Policy Research Working Paper 9160

Identifying Urban Areas by Combining Human Judgment and Machine Learning

An Application to India

Virgilio Galdo
Yue Li
Martin Rama

WORLD BANK GROUP
South Asia Region
&
Latin America and the Caribbean Region
February 2020

Policy Research Working Paper 9160

Abstract

This paper proposes a methodology for identifying urban areas that combines subjective assessments with machine learning, and applies it to India, a country where several studies see the official urbanization rate as an under-estimate. For a representative sample of cities, towns and villages, as administratively defined, human judgment of Google images is used to determine whether they are urban or rural in practice. Judgments are collected across four groups of assessors, differing in their familiarity with India and with urban issues, following two different protocols. The judgment-based classification is then combined with data from the population census and from satellite imagery to predict the urban status of the sample. The Logit model, and LASSO and random forests methods, are applied. These approaches are then used to decide whether each of the out-of-sample administrative units in India is urban or rural in practice. The analysis does not find that India is substantially more urban than officially claimed. However, there are important differences at more disaggregated levels, with "other towns" and "census towns" being more rural, and some southern states more urban, than is officially claimed. The consistency of human judgment across assessors and protocols, the easy availability of crowd-sourcing, and the stability of predictions across approaches, suggest that the proposed methodology is a promising avenue for studying urban issues.

This paper is a product of the Office of the Chief Economist, South Asia Region, and the Office of the Chief Economist, Latin America and the Caribbean 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/prwp. The authors may be contacted at yli7@worldbank.org.

Identifying Urban Areas by Combining Human Judgment and Machine Learning: An Application to India

Virgilio Galdo, Yue Li and Martin Rama[1]

Key words: urban area; urbanization rate; human judgment; Google images; crowd sourcing; population census; satellite imagery; machine learning

JEL Classification: O1, O18, R1

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

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.

Skillful research analysis was provided by Maria Florencia Pinto, Sutirtha Sinha Roy, and Wenqing Zhang. The authors thank Gilles Duranton, Hans Timmer, two anonymous reviewers and participants at the 8th European Meeting of the Urban Economics Association for very helpful comments and suggestions. The authors appreciate operational support provided by Ryan Engstrom and Richard Hinton from the Center for Urban and Environmental Research at George Washington University, and Charles Fox and Keith Garrett from the Geospatial Operational Support Team at the World Bank.

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 fringe, among others—testifies to the absence of a clear divide. Even allegedly urban localities can differ substantially in their key attributes, as they range from compact cities to sprawling low-density areas. In this context, just like beauty, a city is in the eyes of the beholder.

The rural-urban continuum and the heterogeneity of urban settings pose an obvious challenge to identifying urban areas and measuring urbanization rates in a consistent way within and across countries. An objective methodology for distinguishing between urban and rural areas that is based on one or two metrics with fixed thresholds may not adequately capture the wide diversity of places. A richer combination of criteria would better describe the multifaceted nature of a city's function and its environment, but the joint interpretation of these criteria may require an element of human judgment.

In this paper we turn this unavoidable subjectivity into an opportunity, by proposing a methodology to identify urban areas that combines human judgment with machine learning. Human judgment is used to classify a representative sample of places into urban or rural. Reliance on several, qualitatively different groups of assessors, as well as on different classification protocols, provides reassurance that the outcome is stable. This sample is in turn used as the training set for a machine learning exercise allowing to classify out-of-sample places as urban or rural. A comparison between the various classification approaches provides further reassurance that that the prediction outcome is robust.

We illustrate the potential of this methodology by applying it to India. Two reasons motivate this choice. First is scale: accounting for almost a fifth of the world population, India encompasses states with incomes per capita comparable to Mexico and others similar to Benin. Second is the "messiness" of its urbanization process, characterized by a wide variation in the strength of local urban authorities, from high in state capitals to dismal in other cities. Scale and messiness result in an enormous diversity of places, ranging from a glamorous metropolis such as Mumbai, capital of Maharashtra, to a shabby town such as Bagula in West Bengal, to any among hundreds of thousands of villages.

With such wide spatial diversity, not surprisingly there is considerable debate as to how urbanized India actually is. The official urbanization rate for 2011 is 31.2 percent. Several studies using alternative population thresholds or satellite imageries of built-up cover have argued that many areas of India are misclassified as rural by the administrative definition. Depending on the methodology, the urbanization rates estimated by these studies range from 42.0 to 78.0 percent (Denis and Marius-Gnanou 2011; Dijkstra et al. 2018; Ministry of Finance 2017).

Conversely, a study in this special issue relies on a range of parameters to delineate urban markets and suggests that India could be even less urbanized than official figures imply. In this other study the urbanization rate ranges from 14.8 to 33.5 percent when using nighttime light data, and from 27.1 to 39.4 percent when using built-up cover data (Baragwanath et al. 2019). Shedding light on this debate is the second motivation for choosing India to illustrate our proposed methodology.

Subjective assessments are increasingly being used to complement more objective approaches across various fields in economics. For example, good-looking ratings by independent parties have been shown to be correlated with occupational sorting, earnings differentials and physical performance (Beller et al.

1994; Hamermesh and Biddle 1994; Postma 2014). Similarly, self-reported happiness has been adopted as a wellbeing indicator 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).

In the urban economics literature, crowd-sourced assessments of street-level images have been used to determine how safe a neighborhood feels, how clean it looks, or how lively it seems. These are aspects of cities that standard measures, such as income levels, are unable to fully capture (Salesses et al. 2013). Further, this literature has applied machine learning to extend the assessment to other cities, not covered by the original crowd-sourcing (Naik et al. 2016). Machine learning has also been used to predict a neighborhood's socioeconomic characteristics out of its appearance (Glaeser et al. 2018; Naik et al. 2015).

Human judgment is also behind imagery interpretation in the remote sensing literature. Machine learning is often used in this literature to classify billions of "cells" of satellite imageries, for example in terms of their land use. However, satellite imageries are difficult to interpret, because they are two-dimensional, taken directly from above and generally lacking recognizable details. Therefore, an important step is to generate subjective assessments for a subset of cells that is then used as either a training sample or validation data for the machine learning exercise (Campbell and Wynne 2011).

Building on these precedents, our methodology applies subjective assessments to open-source images generated by Google for a representative sample of places in India. We overlay these images with the digitized boundaries of Indian cities, towns and villages, as administratively defined, and use the portion falling within the corresponding boundaries as the subject for human judgment.

Relying on information at the level of cities, towns and villages to shed light on important topics in economic geography has precedents in the literature (Eeckhout 2004; Levy 2009; Michaels, Rauch and Redding 2012). Fairly disaggregated administrative units have also been shown to perform as well as gridded cells when dealing with critical topics in the economic geography literature (Briant, Combes and Lafourcade 2010). An admittedly preferable approach would be to let commuting pattern data delineate local labor markets (Duranton 2015; US Office of Management and Budget 2010). However, data of this sort is rarely available in developing countries, and India is not an exception in this respect.

