantial differences for villages, whose rural status is almost consistently confirmed by the subjective assessments.

Table 2. Subjective assessments versus administrative classification

| Administrative classification | Assessed urban (percent) | Assessed rural (percent) | Total (percent) |
|------------------------------|--------------------------|--------------------------|----------------|
| Statutory urban units        | 24.9                     | 6.8                      | 31.7           |
| Census urban unit            | 24.0                     | 19.6                     | 43.6           |
| Census rural unit            | 0.4                      | 24.3                     | 24.7           |
| Total                        | 49.3                     | 50.7                     | 100.0          |

In light of these discrepancies, a relevant question is whether the size of the sample is large enough to trust the results. In setting the sample size, we relied on the Central Limit Theorem, which implies an asymptotic normal distribution for the statistic of interest. But with a relatively small sample size, the distribution could be different. One way to address this concern is to use the bootstrap method, which generates an approximation of the distribution through a Monte Carlo simulation, in which the sampling is carried out from the fitted distribution of the observed data (Cameron and Trivedi 2005).

The bootstrap method is applied to the sample of 1,277 units, for which we know the urban status based on both the administrative classification and subjective assessment. In each replication, we randomly draw units from this sample with replacement, until arriving at the same sample size as before (1,277). The drawing is done independently for each administrative stratum. For the new sample obtained in each replication, the share classified as urban is computed on the basis of the subjective assessments. After replicating 10,000 times, we estimate the distribution of this urban share, including its mean, standard errors, and confidence intervals (figure 3).

The bootstrap analysis confirms that the sample is sufficiently large for each administrative stratum. For capital cities and municipal corporations, as well as for villages, bootstrap standard errors are below 1 percent. At 5 percent, the standard error is larger for industrial towns, reflecting the small size of the sample for this stratum. However, because the sample includes 25 of 31 units in this stratum, the sampling does not affect the accuracy of the results. Standard errors are below 3 percent for all other administrative strata.
4. Out-of-sample prediction of urban status

The urban status of units derived from the subjective assessments can be used to infer the urban status of the other units in the universe. This involves first learning from the sample and then using the results to make out-of-sample predictions. There are multiple ways to do this, however, and a key difference between more traditional and more modern approaches is whether a specific generation process is assumed for the data. We did assume that such process existed when drawing the sample, as we hypothesized that there was a probability of a unit being urban and that its distribution was known. Consistent with this hypothesis, here we analyze the correlates of urban status using a regression analysis that builds on the classical statistical tradition of data modeling. However, we subsequently check the robustness of the findings using algorithmic modeling.

*Data from the ground and from outer space*

The regression analysis assumes that the likelihood of being urban follows a known stochastic distribution, which allows fitting a model to data from the 1,277 units in the sample. We choose as explanatory variables in this model a set of indicators which appear recurrently in the urban literature and are observable for all administrative units in the country. We then use the estimated model to predict the likelihood of being urban for all the remaining administrative units in the universe.

The chosen set of indicators is relatively narrow to ensure that the proposed approach can be applied to other country settings. Because of its size and diversity, as well as the overall “messiness” of its urbanization process, India provides a very good testing ground for our approach. But measuring the
urban extent is a challenge across the world, and results that are internationally comparable is the goal. This requires focusing on a set of key indicators available for the vast majority of countries.

At the same time, the set of indicators chosen is broader than is generally the case when measuring the urban extent. Official estimates of urbanization rates by national governments and international organizations rely almost exclusively on data from the ground. The key indicators considered are population size, population density and administrative category. Conversely, products recently developed by the remote-sensing literature focus almost exclusively on data from outer space. Built-up cover and nighttime lights are the preferred indicators in this respect. But rather than choosing one set of indicators or the other, our analysis combines data from the ground with data from outer space.

Data from the ground are taken from the Spatial Database for South Asia, which georeferenced both the 2001 and 2011 Census of India to the digitized official boundaries of administrative units (Li and others 2015). A range of georeferenced indicators from the ground are available for 2001 and 2011 across all units in the country. The indicators are population size, population density, employment structure, access to services, dependency ratio, and literacy rate. The administrative type of each unit is also available, as described earlier.

Following the practice of national governments and international organizations, we use population size and population density at the town and village level as our explanatory variables from the ground. This choice is consistent with that of studies for more advanced economies that rely on high-quality microdata from the ground (Michaels, Rauch, and Redding 2012; Rozenfeld and others 2011). Population density is computed as the ratio between the population of the administrative unit and the size of the area within the corresponding polygon boundary. In some specifications, we also consider the units’ administrative type as defined by the 2011 Census.

Data from outer space are selected building on the remote sensing literature, as well as on recent developments in the economics literature that use satellite imagery (Burchfield and others 2006; Donaldson and Storeygard 2016; Henderson, Storeygard, and Weil 2012; Jean and others 2016). Our indicators are derived from the best open-source satellite imagery available for 2001 and 2011, which are the years when population was counted. The indicators are computed by overlaying the boundaries of towns and villages on the processed built-up or nighttime light imagery.

In line with the literature, we choose built-up share and lit-up share at the town or village level as the two key indicators from outer space. Built-up share is computed as the ratio between the built-up area in a specific polygon boundary and the total area of the polygon. Lit-up share is the share of the surface of the polygon whose nighttime light intensity exceeds some critical threshold. In some specifications, built-up area and lit-up area—without dividing by the surface of the polygon—are used as alternative indicators from outer space.