In our methodology, four groups of assessors independently judge whether the Google images from a place correspond to an urban or a rural settlement. The four groups differ in their familiarity with India and in their expertise in urban issues. The most knowledgeable group is made of in-house research analysts. A second group comprises university students from the US who have experience in land use classification but no exposure to India. The last two groups are made of crowdsourced anonymous viewers hired through Amazon Mechanical Turk (MTurk) who are unlikely to have expertise in urban issues. In the third group all viewers are from India while in the fourth one they are all from the US.

Assessments also follow two different protocols. The more structured one requests the assessors to sequentially evaluate the density of construction, the nature of transport infrastructure, and the availability of urban amenities, before making their judgment. All four groups of assessors follow this protocol. But before doing so, the three outsider groups are also asked to have an impromptu judgment of the urban or rural nature of the place, without any guidance.

By proceeding this way, every place in our sample of cities, towns and villages is classified a total of seven classification rounds. We find that the classifications are highly consistent, with two thirds of the localities

having the same status regardless of the group of assessors and the protocol, and with almost 90 percent of them being classified in the same way in at least five of the seven rounds. We also show that the characteristics of the assessors and the protocol followed affect judgment results, but their overall influence is very small. Given this high level of consistency, we pool all seven rounds of assessments which results in 50 or more human judgments for each place in the sample, and then classify every place as urban or rural based on a majority rule. We treat this pooled classification as the outcome of our human judgment exercise.

We then to develop an approach to predict urban status, as assessed through human judgment, based on observable characteristics of the corresponding places. We emphasize four characteristics for each place. Total population and population density, both drawn from the population census, are key indicators in the urban economics literature. Built-up cover and nighttime lights, both from satellite imageries, have been at the core of recent attempts to estimate a country's "true" urbanization rate. We also consider other indicators, but the four key characteristics just listed remain the most important predictors of subjective urban status.

Several approaches, drawn from classical data modeling and from machine learning, are applied for the prediction exercise as well. We first use a standard Logit model linking the urban status of a place with the values of its characteristics. We then apply more sophisticated machine learning methods, including LASSO and random forests. We find that all three approaches yield a similar prediction accuracy, with random forests performing slightly better. These approaches are then used to predict whether each of the other places in the country is urban or rural in practice.

The urbanization rate emerging from this exercise for India's is 29.9 percent. Unlike the remarkably high rates reported by some studies, this figure is quite close to the corresponding official estimate. However, we also find important gaps with the official estimates at more disaggregated levels. Consistent with those other, recent studies, we show that many places administratively defined as villages have urban characteristics. But we also find that a significant share of "statutory towns" and "census towns" would be better classified as rural. And there are differences across states as well.

While these subnational gaps with official estimates could be interpreted as the outcome of prediction error, we show that the sign of the gaps is consistent with the fiscal incentives and statistical innovations characterizing the administrative classification of places in India. We also show that our estimate of the urbanization rate is not an outlier relative to estimates based on global urban layer products, as the main discrepancy with them arises from the population data used, rather from the underlying land use classification.

Beyond offering new insights into the debate on urbanization in India, the exercise in this paper illustrates several strengths of the proposed methodology. First, the methodology is holistic. Because it relies on human judgment to assess the "urbanness" of places, it allows the multiple characteristics of a place to be evaluated jointly. Second, it makes the most of existing data. A growing number of studies are using satellite imagery data to identify urban areas, contributing to urban studies (Donaldson and Storeygard 2016). However, our analysis shows that relying on built-up cover or nighttime lights alone can be insufficient. And third, our methodology is scalable. High-quality Google images are becoming available for an increasing number of places, and human judgment can be crowd-sourced efficiently nowadays.

A legitimate concern with our methodology is its reliability, in the sense that small changes to its design could result in important changes in its outcomes. However, the variants tried in this paper suggest that

the results are robust. Our analysis shows that assessor characteristics and decision protocols do not significantly affect the classification of places. And the prediction approaches used to extend the classification beyond the original sample do not make a major difference either. Therefore, the proposed methodology may contribute to a growing literature in urban economics that applies innovative approaches to the identification of urban areas (de Bellefon et al. 2018, Diegel et al. 2019, Rozenfeld et al. 2011).

The paper is organized as follows. In Section 2 we present the stratified sampling strategy. In Section 3 we describe the images used for assessment and the collection of human judgments from different assessors. We also compare classifications based on different human judgment exercises. In Section 4 we introduce the additional sources of data and apply three approaches—Logit model, and LASSO and random forests methods—to predict the urban status of places in the sample. In Section 5 we show the results on the estimated urbanization rate for India in 2011, and also for fine administrative types and for states. Finally, we compare our results with urbanization rates from other studies.

2. A sample of places

Important information about cities, towns and villages is captured by the administrative boundaries of localities. These boundaries are generally built on historical data and tend to reflect the best knowledge available on the spatial distribution of economic activities and people. They also define the constituency to which each local government provides services and is accountable. In this study, we used the administrative units adopted by the Census of India as our unit of analysis.

The universe of places

To compile the universe of places we took advantage of newly digitized boundaries of administrative units in India, available down to the town or village level, to define the place of analysis. 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 et al. 2015). The 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 that present the boundaries of towns and villages. The location, size, and shape of these towns and villages were digitized in the form of 649,818 boundary polygons. Attributions such as place codes, names, and administrative types were added to the polygons.

Administrative units need to be integrated further in the case of metropolitan areas, which represent broader integrated labor markets. The preferred approach in this respect is to base the integration process on commuter pattern data (Duranton 2015; US Office of Management and Budget 2010). This type of data is unfortunately not available for most developing countries, including India. In its absence, we simply merged the polygons of cities spreading over multiple administrative units, based on their names and unique geo-codes.

We made two other adjustments to the digitized administrative boundaries. In the original Spatial Database some villages were presented as points on the source maps as their boundaries were unavailable —they are mostly in the mountainous or forest areas of the states of Arunachal Pradesh, Chhattisgarh, Himachal Pradesh, and Meghalaya. And 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. We therefore excluded villages without boundaries or population data.

As a result of these mergers and exclusions, the number of administrative units we retained as the universe of places for the analysis fell to 564,052.

Two concerns can be raised on the use of such administrative units as the unit of analysis. The first one, known as the modifiable areal unit problem (MAUP), is the potential sensitivity of results to their considerable diversity in shape and size. It has been shown that shape does not make much of a difference for three important topics in the economic geography literature (namely, spatial concentration, agglomeration economies, and inter-regional trade). As for size, administrative boundaries do not suffer significantly from the MAUP, except in the case of large-scale spatial units, such as states or provinces (Briant, Combes and Lafourcade 2010). However, none of the administrative units considered here reaches such large scale.

A second concern is whether administrative units, self-standing or merged in the case of bigger cities, provide an accurate representation of integrated labor markets. An alternative is to delineate the urban extent by aggregating contiguous cells. Depending on the studies, the aggregation of cells is based on criteria such as the density of their built-up or the intensity of their nighttime light (Baragwanath et al. 2019; Ch et al. 2018; Rozenfeld et al. 2011). However, the nature of the links between cells in a broader spatial context may vary across localities. The aggregation process can introduce other biases (Bosker et al. 2018).

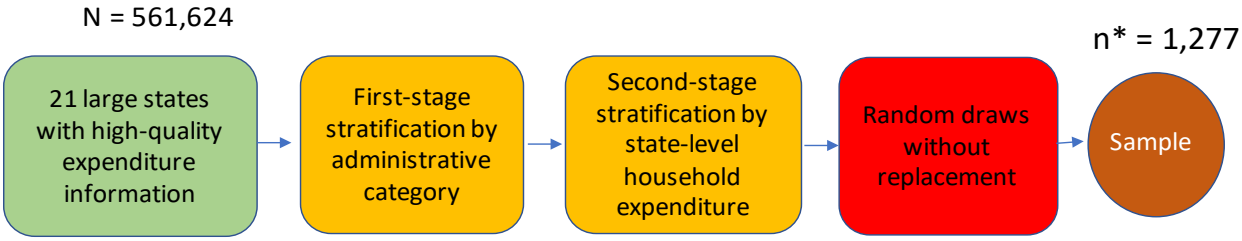
For selecting a representative sample, we further restricted the set of administrative units to those of the 21 largest states and union territories (hereafter states). These large states are covered by the high-quality household surveys whose data are needed for a proper stratification. Focusing exclusively on them resulted in 561,624 places being considered for sampling.

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 of India for the country as a whole (ORGI 2011a). Almost a third of this population gap—about 11.3 million people—is accounted for by the small states excluded from the analysis. Some of these small states 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 in the official sources. These villages are most likely rural in practice, and their population 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 sampling

We used a two-stage disproportionate stratification strategy to select the sample of administrative units for human judgment (figure 1). In the first stage, we divided the universe into seven strata by administrative categories, following the definitions of the 2011 Census of India. Four of these categories —state capitals and municipal corporations, municipalities, industrial towns, and other towns—are statutorily designated as urban. Two of the categories—census towns and outgrowths—are statutorily rural but recognized by the 2011 Census as urban. The seventh category is villages, or areas that are rural according to both the Constitution and the 2011 Census of India (ORGI 2011a).

Figure 1. Two-stage stratification sampling



In the first stage we relied on a disproportionate sampling strategy. The 561,624 administrative units comprise 3,852 statutory urban places, 4,645 places that are urban in terms of the census, and 553,127 villages. If a proportionate sampling strategy had been adopted, the share of statutory and census urban places in the sample would have been 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 places. It therefore made sense to overrepresent these two groups.

For each administrative category stratum, we chose 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. Building on the survey design literature, we rely on the Central Limit Theorem to determine the sample size of each stratum (appendix 1).

In the second stage, we divided each of the seven strata into three substrata defined by average consumption per capita at the state level. Consumption data were from the household consumer expenditure module of the 68th round of National Sample Survey of India (NSSO 2012). Monthly household expenditure was divided by household size to estimate consumption per capita. To account for the spatial variation in prices, the result was deflated by a year-specific price index that differs across states and between rural and urban areas. We ranked states on the basis of their average consumption per capita and divided them into three groups of equal size: richer, intermediate, and poorer. The sample for each administrative stratum was split evenly across these three groups of states. The sampling process is completed by conducting random draws without replacement for each of the 21 substrata.

The resulting sample of places

The sample drawn through this process comprises 1,277 places, of which 405 are statutorily urban and 557 are urban in terms of the census; the remaining 315 places are villages (table 1). For the full sample we achieved a relative standard error of 8 percent, which is below the 10 percent threshold required for a survey design to be considered good (Knottnerus 2003; UN 2005).

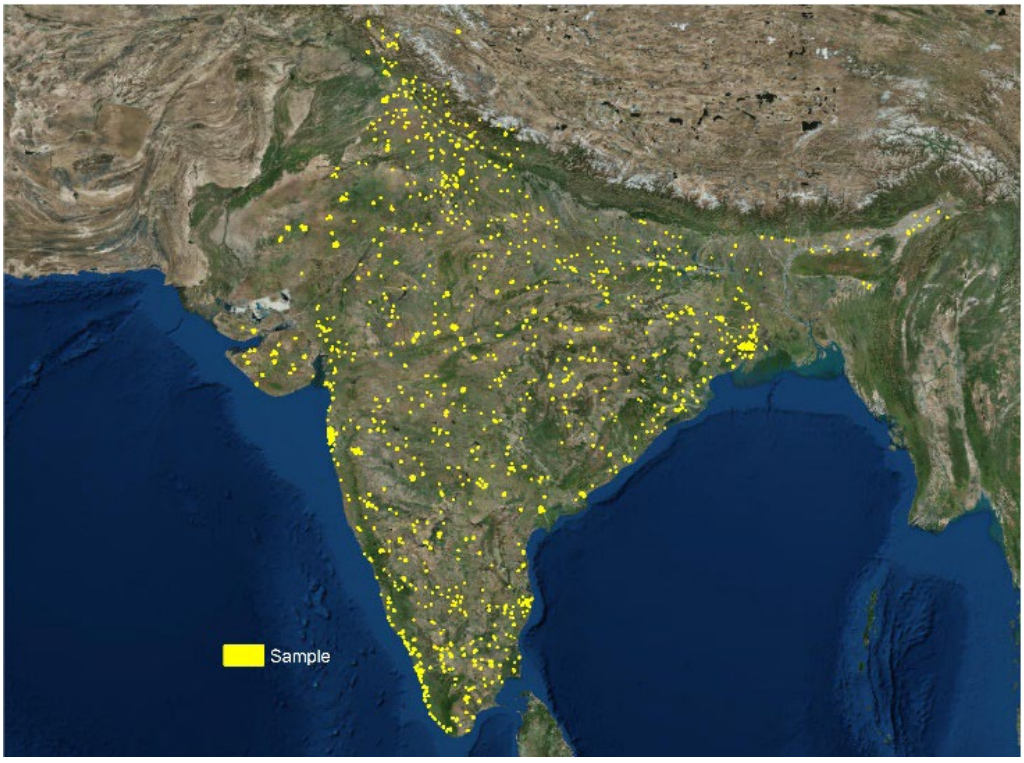
Recognizing that the Central Limit Theorem may not apply for a finite sample, we used bootstrapping to generate an approximation of the distribution through a Monte Carlo simulation (Cameron and Trivedi 2005). The analysis confirmed that the sample was sufficiently large for each administrative stratum. Standard errors were below 3 percent for six administrative strata, and around 5 percent for industrial towns.

Table 1. The distribution of the sample across administrative categories

	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			1,277
Mean probability of being urban			0.11	
Relative standard error			0.08	

Note: N is the total number of places belonging to an administrative category and n* is the selected sample size for an administrative category.