The assessment of land cover has improved rapidly over the last decade thanks to more precise spatial, spectral and radiometric resolutions of satellite imageries, and to much greater computational power to process the data. Three open-source products on built-up cover stand out as providing the best quality at the global scale:

- **MODIS** relies on images taken by sensors aboard satellites from the Earth Observing System program. These images—known as the Land Cover Type Yearly Grid—are provided by the Land Processes Distributed Active Archive Center managed by the U.S. National Aeronautics and Space
Administration. The product has a resolution of about 500 meters per cell and is available from 2001 onward (Friedl and others 2010).

- **GHSL-Landsat** is derived from images taken by sensors aboard satellites that are part of the Landsat program. As one of the first civilian satellite programs on land cover, Landsat has produced the longest continuous space-based record of the Earth’s land surface. Data for the years 1975, 1990, 2001, and 2013/14 are available from the Global Human Settlement Layer project (Pesaresi, Ehrlich, and others 2016). The resolution of the product is about 38 meters per cell.

- **GUF-TerraSAR** is based on radar images taken by sensors aboard the twin satellites TerraSAR-X and TanSAR, which are part of the more recent WorldDEM program. Reliance on radar images implies lower sensitivity to weather conditions, cloudiness and vegetation density. Data for 2011 are available from the German Aerospace Center at a resolution of about 12 meters per cell (Esch and others 2017).

After a careful comparison of the three data sources, we selected the GHSL-Landsat product for this study. This choice is consistent with that of Burchfield and others (2006), who use the Landsat product to study urban development in the United States.

In the case of India, MODIS is limited by its coarse resolution. In particular, it does not capture built-up cover in forest areas as well as the other two data sources. Data from both GHSL-Landsat and GUF-TerraSAR yield very similar results while GUF-TerraSAR has a higher resolution. The correlation coefficient between the two sources for the built-up share at the town or village level is 0.96. In addition, GHSL-Landsat data are available for multiple years, whereas GUF-TerraSAR data are available only for 2011. This is a constraint, because we want to estimate not only the urban extent, but also the speed of urbanization over time. We use GHSL-Landsat data for 2001 and 2013/14 for the census years 2001 and 2011, respectively. But we check that the results remain the same when using GUF-TerraSAR data for the year 2011.

For the measurement of the lit-up area we rely on the Global Radiance Calibrated Nighttime Lights product. This product is derived from images taken by the Operational Linescan System (OLS) sensors aboard satellites from the U.S. Air Force Defense Meteorological Satellite Program (DMSP). Data are available from the National Geophysical Data Center at the National Oceanic and Atmospheric Administration (NOAA).

This is not the only product available for nighttime light images, but it has a higher quality than its competitors because it captures stable lights, addresses the saturation problem, and is intercalibrated to reduce inconsistencies across satellites (Elvidge and others 2009; Hsu and others 2015; NOAA 2014). The intensity of nighttime light is reported in Digital Numbers. We use data for 2001 and 2011, thereby matching the census years. Because the data are not generated in the same way in both years, we apply the intracalibration formula proposed by Hsu and others (2015) to ensure consistency over time.

A key challenge when using data on nighttime light is that even areas without human activity may appear to be lit up (Henderson and others 2003; Li and Zhou 2017; Zhou and others 2014). This may be due to blooming, or to the reflection of light from surrounding water and other land covers, or to the sensitivity of the sensors, or to the cumulative effects of geo-referencing errors made in the capture and composting of satellite imagery. To avoid overestimation of the lit-up share, a threshold is generally set up for nighttime light intensity, and only observations above the threshold are considered urban. However, the
appropriate threshold varies considerably across countries, and it is correlated with their levels of economic development.

To identify the appropriate threshold in India’s case, we review a dozen studies that together present 16 threshold estimates at the country level and 13 at the city level. We then match their results with data on gross domestic product (GDP) per capita at the country and city level (Brookings Institution 2015; WDI 2016). Consistent with the literature, we find a positive and statistically significant correlation between the identified thresholds and the corresponding incomes per capita (figure 4).

**Figure 4. The nighttime light threshold increases with income per capita**

Given India’s GDP per capita in 2001 and 2011, the estimated relationship suggests that the appropriate thresholds in its case are 14.5 and 15 Digital Numbers respectively. For each of the two years, we delineate the cells with a nighttime light intensity above the relevant threshold and create a new file on urban light for the whole of India. We then overlay the boundaries of towns and villages over this file and compute the lit-up area and the lit-up share of every unit.