Figure 2. The spatial distribution of sample places



Our final sample with 1,277 administrative units is quite evenly distributed across India's territory (Figure 2). The selected places cover five of the six regions of the country— northern, central, eastern, western,

and southern—relatively well. Only for the northeastern region the coverage of the sample is limited. This is partially due to the exclusion from the universe of Manipur, Mizoram, Tripura, Meghalaya, and Assam, five small states with relatively poor household survey data. It is also due to the shortcomings of village boundaries in the states of Arunachal Pradesh and Meghalaya. Not surprisingly, the places in the sample are quite different in size, with their surface varying from 0.1 to 464.7 km^2.

3. Human judgment

Places are usually classified as urban or rural based on a narrow set of indicators that most often includes population and administrative status. Sometimes additional metrics such as population density or the share of employment out of agriculture are considered as well. But even this may not be enough to ensure that the classification is consistent. Cities have diverse characteristics both across and within countries. This multiplicity of aspects to consider calls for an element of human judgment.

Google images as virtual site visits

A thorough subjective assessment would require visiting every city, town, and village in the sample and literally seeing how they look like. However, the time and cost implications of visiting 1,277 places scattered across a large country are dissuasive. To resolve a similar challenge, an emerging literature uses Google images as a proxy for information gathered through field surveys. Salesses et al. (2013) rely on over 4,000 Google street view images to crowdsource perceptions on safety and livability, for two major cities of the US. In a scaling up effort, Dubey et al. (2016) expand this approach to over 110,000 images from 56 cities. Naik et al. (2016) use the images associated with quantified human judgment as the training set and develop a model to further quantify neighborhood appearance for 19 cities in the US through machine learning. And Naik et al. (2015) apply the machine learning model to study the changes in neighborhood appearance and their relationship with neighborhood socioeconomic characteristics.

Following the literature, we relied on open-source images accessible through Google Earth and Google Maps as a proxy for real site visits. The images provided by Google are an integrated collection of processed satellite imagery, aerial photos and street view pictures. A zoom-in option allows exploring details with a resolution ranging from 15 cm to 15 m. With these images a layperson, with no special training in remote sensing, can judge the characteristics of large numbers of places.[2]

The views convey information on the type of land cover, the characteristics of man-made structures, the layout of these structures, and transportation vehicles on roads. While three-dimensional images are not available for India, the combination of rotation and tilt options allows examining each place from different angles, giving a sense of the height and quality of buildings. A street view option providing 360-degree panoramic ground-level photos is also available for selected places in India. And tags highlight amenities such as educational institutions, health facilities, religious buildings and recreational parks, among others.

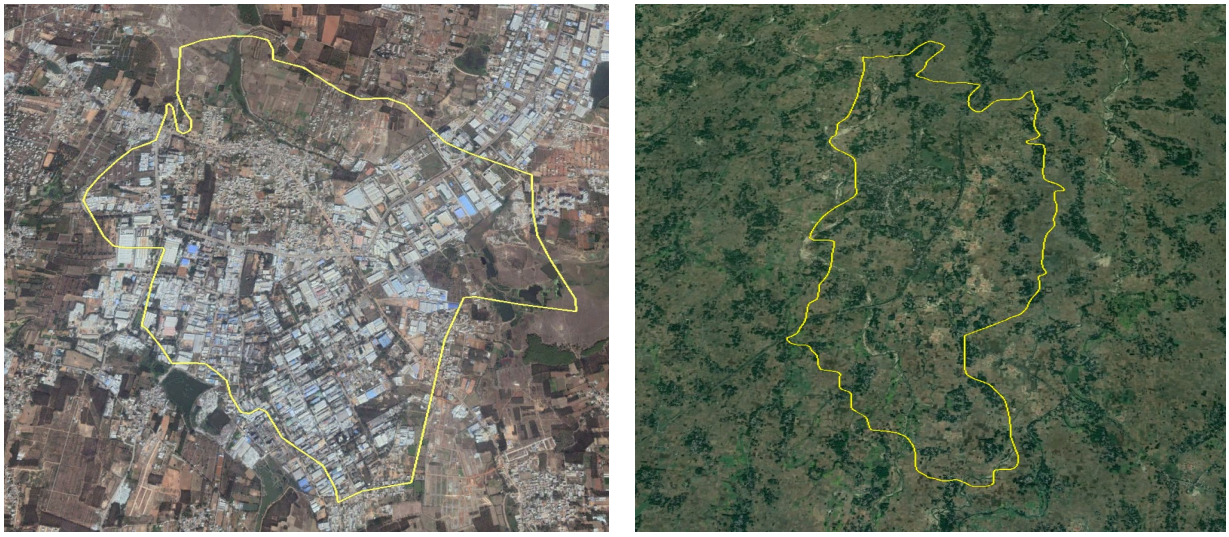
To support the virtual assessment of the places in the sample, we overlaid the digitized boundary of each of these places on top of Google images, and used the images falling within the corresponding boundary as the subject for human judgment (figure 3). For the vast majority of places, we relied on images taken

[2] Information on these images is available at https://www.google.com/streetview/explore/, https://support.google.com/earth/answer and https://support.google.com/mapsdata/.

in 2010–12, to be consistent with the population data from the 2011 Census of India. However, when the images for these three years were of poor quality, we complemented them with images from more recent years. Similarly, when Google Earth images were not available, we used Google Maps images that were composites from recent years.

Figure 3. Examples of Google images overlaid with administrative boundaries

a. Jigani (Karnataka) **b. Bahadurganj (Bihar)**



Note: The yellow lines depict the digitized administrative boundaries for the respective administrative units.

Diverse groups of assessors

To collect subjective assessments of Google images of the 1,277 places in the sample we mobilized a broad pool of assessors, emphasizing heterogeneity along two dimensions: familiarity with India and knowledge of urban issues. Heterogeneity is important to explore whether the assessors' characteristics matter. The two dimensions considered lead to the constitution of four distinct groups of assessors.

Our first group of assessors were three in-house research analysts who were both familiar with India and knowledgeable on urban issues. These analysts all had graduate-level degrees in economics and were given background information on the purpose of the exercise. While they were originally from three different developing countries – China, India and Peru – they were all familiar with India as they worked at the time at the South Asia region at the World Bank.

The second group of assessors comprised 15 university students who were knowledgeable on urban issues but unfamiliar with India. They all had experience in land use classification using satellite images, because of their activities at the Center for Urban and Environmental Research at George Washington University (GWU). These students participated on a voluntary basis, for a fixed payment upon completing the classification of all the places assigned to them for judgment. The students were all from the US.

The third and fourth group of assessors were anonymous online workers mobilized through Amazon's Mechanical Turk (MTurk). This is a crowdsourcing marketplace that facilitates the virtual breakdown and distribution of manual time-consuming tasks among thousands of anonymous workers and has become increasingly popular as a tool for research.[3] While MTurk workers often have a college degree or above, they are unlikely to be urban experts (Ross et al. 2010).

To ensure quality, we required participants in the third and fourth groups of assessors to have MTurk "master" accreditation, which reflects substantial experience and low rates of rejection. MTurk allows to restrict the nationality of participants in crowdsourced tasks. Our third group of assessors was thus made exclusively of MTurk workers from India, and the fourth one exclusively from the US. In all, the third group had 72 members and the fourth one 207.

Two assessment protocols

All four groups of assessors followed a tightly structured protocol for the classification of places. However, before being exposed to this structured approach, assessors in the second, third and fourth groups were also asked to provide an impromptu judgment. Without any guidance they had to decide by themselves whether the Google images shared with them corresponded to urban or rural places. As a result, each of the places in the sample was subject to seven rounds of classification: three of them impromptu and four following a structured protocol.

Under the structured protocol, assessors were requested to rely on a decision tree to classify the places allocated to them. The decision tree had three nodes, each involving the interpretation of Google images along one dimension. The first node referred to the distribution of land cover types. The second node was about the characteristics of the buildings and the relationship between them. And the third one concerned the presence of transportation networks and the availability of amenities such as schools, universities, hospitals or cultural sites (appendix 2).

The decision tree was built on the economic geography and urban economics literatures. Indeed, models of land use predict that the density of construction declines as one moves away from a city center to more rural areas (Alonso 1964; Mills 1967; Muth 1969; Brueckner 1987; Duranton and Puga 2015). It has also been shown that large infrastructure investments facilitating access to markets spur the concentration of economic activities (Duranton and Puga 2004; Fujita et al. 2001; Krugman 1991; Scotchmer 2002). Finally, services and amenities are crucial in explaining where firms prefer to operate, and where households prefer to live (Ahlfeldt et al. 2015; Combes et al. 2010; Straszheim 1987).

While the structured protocol was the same across all four groups, the diversity of the assessors required slight adjustments to its implementation in each case. With the first group, the sample of cities, towns, and villages was randomly split into three subsets, each allocated to one of the three analysts. A randomly selected subset of 10 places was reallocated from the original research analyst to a different one to assess the robustness of the classifications. When one research analyst could not reach an unambiguous conclusion on a specific place, another analyst was called in to consult and make a joint decision.

With the second group we used an open-source data collection tool, Collect Earth, to be able to work remotely. Collect Earth enables data gathering and image analysis through Google Earth.[4] For each

[3] Information on Amazon's Mechanical Turk is available through https://www.mturk.com/.

[4] Information on Collect Earth is available through http://www.openforis.org/tools/collect-earth.html.

university student in the second group we created two Collect Earth "projects" including 640 randomly chosen places each, and we shared them sequentially. For places in the first project, the assessors were asked to decide without any guidance whether they were urban or rural. For the second project, on the other hand, they had to follow the structured protocol.

The procedure was similar for anonymous MTurk workers in the third and fourth groups of assessors. We created two MTurk "tasks" including links to the Google images of all 1,277 places with overlaid administrative boundaries. The first task requested impromptu assessments while for the second one the assessors had to follow the structured protocol. Because MTurk workers are anonymous, we could not ensure that the assessors participating in the first and second round were the same. But for each round, we could specify that each of the 1,277 places needed to be assessed exactly 10 times and that the same MTurk workers could not classify any place more than once.

Assessing consistency

Relying on human judgment to classify places raises understandable concerns about the consistency of the outcomes depending on the characteristics of the assessors and the protocols they follow. Relying on administrative boundaries as the unit of analysis also implies that places can have very different sizes and this diversity could affect human judgment. We addressed these two concerns sequentially and found that the classification of places generated by our methodology was reassuringly stable.

First, we conducted a regression analysis linking the probability of a place being judged urban by an assessor with the characteristics of the accessor and with the protocol followed. We used a standard Logit model, with one observation per place and assessor, and with errors clustered at the place level. The two assessor characteristics considered were familiarity with India and knowledge of urban issues.

The results show that more places are considered urban by assessors familiar with India, while fewer places are judged to be urban among assessors who are knowledgeable on urban issues (table 2). It also appears that following a structured protocol leads to more places being classified as urban. However, while all coefficients are significant, the overall explanatory power of the regression is very small, with the pseudo R-square never exceeding 0.01. This means that subjective biases exist but do not have a substantial impact on the classification of places.

Another way to check whether human judgment leads to consistent results is to compare the classification of places across the seven rounds of assessments. To determine whether a place is urban or rural, we applied the majority rule to all the available assessments within each round. The few places for which we face a tie were excluded from the analysis. Overall, assessments are quite consistent. Two thirds of places have the same status in all seven rounds, an additional 13 percent are classified identically in six rounds, and yet another 10 percent in five rounds.

The classifications also appear to be highly consistent when considering pairwise comparisons between the seven rounds. The percentage of places assessed in the same way in two different rounds of classification varies between 82.3 and 96.7 (table 3). This is regardless of whether the few tied cases are included or not. The result is similar if Tetrachoric correlation coefficients are considered instead, with the estimated coefficients hovering around 0.9.

Table 2. The relationship between classifications, assessor characteristics and protocols

	(1)		(2)		(3)		(4)	
	coef.	dy/dx	coef.	dy/dx	coef.	dy/dx	coef.	dy/dx
Familiar with India	0.371 ***	0.091 ***					0.131 ***	0.032 ***
	(0.015)	(0.004)					(0.017)	(0.004)
Knowledgeable on urban issues			-0.296 ***	-0.073 ***			-0.193 ***	-0.047 ***
			(0.018)	(0.004)			(0.018)	(0.004)
Structured protocol					0.397 ***	0.097 ***	0.297 ***	0.072 ***
					(0.014)	(0.003)	(0.017)	(0.004)
Observations	68,978		68,978		68,978		68,978	
Log pseudolikelihood	-47030		-47129		-46904		-46857	
Pseudo R2	0.004		0.0019		0.0067		0.0077	

Note: Results are from a Logit model with the dependent variable being whether a place is judged to be rural (0) or urban (1) by an assessor, with one observation per assessor and place, and with standard errors clustered at the place. Coef. stands for estimated coefficient, dy/dx is the marginal effect, the numbers in the parentheses are standard errors clustered at the place level, and statistical significance is reported by *** if $p < 0.01$, ** if $p < 0.05$, and * if $p < 0.1$.

Table 3. Percentage of agreement between rounds of assessments

	World Bank analysts	GWU students		MTurk- India		MTurk- USA	
		Impromptu judgment	Structured protocol	Impromptu judgment	Structured protocol	Impromptu judgment	Structured protocol
World Bank analysts							
Structured protocol	100	82.3	83.7	85.6	88.2	87.6	89.1
GWU students							
Impromptu judgment		100	96.7	91.9	86.0	89.3	87.0
Structured protocol			100	92.0	86.0	89.3	86.6
MTurk- India							
Impromptu judgment				100	89.5	93.5	91.3
Structured protocol					100	92.6	93.0
MTurk- USA							
Impromptu judgment						100	93.6
Structured protocol							100

Given this high level of consistency across assessors and protocols, we pooled all seven rounds of assessments which resulted in 50 or more human judgments for each place in the sample. We then classified every place as urban or rural based on a majority rule over these 50 or more judgments and got only one tie among the 1,277 places in the sample.