As expected, the mean values of the six indicators chosen for this study differ between the universe and the sample because, by design, the sample overrepresents the units that are more likely to be urban in practice (table 3). Mean values are also higher in 2011 than in 2001, and this holds for all six indicators and for both the sample and the universe. This consistent upward trend is a clear indication that India’s urbanization rate must be increasing. The only question is by how much.
### Table 3. Summary statistics of key indicators from the ground and from outer space

|                  | Obs  | Mean   | Std. Dev. | Mean   | Std. Dev. |
|------------------|------|--------|-----------|--------|-----------|
|                  |      | 2001   |           | 2011   |           |
| **Sample**       |      |        |           |        |           |
| Population size  | 1,277| 38.73  | 371.39    | 46.79  | 392.48    |
| (thousands)      |      |        |           |        |           |
| Population density | 1,277| 2252   | 3983      | 3085   | 6057      |
| (people per square kilometer) |      |        |           |        |           |
| Built-up share   | 1,277| 9.91   | 17.46     | 13.05  | 19.99     |
| (percent)        |      |        |           |        |           |
| Lit-up share     | 1,277| 34.06  | 41.92     | 52.97  | 45.45     |
| (percent)        |      |        |           |        |           |
| Built-up area    | 1,277| 1.65   | 10.80     | 2.07   | 11.84     |
| (square kilometers) |      |        |           |        |           |
| Lit-up area      | 1,277| 4.74   | 20.87     | 6.83   | 23.24     |
| (square kilometers) |      |        |           |        |           |
| **Universe**     |      |        |           |        |           |
| Population size  | 564,052| 1.75 | 27.59   | 2.10  | 33.53    |
| (thousands)      |      |        |           |        |           |
| Population density | 564,052| 580  | 2339   | 693   | 2700     |
| (people per square kilometer) |      |        |           |        |           |
| Built-up share   | 564,052| 0.50 | 3.47   | 0.81  | 4.43     |
| (percent)        |      |        |           |        |           |
| Lit-up share     | 564,052| 2.57 | 14.18  | 8.57  | 25.93     |
| (percent)        |      |        |           |        |           |
| Built-up area    | 564,052| 0.04 | 1.13   | 0.05  | 1.35     |
| (square kilometers) |      |        |           |        |           |
| Lit-up area      | 564,052| 0.13 | 2.46   | 0.40  | 3.09     |
| (square kilometers) |      |        |           |        |           |

#### A model for the probability of being urban

Following the data modeling tradition, we specify an explicit probabilistic model on whether an administrative unit is urban or rural in practice:

\[
U_i = \begin{cases} 
1 & \text{with probability } p_i \\
0 & \text{with probability } 1 - p_i 
\end{cases} 
\]

\[
p_i = \Pr[U_i = 1 | X_i] = F(\alpha + X_i \cdot B)
\]

where \(U_i\) is the subjective assessment of unit \(i\), \(p_i\) is the expected likelihood that a unit is urban given the values of its key indicators \(X_i\), and \(F(\cdot)\) is the cumulative distribution function. The estimation of equation (2) requires that a functional form be assumed for \(F(\cdot)\). Here we use the standard logit model linking the urban status of a unit in a specific year with the values of its key indicators at a similar point in time. The justification for this choice, as opposed to the also standard Probit model, is its higher goodness of fit and better prediction accuracy. The results obtained using the Probit model are available upon request.

A linear function of \(X_i\) is assumed for the function \(F(\cdot)\). The benchmark specification combines data from the ground and from outer space using population size, population density, built-up share, and lit-up share as the key indicators. The model is also run retaining only indicators from the ground or from outer space, either individually or combined (table 4).
Table 4. Logit regressions on the probability of being urban

|                        | Benchmark specification | Data from the ground | Data from outer space |
|------------------------|-------------------------|----------------------|-----------------------|
| **Population**         | 0.108*** (0.009)        | 0.104*** (0.008)     | 0.0925*** (0.008)     |
| **Population density** | 1.63e-05 (0.000)        | 0.00034*** (0.000)   | 0.000146*** (0.000)   |
| **Built-up share**     | 0.0725*** (0.009)       |                      | 0.147*** (0.011)      |
| **Lit-up share**       | 0.0213*** (0.002)       |                      | 0.0276*** (0.002)     |
| **Pseudo R2**          | 0.50                    | 0.26                 | 0.30                  |
| **Log likelihood**     | -442                    | -658                 | -622                  |
| **Bayesian information criterion** | 920                  | 1330                 | 1257                  |
| **Correctly classified** | 84.3                  | 72.9                 | 80.0                  |

Note: Standard errors are reported in parentheses, with one, two and three asterisks indicating significance at the 10, 5 and 1 percent level.
All coefficients for the key indicators have the expected signs, and most of them are statistically significant regardless of the specification chosen. The exception is population density, which loses significance in the benchmark regression. The estimated marginal effects, not reported here, also appear to be reasonable in magnitude. However, these marginal effects should not be interpreted literally because omitted-variable bias is likely in parsimonious specifications like the ones chosen for this study.

A more telling evaluation of the probabilistic model proposed in this paper is based on its goodness of fit. The metrics used to assess goodness of fit are log likelihood, pseudo R-squared, the Bayesian information criterion (BIC), and correct classification rates. Based on all four metrics, the benchmark specification has the highest goodness of fit, closely followed by those that include the built-up share indicator.

A bootstrap analysis helps us evaluate whether these differences in performance between specifications are statistically significant because it allows us to compute standard errors and confidence intervals for each of the four metrics used to assess goodness of fit (Cameron and Trivedi 2005; Shmueli 2010; Wooldridge 2002). The bootstrap analysis follows the same procedure used to assess the quality of the sample—that is, we resample the 1,277 units with replacements and replicate the whole exercise 10,000 times. The results confirm that combining data from the ground and from outer space produces better results than relying exclusively on one type of data (figure 5). For readability, only the bootstrap results for log likelihood and correct classification rate are presented; bootstrap results for the other metrics are available upon request.