Second, we assessed whether the size of administrative units affected human judgment. To do this, we classified the 1,277 places in the sample into 20 size quantiles and computed the share of places in each quantile that was considered urban based on the pooled classification. For the top three size quantiles the share of urban places is high and increases with their size. But for the other 17 quantiles there is no clear correlation between size and urban share (appendix 3).

We also investigated the relationship between the size of a place and the share of its surface that is built-up, which provides the first node in the decision tree followed by the assessors under the structured protocol. We found that the median built-up share varies across quantiles, but it displays no consistent relationship with sizes. This finding gave us further assurance that human judgment is not affected by the size of administrative units.

The classification of the sample

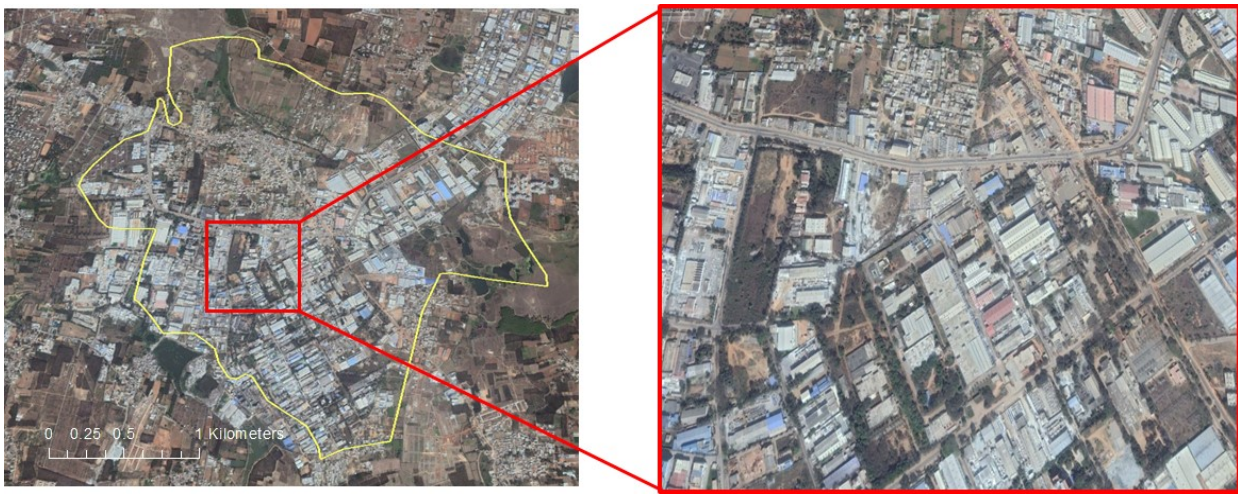
Given the consistency of outcomes across assessors, protocols and sizes, we treated the classification of places resulting from the pooled assessments as the outcome of our human judgment exercise. This pooled classification turns out to be generally aligned with the official status of the places in the sample, but there are also some significant discrepancies.

For example, Jigani of Karnataka is administratively rural but meets most of the criteria to be considered urban in practice (figure 4a). The built-up area covers a substantial share of the place's surface, the buildings are compact, there is a network of roads both inside the place and linking it to external markets,

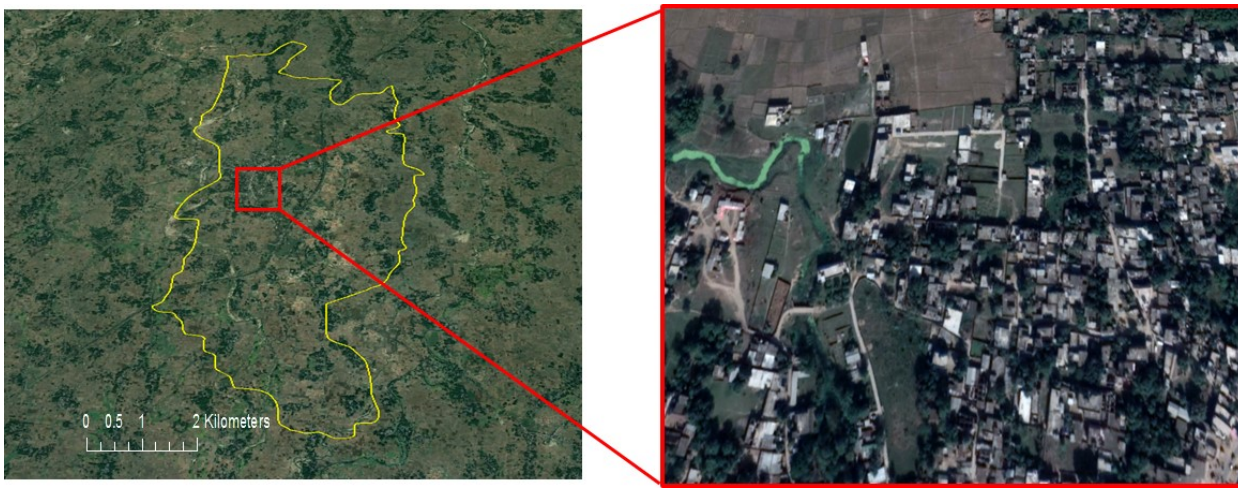
and signs of amenities are clearly visible. The place is classified as urban in 56 of the 57 assessments we have for it. Conversely, Bahadurganj of Bihar is administratively urban, but its cropland areas are vast, the built-up area is relatively small, buildings are scattered, amenities are few, and the road network is limited both within and outside of the place (figure 4b). This place is considered rural in 42 out of 57 assessments available.

Figure 4. Examples of discrepancies with the administrative classification

a. **Jigani (Karnataka) is reclassified as urban**



b. **Bahadurganj (Bihar) is reclassified as rural**



Note: The yellow lines depict the digitized administrative boundaries for the respective administrative units.

According to the pooled classification, 43 percent of the 1,277 places in the sample are urban in practice, and 57 percent are rural (table 4). Therefore, the sampling strategy does ensure a balanced representation of rural and urban places. A nonnegligible fraction of statutorily urban places is classified as rural in practice. Discrepancies are also large for places deemed urban in terms of the census, many of which are

assessed as rural in practice. But there are no substantial differences for most villages, whose rural status is confirmed by the subjective assessments.

Table 4. Classification based on human judgments versus administrative classification

Administrative classification	Assessed urban (percent)	Assessed rural (percent)	Total (percent)
Statutory towns	23.1	8.6	31.7
Census towns	20.0	23.6	43.6
Villages	0.2	24.5	24.7
Total	43.3	56.7	100.0

4. Machine learning

The urban status of places derived from human judgments can be used to infer the urban status of the other places in the universe. Doing this requires first learning from the sample and then using the results to make out-of-sample predictions. A first question is which indicators to consider for this exercise. Official measures of urbanization are based on traditional data sources, such as population censuses, whereas more recent studies rely on modern data, and especially on satellite imageries. A second question refers to the statistical approach to be used. A specific data generation process is assumed in the classical statistical tradition, but supervised machine learning is becoming increasingly popular. Our methodology brings together traditional and modern data, and compares the outcomes depending on whether classical econometric models or machine learning are used for the prediction.