**Figure 5. Goodness of fit is highest when combining data from the ground and from outer space**

The most critical step in the analysis is the classifications of the out-sample administrative units. Inferring prediction power from explanatory power is potentially misleading because there is a risk of overfitting the sample. With this concern in mind, we evaluate the different specifications on their prediction accuracy in addition to their goodness of fit.
A popular way of doing this is to separate the sample into a training set, on which the model is reestimated, and a validation set, on which the model's performance is evaluated. Following this approach, we apply the Monte Carlo cross-validation method, randomly splitting the sample into training and validation sets—in a fixed proportion—multiple times. Asymptotically, the distribution of the estimated prediction accuracy has been shown to converge toward that of exhaustive cross-validation (Picard and Cook 1984; Shao 1993; Shmueli 2010).

In this study, at each replication we randomly draw 70 percent of the sample (the training set), estimate all seven specifications on it, and use the results to assess prediction accuracy on the remaining 30 percent of the sample (the validation set). We repeat the exercise 10,000 times and compute the mean value and the confidence interval of two metrics. The first one is the percentage of observations correctly predicted, which is directly comparable to the percentage correctly classified used for the assessment of goodness of fit. A second metric used is the Akaike information criterion (AIC), which resembles BIC but is derived from a predictive viewpoint (Shmueli 2010; Sober 2002).

The results of this Monte Carlo cross-validation confirm that the benchmark model has the highest prediction accuracy (figure 6). Again, the specifications based exclusively on key indicators from outer space are a close second. For the AIC, the difference with the benchmark specification is not statistically significant. However, when considering the percentage of observations correctly predicted, the benchmark specification does significantly better than the others.

**Figure 6. Prediction accuracy is highest when combining data from the ground and from outer space**

Following the literature, we also compute the shrinkage, which decomposes the potential prediction bias into out-sample bias, in-sample bias, and overfitting (Bilger and Manning 2015; Copas 1983). The results,
available upon request, confirm that the best results overall are associated with the benchmark specification.

Robustness checks

A series of robustness checks confirms that combining data from the ground and from outer space leads to both a higher goodness of fit and better prediction accuracy than any of the two sets of indicators taken separately. The checks include:

- **Alternative indicators for data from outer space.** When built-up share is replaced by built-up area and lit-up share by lit-up area, the performance of the model on the probability of being urban declines in terms of both goodness of fit and prediction accuracy.

- **Dummies for administrative type.** The introduction of dummies only slightly improves goodness of fit and prediction accuracy. For example, the percentage of units correctly predicted in the benchmark specification increases by a mere 0.7 percentage points.

- **Nonlinear probability function.** When a quadratic term is introduced for each of the four key indicators, the goodness of fit, as expected, increases. This comes at the cost of out-of-sample bias because of overfitting, as shown by the shrinkage indicator.

- **Blending of specifications.** Following the “ensemble” method proposed by the statistical learning literature, we consider that a unit is urban if it is predicted to be so by at least four of the seven specifications. The percentage correctly predicted does not improve relative to the benchmark specification.

- **Inconclusive subjective assessments.** Because research analysts were unable to unambiguously classify 46 units in the sample, we reestimate the probabilistic model removing these observations, and the results barely change.

- **Potentially inaccurate polygons.** The subjective assessments suggested that the polygons for a few units—mainly outgrowths and villages—may not be appropriately geo-referenced. We had replaced these polygons when compiling the sample of 1,227 units. A subjective assessment for these units and a prediction of their status using the regression model reveals that the out-of-sample prediction results are consistent with the subjective assessments.

These robustness checks, whose results are available upon request, give us confidence that the benchmark specification of the probabilistic model is a credible instrument for predicting the urban status of the broader universe of administrative units in India.

We also run a robustness check based on a different approach to statistical learning. In the analyses described earlier, a stochastic model for the data generation process was explicitly assumed. But a literature has emerged recently that dispenses with this assumption and proposes algorithmic modeling. This approach has gained prominence over last the two decades thanks to the rapid increase in computing power and the availability of big data. In computer science, the method is known as machine learning. Algorithmic modeling uses tools such as classification trees, in which data are split into groups at various nodes, each defined by a threshold for an explanatory variable. Compared with data modeling, these tools do not lend themselves to straightforward interpretation. But they can offer higher prediction accuracy (Breiman 2001b; James and others 2013).
The random forests method we use works as a combination of individual classification trees. Each of these trees focuses only on a subset of the explanatory variables to develop its nodes. That way, the classification trees are not correlated and offer different information on the underlying data generation process (Breiman 2001a; James and others 2013). Implementation of the random forests method is based on the package provided by Liaw and Wiener (2002) for the R program. As before, to compute prediction accuracy we split the sample between a training set and a validation set and repeat the exercise 10,000 times to estimate the mean and the confidence interval of the relevant metrics.

The benchmark specification under the logit regression framework performs as well as the random forests method, and the results from the two methods are highly correlated. The share of out-of-sample units correctly predicted when using the random forests method is 85.2 percent, or only 1 percentage point higher than when using the benchmark specification. This difference is less than the estimated standard error, which is around 1.4 percent under both the benchmark probabilistic model and the random forests method.

5. The rate and speed of urbanization

The probabilistic model built from the sample of 1,277 units using subjective assessments can be used to predict the urban status of all 564,052 units in the universe of India in 2011. The key indicators for each unit are plugged into the model to compute the unit’s likelihood of being urban. If this likelihood is above 0.5, the unit is deemed urban; otherwise, it is considered rural. Combining indicators from the ground and from outer space, this procedure yields an urbanization rate that is almost identical to that in the 2011 Census of India. However, this similarity hides important discrepancies across administrative types, especially in the midrange between villages and municipal corporations. The exercise also shows that relying exclusively on one type of indicators, such as nighttime light, can yield potentially misleading results for India’s rate and speed of urbanization.

The urbanization rate in 2011

According to the 2011 Census of India, the share of the population living in urban areas was 31.2 percent. But this figure is computed for a set of units broader than our universe because we excluded small states and union territories as well as villages for which information was missing. If only the 564,052 units considered in our analyses were retained, the 2011 Census of India would yield an urbanization rate of 31.6 percent. This is almost identical to the 31.4 percent rate predicted by the benchmark specification of the probabilistic model (figure 7).

Predictions using other specifications vary, but they also fall within a close range of the official estimate by the 2011 Census of India. The urbanization rate varies between 27.0 percent using built-up share as the only key indicator and 33.9 percent using only population size. The predicted urbanization rate is similar—31.8 percent—when using the random forests method.
Figure 7. The predicted urbanization rate of India in 2011 varies little across specifications
These results are in stark contrast with those reported in recent studies. Predictions based on different cutoffs for population size and population density yield urbanization rates that range from 47.2 to 78.0 percent (IDFC Institute 2015). A measure based on an agglomeration index that combines population density, population of the nearby main center, and travel time to that center suggests an urbanization rate of 55.3 percent (Uchida and Nelson 2009; World Bank 2015). And measures based on satellite imagery and agglomeration criteria produce urbanization rates of from 37.5 to 63.1 percent (Denis and Marius-Gnanou 2011). All of these results exceed our highest estimate.

**Gaps at the subnational level**

Although we find no major discrepancy with the official urbanization rate for India as a whole, our results suggest that the 2011 Census of India misclassified large numbers of units and people. Based on the benchmark specification, the misclassification affected 62 million people.

The misclassification is substantial among units falling within three of India’s administrative types (figure 8). Among statutory urban units, the gap is highest for other towns. According to our assessment, 1,317 of 2,861 lack sufficient urban characteristics and so should be considered rural. Thus 17 million people are incorrectly considered urban residents.

![Figure 8. Misclassification of urbanization in India is highest among three administrative types](image)

Misclassification is also substantial among the census towns of India. These units are administratively rural, but they meet three criteria to be identified as urban areas according to the 2011 Census of India. They have more than 5,000 inhabitants, a population density exceeding 400 persons per square kilometer, and over 75 percent of their male working population is engaged in non-agricultural activities. We find
that a considerable share of census towns has sufficient urban characteristics, but this is not true for all of them. Out of 3,847 “census towns”, 1,674 had mainly rural characteristics, resulting in another 14 million people being erroneously classified as urban. This marks a difference with other studies, which take for granted that all of India’s census towns are urban.

The third administrative type suffering from substantial misclassification is villages. The gap may seem small in relative terms, as discrepancies affect only 4,708 of 555,292 villages. But this apparently small gap has nontrivial implications when estimating the urban population share. Overall, 31 million people lived in villages with urban characteristics in 2011, but they were classified as living in rural areas by the census.

The substantial misclassification of people observed in other towns, census towns, and villages is corroborated by predictions based on other specifications of the probabilistic model, as well as by the classification based on the random forests method. Measured in millions of people, both the preferred specification in the probabilistic model and the random forests method yield very similar results.

However, the other specifications of the probabilistic model indicate that the number of people misclassified as urban or rural could be very large (figure 9). For example, the specifications including only built-up share or lit-up share would suggest that over 35 million people from other towns were misclassified as urban residents. Similarly, the specification relying only on population size implies that 54 million village residents were in urban areas. If applying the model using only the lit-up share, the figure would climb to 84 million.

Figure 9. Population misclassification depends on the method considered

Note: For each administrative type, misclassification by India’s census is computed as the difference between the urban population reported by the 2011 Census of India and that estimated by the particular specification of the probabilistic model (or by the random forests method).
Despite the similarity between our aggregate urbanization rate and that of the 2011 Census of India, there are important discrepancies at the state level. Three coastal states (Kerala, Andhra Pradesh, and Maharashtra) and two states (Haryana and Punjab) between Delhi and the India-Pakistan border have large numbers of villages with urban characteristics. Based on the benchmark specification of the probabilistic model, 16 million people from these five states were misclassified as living in rural areas by the 2011 Census of India. In contrast, many other towns and census towns lack urban characteristics in Madhya Pradesh, Uttar Pradesh, Orissa and Chhattisgarh. In all, 11 million people from these four states were misclassified as urban residents by the census.

The speed of urbanization over 2001–11

The benchmark specification of the probabilistic model can be used to predict the urban status of all 564,052 units in 2001. The process is the same as before, but the key indicators are from 2001. Computing the speed of urbanization over 2001–11 reveals that India urbanized slightly faster than acknowledged by population censuses (figure 10).