Traditional and modern data

We chose as key covariates for the prediction a set of indicators which appear recurrently in the urban literature and are observable for all administrative units in India.

Traditional data were from the 2011 Census of India. These data were georeferenced to the digitized boundaries of administrative units as part of the Spatial Database for South Asia project. The range of georeferenced indicators available for all places in the country as of 2011 is wide (Li et al. 2015). Following the practice of national governments and international organizations, we selected *population* and *population density* at the town and village level as the key indicators. This choice is consistent with that of studies relying on high-quality microdata from the ground for more advanced economies (Michaels, Rauch and Redding 2012; Rozenfeld et al. 2011). Population density was computed as the ratio between the population of the administrative unit and the size of the area within the corresponding boundary.

Modern data were obtained from open-source products built on satellite imagery. Following more recent global products, built-up cover and nighttime lights are the preferred data in this respect. For built-up cover, we selected the Landsat-based product provided by the Global Human Settlement Layer, hereafter referred to as GHSL (Pesaresi et al. 2016). The product is derived from images taken by sensors aboard Landsat 7 and Landsat 8 satellites. 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. Landcover products

derived from Landsat 7 and Landsat 8 satellites have been widely used for academic research, because of their high spatial resolution, high spectral resolution and long temporal coverage (Burchfield et al. 2006).[5]

We compared GHSL with two other products reporting built-up cover. One of them is based on imagery from MODIS and the other on imagery from the twin satellites TerraSAR-X and TanSAR (Friedl et al. 2010; Esch et al. 2017).[6] Because of its higher resolution GHSL captures built-up cover in forest areas, such as Kerala, better than MODIS. The correlation between built-up data based on GHSL and those on TerraSAR-X and TanSAR imagery is very high. For the built-up share at the town and village level it reaches 0.96. In light of this, and also given the wide use of the Landsat-based landcover product, we chose GHSL 2013/14 as our data source. Our *built-up share* indicator was computed for every place as the ratio between its built-up area and its total area.

For nighttime lights, we relied on the Global Radiance Calibrated Nighttime Lights product. Data in this case are derived from images taken by the Operational Linescan System (OLS) sensors aboard satellites from the US Air Force Defense Meteorological Satellite Program (DMSP). The product has a higher quality than other products because it captures stable lights, addresses the saturation problem, and is intercalibrated to reduce inconsistencies across satellites (Elvidge et al. 2009; Hsu et al. 2015; NOAA 2014). The intensity of nighttime lights is reported in Digital Numbers (DN) as of 2011.

A well-known challenge when using data on nighttime lights is that even areas without human activity may appear to be lit up (Henderson et al. 2003; Li and Zhou 2017). This may be due to blooming, to the reflection of light from surrounding water, to the sensitivity of the sensors, or to geo-referencing errors made while capturing and composing the data. To avoid an 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 level of economic development.

To identify the appropriate threshold, we reviewed a dozen studies that together present 16 threshold estimates for different countries. We then matched their results with the corresponding gross domestic product per capita. We found a positive and statistically significant correlation between threshold estimates and product per capita. Using the estimated relationship, we concluded that the appropriate threshold for India in 2011 was 15 DN (Galdo et al. 2018). Our *lit-up share* indicator was thus computed as the fraction of a place's surface with light intensity above this threshold.

Beyond these four indicators, our framework is flexible to the inclusion of other metrics providing information on how urban a place may be in practice. Including these other metrics in the prediction model helps assess whether the four indicators we consider are informative enough on their own. Following the literature, in some of the analyses we added the normalized difference vegetation index

---

[5] The Landsat 7 satellite was launched in 1999. It has seven spectral bands at a spatial resolution of 30 m and a temporal frequency of 16 days. Landsat 8 was launched in 2013. It includes nine spectral bands with the same spatial resolution and temporal frequency as Landsat 8.

[6] MODIS based product relies on images taken by sensors aboard satellites from the Earth Observing System, known as the Land Cover Type Yearly Grid. The product has a resolution of 500 m and is available from 2001 onward. The imagery from the twin satellites TerraSAR-X and TanSAR are radar images, which are part of the recent WorldDEM program. Built-up cover data for 2011 are available at a resolution of 12 m.

(NDVI) and the normalized difference water index (NDWI) to improve prediction accuracy. Both of these metrics are derived from Landsat satellite imagery, with NDVI being related to the coverage of vegetation and NDWI related to the water content on Earth's surface.

The summary statistics for our key indicators differ between the sample and the universe, because the former was designed so as to over-represent urban places (table 5). As expected, the mean values for population, population density, the built-up share and the lit-up share are substantially higher in the sample than in the universe. On the other hand, the mean values of NDVI and NDWI are not statistically different, suggesting that they may not help improve the prediction substantially.

Table 5. Summary statistics of key indicators

	Observations	Mean	Std. Dev.
Selected sample			
Population size (*thousands*)	1,277	46.79	392.48
Population density (*people per km*^{2})	1,277	3085	6057
Built-up share (*percent*)	1,277	13.05	19.99
Lit-up share (*percent*)	1,277	52.97	45.45
NDVI (-1 to 1)	1,277	0.29	0.10
NDWI (-1 to 1)	1,277	0.09	0.09
All			
Population size (*thousands*)	564,052	2.10	33.53
Population density (*people per km*^{2})	564,052	693	2700
Built-up share (*percent*)	564,052	0.81	4.43
Lit-up share (*percent*)	564,052	8.57	25.93
NDVI (-1 to 1)	564,052	0.31	0.11
NDWI (-1 to 1)	564,052	0.08	0.09

Classical data modeling

Subjective assessments of the places in the sample can be used to infer the status of all other places in the universe. In line with the classical statistical tradition, this can be done using an econometric model, in which case an explicit probabilistic distribution is assumed for the data. The most common functional choices in the classical statistical tradition are the Logit and the Probit models, which respectively assume a logistic and a normal distribution for the data (Cameron and Trivedi 2005).

Following the classical statistical tradition, we specified a probabilistic model on whether a place is urban or rural in practice. We used the standard *Logit* model linking the urban status of a place with the values of its key indicators. The specification took the following form:

$$U_i = \begin{cases} 1 \ \ with\, probability \ \ p_i \\ 0 \ \ with\, probability \ \ 1 - p_i \end{cases}$$
$$p_i = \Pr[U_i = 1 | \, X_i] = F(X_i^T \beta \,) \qquad (1)$$

where U_i is the subjective assessment of place i, p_i is the expected likelihood that the place is urban given the values of its key covariates X_i, and $F(.)$ Is the logistic distribution function. A linear function of X_i was assumed for the entrance to $F(.)$.