Figure 10. The speed of urbanization in India depends on the method considered

According to the censuses, India’s urbanization rate increased from 27.4 to 31.6 percent from 2001 to 2011, or 4.2 percentage points. With the benchmark specification, the predicted increase is from 26.6 to 31.4 percent, or 4.8 percentage points. Overall, the difference is small. And it is only slightly higher when using the random forests method, which implies that the urbanization rate would have increased from 27.0 to 32.2 percent, or 5.2 percentage points.

The differences become much larger when relying only on indicators from the ground or from outer space without combining them. For example, using only indicators from the ground, the speed of urbanization is similar to that obtained when relying on the benchmark specification. But the share of urban population
would be higher in both 2001 and 2011. The speed of urbanization would also be the same in both years when focusing only on the built-up share, but the urbanization rate would be much lower than using the benchmark specification. And if the lit-up share were the only key indicator used, the change in the urbanization rate over a decade would be 11.3 percentage points, which is more than double the estimate derived from applying the preferred specification to the probabilistic model.

6. Conclusion

The analysis in this paper yields a more accurate picture of urbanization in India than was previously available. We confirm that India’s urbanization rate in 2011 was close to that reported by the Census of India. However, this finding contrasts sharply with those of several studies reporting much higher urbanization rates (Denis and Marius-Gnanou 2011; IDFC Institute 2015; Uchida and Nelson, 2009; World Bank 2015). We also find substantial gaps between the official classification and our results for several administrative categories of the census. Overall, the misclassification affects 62 million people, which is roughly equivalent to the population of France or the United Kingdom.

Among statutory urban categories, misclassification is highest for other towns. Although a large share of them have sufficient urban characteristics, many should be considered rural. The misclassification rate is much lower for villages. But given that there are more than half a million of them, the absolute number of misclassifications has non-trivial implications.

Another important finding concerns the so-called census towns of India, which are administratively rural but meet three criteria to be identified as urban areas according to the population census. In line with the population census and a growing literature, we confirm that most census towns are urban in practice. But this is not true of all of them.

Finally, our analysis shows that relying only on indicators from the ground or from outer space results in substantially different rates and paces of urbanization. For example, if the lit-up share were the only key indicator used, the change in India’s urbanization rate over a decade would be 11.3 percentage points, which is more than double the figure obtained by combining data from the ground and from outer space.

Although this analysis centers on India, it has broader implications. Several recent studies have made a special effort to link the identification of urban areas more closely to the functionalities of cities. Examples of this new generation of studies are those by Davis, Dingel, and Miscio (2018); de Bellefon and others (2018); and Vogel and others (2018). This paper adds to this literature.

Our results show that there is value in relying on what one “sees,” especially where urbanization is messy. Subjective assessments can capture the multifaceted nature of cities—relatively large spaces with a higher density of construction, better access to transportation networks, and greater availability of residential amenities. In this paper, we develop what we believe is a credible structured approach to the subjective assessment of cities.

We also confirm the importance of combining indicators from the ground and from outer space when assessing the urban status of a location. Putting different sources of data together and comparing their performance in identifying urban areas reveals the weakness of some of the key criteria used traditionally. Population density, one of the workhorses of administrative approaches, performs quite poorly. And there are also pitfalls in relying exclusively on nighttime light, a method that is popular nowadays.
Finally, the value added from combining multiple sources of information highlights the importance of reliable administrative boundaries. Digitized administrative boundaries—the polygons delimiting the jurisdiction of cities, towns, and villages—are the true anchor of data integration. They can be used as the basis for combining information from different sources and for establishing ground truth. But for this approach to work, credible digitized files (often in the form of shapefiles) are needed for the lowest layers of the administrative hierarchy of a country. However, this issue is often overlooked by modern approaches based on remote sensing technology.

References

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. “The Economics of Density: Evidence from the Berlin Wall.” *Econometrica* 83 (6): 212789.

Alonso, W. 1964. “Location and Land Use: Toward a General Theory of Land Rent.” Harvard University, Cambridge, MA.

Bilger, Marcel, and Willard G. Manning. 2015. "Measuring Overfitting in Nonlinear Models: A New Method and an Application to Health Expenditures." *Health Economics* 24 (1): 75–85.

Bontemps, S., P. Defourny, E. V. Bogaert, O. Arino, V. Kalogirou, and J. R. Perez. 2011. "GLOBCOVER 2009 Products Description and Validation Report.” Google Scholar.

Breiman, Leo. 2001a. "Random Forests." *Machine Learning* 45 (1): 5–32.

___________. 2001b. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3): 199–231.

Brookings Institution. 2015. *Global Metro Monitor 2014*. Washington, DC: Brookings.

Brueckner, J. K. 1987. “The Structure of Urban Equilibria: A Unified Treatment of the Muth-Mills Model.” *Handbook of Regional and Urban Economics*, Vol. 2, edited by E. S. Mills, 821–45. Amsterdam: North-Holland.

Burchfield, M., H. G. Overman, D. Puga, and M. A. Turner. 2006. “Causes of Sprawl: A Portrait from Space.” *Quarterly Journal of Economics* 121 (2): 587–633.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, U.K.: Cambridge University Press.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. 2010. “Estimating Agglomeration Economies with History, Geology, and Worker Effects.” In *Agglomeration Economics*, edited by Edward L. Glaeser, 15–66. Chicago: University of Chicago Press.

Copas, John B. 1983. "Regression, Prediction and Shrinkage." *Journal of the Royal Statistical Society. Series B* (Methodological): 311–54.

Davis, Donald, Jonathan Dingel, and Antonio Miscio. 2018. “Cities, Skills, and Sectors in Developing Economies.” Paper presented at 8th European Meeting of Urban Economics Association, Düsseldorf, Germany, May 11.
de Bellefon, Marie-Pierre, Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. 2018. “Delineating Urban Areas with Building Density.” Paper presented at 8th European Meeting of Urban Economics Association, Düsseldorf, Germany, May 11.

Denis, Eric, and Kamala Marius-Gnanou. 2011. “Toward a Better Appraisal of Urbanization in India: A Fresh Look at the Landscape of Morphological Agglomerates.” *Cybergeo: European Journal of Geography*.

Donaldson, Dave, and Adam Storeygard. 2016. “The View from Above: Applications of Satellite Data in Economics.” *Journal of Economic Perspectives* 30 (4): 171–98.

Duranton, G., and D. Puga. 2004. “Micro Foundations of Urban Agglomeration Economies.” In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. V. Henderson and J. F. Thisse, 2063–117. Amsterdam: North-Holland.

______. 2015. “Urban Land Use.” In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William Strange, 467–560. Amsterdam: North-Holland.

Elvidge, Christopher D., Benjamin T. Tuttle, Paul C. Sutton, Kimberly E. Baugh, Ara T. Howard, Cristina Milesi, Budhendra Bhaduri, and Ramakrishna Nemani. 2007. "Global Distribution and Density of Constructed Impervious Surfaces." *Sensors* 7 (9): 1962–79.

Elvidge, Christopher D., Daniel Ziskin, Kimberly E. Baugh, Benjamin T. Tuttle, Tilottama Ghosh, Dee W. Pack, Edward H. Erwin, and Mikhail Zhizhin. 2009. "A Fifteen Year Record of Global Natural Gas Flaring Derived From Satellite Data." *Energies* 2 (3): 595–622.

Esch, Thomas, Wieke Heldens, Andreas Hirner, Manfred Keil, Mattia Marconcini, Achim Roth, Julian Zeidler, Stefan Dech, and Emanuele Strano. 2017. "Breaking New Ground in Mapping Human Settlements from Space—The Global Urban Footprint." *ISPRS Journal of Photogrammetry and Remote Sensing* 134: 30–42.

Frey, Bruno S., and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40 (2): 402–35.

Friedl, Mark A., Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. 2010. "MODIS Collection 5 Global Land Cover: Algorithm Refinements and Characterization of New Datasets." *Remote Sensing of Environment* 114 (1): 168–82.

Fujita, M., P. R. Krugman, and A. J. Venables. 2001. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

GADM (Global Administrative Areas). 2012. "GADM Database of Global Administrative Areas, Version 2.0." http://www.gadm.org/.

Henderson, J. V. 1974. “The Sizes and Types of Cities.” *American Economic Review* 64 (4): 640–56.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Henderson, M., E. T. Yeh, P. Gong, C. Elvidge, and K. Baugh. 2003. “Validation of Urban Boundaries Derived from Global Night-Time Satellite Imagery.” *International Journal of Remote Sensing* 24 (3): 595–609.
Hsu, Feng-Chi, Kimberly E. Baugh, Tilottama Ghosh, Mikhail Zhizhin, and Christopher D. Elvidge. 2015. “DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration.” Remote Sensing 7: 1855–76.

Hwang, Jackelyn, and Robert J. Sampson. 2014. "Divergent Pathways of Gentrification: Racial Inequality and the Social Order of Renewal in Chicago Neighborhoods." American Sociological Review 79 (4): 726–51.

IDFC Institute. 2015. “Chasing Definitions in India.”
http://www.idfcinstitute.org/knowledge/publications/op-eds/chasing-definitions-in-india/.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning, Vol. 112. New York: Springer.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” Science 353 (6301): 790–94.

Knottnerus, Paul. 2003. Sample Survey Theory. Berlin: Springer Science+Business Media.

Krugman, Paul. 1991. “Increasing Returns and Economic Geography.” Journal of Political Economy 99 (3): 483–99.

Lewis, W. A. 1954. “Economic Development with Unlimited Supplies of Labour.” The Manchester School 22 (2): 139–91.

Li, Xuecao, and Yuyu Zhou. 2017. "Urban Mapping Using DMSP/OLS Stable Night-Time Light: A Review.” International Journal of Remote Sensing 38 (21): 6030–46.

Li, Yue, Martin Rama, Virgilio Galdo, and Maria Florencia Pinto. 2015. “A Spatial Database for South Asia.” World Bank, Washington, DC.

Liaw, Andy, and Matthew Wiener. 2002 "Classification and Regression by Random Forest." R News 2 (3): 18–22.

Mertes, Carly M., A. Schneider, D. Sulla-Menashe, A. J. Tatem, and B. Tan. 2015. "Detecting Change in Urban Areas at Continental Scales with MODIS Data." Remote Sensing of Environment 158: 331–47.

Michaels, G., F. Rauch, and S. J. Redding. 2012. “Urbanization and Structural Transformation.” Quarterly Journal of Economics 127 (2): 535–86.

Mills, E. S. 1967. “An Aggregative Model of Resource Allocation in a Metropolitan Area." American Economic Review 57 (2): 197–210.

Muth, R. F. 1969. Cities and Housing: The Spatial Patterns of Urban Residential Land Use. Chicago: University of Chicago Press.

NOAA (National Oceanic and Atmospheric Administration). 2014. Global Radiance Calibrated Night Lights, National Geophysical Data Center, https://ngdc.noaa.gov/eog/dmsp/download_radcal.html

NSSO (National Sample Survey Office). 2012. “The 68th Round of National Sample Survey of India.” National Sample Survey Office, Ministry of Statistics and Programme Implementation, Government of India.
Pesaresi, Martino, Daniele Ehrlich, Stefano Ferri, Aneta Florczyk, Sergio Freire, Matina Halkia, Andreea Julea, Thomas Kemper, Pierre Soille, and Vasileios Syrris. 2016. Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014. Luxembourg: Publications Office of the European Union.

Pesaresi Martino, Michele Melchiorri, Alice Siragusa, and Thomas Kemper. 2016. Atlas of the Human Planet 2016: Mapping Human Presence on Earth with the Global Human Settlement Layer. JRC Science for Policy Report, European Commission, Brussels. https://ghsl.jrc.ec.europa.eu/atlas2016Overview.php.

Picard, Richard R., and R. Dennis Cook. 1984. "Cross-validation of Regression Models." Journal of the American Statistical Association 79 (387): 575–83.

Potere, David, Annemarie Schneider, Shlomo Angel, and Daniel L. Civco. 2009. "Mapping Urban Areas on a Global Scale: Which of the Eight Maps Now Available Is More Accurate?" International Journal of Remote Sensing 30 (24): 6531–58.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse. 2011. “The Area and Population of Cities: New Insights from a Different Perspective on Cities." American Economic Review 101 (S): 2205–25.

Salesses, Philip, Katja Schechtner, and César A. Hidalgo. 2013. "The Collaborative Image of the City: Mapping the Inequality of Urban Perception." PLOS One 8 (7): e68400.

Schneider, Annemarie. 2012. "Monitoring Land Cover Change in Urban and Peri-urban Areas Using Dense Time Stacks of Landsat Satellite Data and a Data Mining Approach." Remote Sensing of Environment 124: 689–704.

Schneider, Annemarie, Mark A. Friedl, and David Potere. 2009. "A New Map of Global Urban Extent from MODIS Satellite Data." Environmental Research Letters 4 (4): 044003.

Scotchmer, S. 2002. “Local Public Goods and Clubs." Handbooks in Economics 4 (4): 1997–2044.

Shao, Jun. 1993. "Linear Model Selection by Cross-Validation." Journal of the American statistical Association 88 (422): 486–94.

Shmueli, Galit. 2010. "To Explain or to Predict?" Statistical Science 25 (3): 289–310.

Sober, Elliott. 2002. "Instrumentalism, Parsimony, and the Akaike Framework." Philosophy of Science 69 (S3): S112–S123.

Straszheim, M. 1987. “The Theory of Urban Residential Location.” In Handbook of Regional and Urban Economics 2: 717–57. Amsterdam: Elsevier.

Uchida, Hirotugu, and Andrew Nelson. 2009. "Agglomeration Index: Towards a New Measure of Urban Concentration." Background paper, World Development Report 2009: Reshaping Economic Geography, World Bank, Washington, DC.
UN (United Nations). 2005. “Designing Household Survey Samples: Practical Guidelines.” Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

______. 2015. “World Urbanization Prospects: The 2014 Revision.” Department of Economic and Social Affairs, Population Division, United Nations, New York.

Veenhoven, Ruut. 2004. “Happiness as a Public Policy Aim: The Greatest Happiness Principle.” In Positive Psychology in Practice, edited by P. A. Linley and S. Joseph. Hoboken, N.J.: John Wiley.

Vogel, Kathryn Baragwanath, Ran Goldblatt, Gordon H. Hanson, and Amit K. Khandelwal. 2018. “Detecting Urban Markets with Satellite Imagery: An Application to India.” NBER Working Paper No. w24796, National Bureau of Economic Research, Cambridge, MA.

Wooldridge, Jeffrey. 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

World Bank. 2013. Urbanization beyond Municipal Boundaries: Nurturing Metropolitan Economies and Connecting Peri-urban Areas in India. Washington, DC: World Bank.

______. 2015. Leveraging Urbanization in South Asia: Managing Spatial Transformation for Prosperity and Livability. Washington, DC: World Bank.

______. 2016. World Development Indicators (database). Washington, DC. http://data.worldbank.org/data-catalog/world-developmentindicators.

Yu, Le, and Peng Gong. 2012. “Google Earth as a Virtual Globe Tool for Earth Science Applications at the Global Scale: Progress and Perspectives.” International Journal of Remote Sensing 33 (12): 3966-86.

Zhou, Y., S. J. Smith, C. D. Elvidge, K. Zhao, A. Thomson, and M. Imhoff. 2014. “A Cluster-Based Method to Map Urban Area from DMSP/OLS Nightlights.” Remote Sensing of Environment 147: 173–85.