We relied on maximum likelihood estimation, implying that parameters β were approximated by the values β_{Logit} that maximize the log likelihood function $\ln L(U_i|X;\beta)$:

$$\begin{aligned}\beta_{Logit} &= \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[U_i ln F\left(X_i^T \beta\right) + (1-U_i) ln\left(1 - F\left(X_i^T \beta\right)\right)\right] \right\} \\ &= \underset{\beta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{N} \left[U_i \left(X_i^T \beta\right) - ln\left(1 + exp\left(X_i^T \beta\right)\right)\right] \right\}\end{aligned} \tag{2}$$

The Logit model was applied to the pooled classification of the 1,277 places in the sample. The key indicators were population, population density, built-up share, and lit-up share. The corresponding coefficients had the expected signs, and most of them were statistically significant regardless of the specification chosen (table 6). The specification including these four key indicators provided the benchmark for our analysis. However, to enrich the model for prediction, we further expanded the covariates to include NDVI and NDWI, quadratic terms of individual indicators and terms interacting two indicators at a time. A comparison of results across specifications suggests that overall built-up share and population size are the most stable predictors of urban status.

The most critical step in the analysis is the classification of the out-sample places. Inferring prediction power from explanatory power is potentially misleading because there is a risk of overfitting the sample. Following the literature, we evaluated the different specifications on their prediction accuracy. A popular way of doing this is through cross-validation. This involves separating the sample into a training set, on which a prediction model is built, and a validation set, on which the model's performance is evaluated.

In our analysis we randomly partitioned the sample of 1,277 places into 10 equally sized subsamples, took one of the subsamples as the validation data and the rest as training data, and computed the percentage of observations correctly predicted. The process was repeated ten times, one with each of the ten subsamples used as the training data (Hastie et al. 2017; Mullainathan and Spiess 2017). Varying the number of folds or applying the Monte-Carlo cross-validation did not affect the conclusions.

The results of this 10-fold cross-validation confirmed that combining multiple sources of data increases prediction accuracy in the Logit model (figure 5a). The prediction accuracy of the benchmark specification reached 85.9 percent. This specification performed significantly better than the more parsimonious ones based exclusively on indicators from traditional data or from modern data. On the other hand, the richer specifications improved prediction accuracy over the benchmark specification only at the margin.[7]

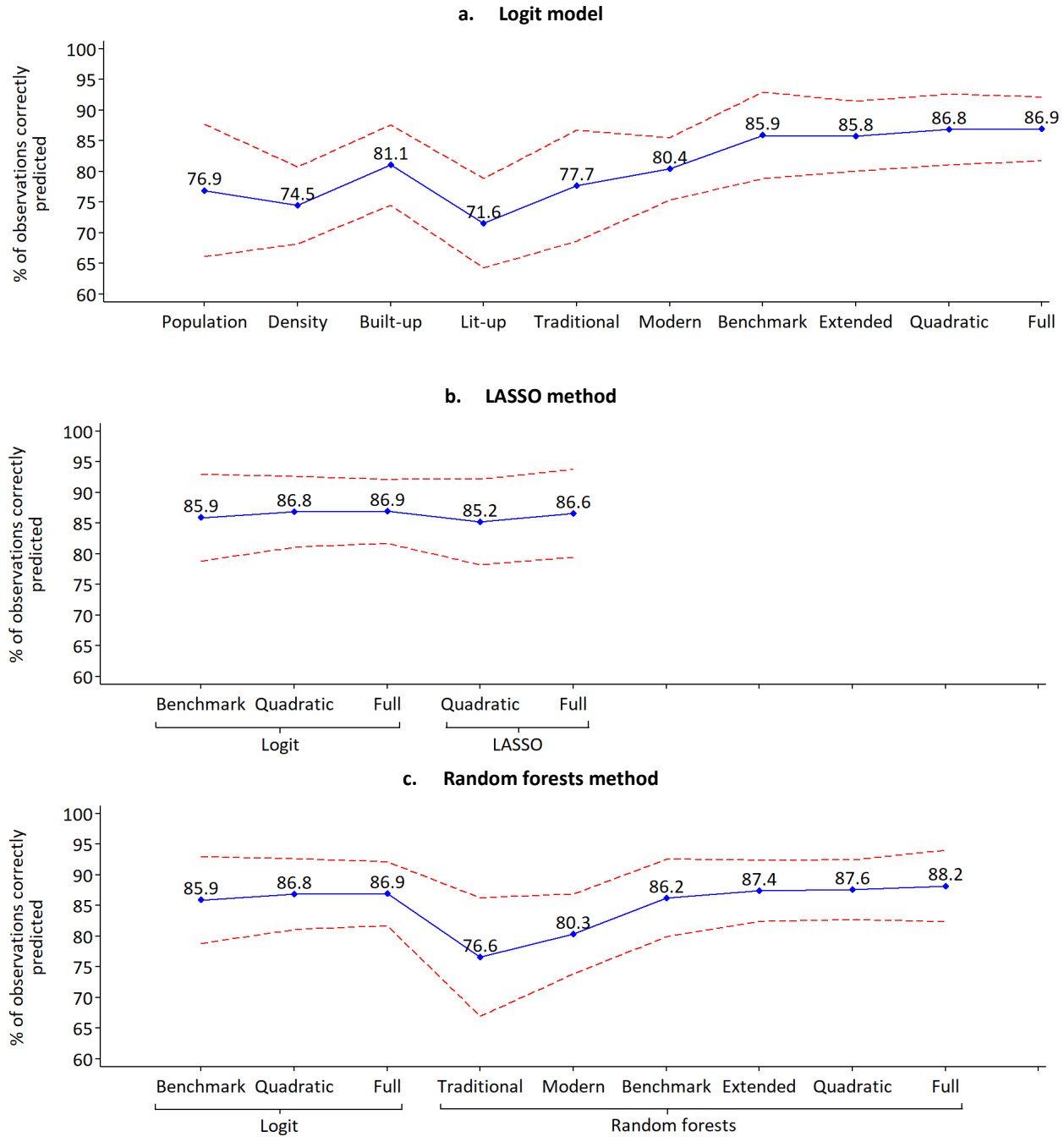
[7] We replicated all prediction exercises of all three approaches with even higher-order polynomials but found no improvement in prediction accuracy. Results are available upon request.

Table 6. Logit models on the probability of being classified as urban

	Traditional data			Modern data					Benchmark	Extended	Quadratic	Full
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Population	0.101***		0.0916***						0.105***	0.106***	0.0994***	0.163***
	(0.00715)		(0.00720)						(0.00848)	(0.00858)	(0.00984)	(0.0431)
Population density		0.346***	0.127***						0.0146	0.0171	0.144***	0.461***
		(0.0291)	(0.0243)						(0.0139)	(0.0150)	(0.0447)	(0.142)
Built-up share				0.134***				0.110***	0.0837***	0.0841***	0.168***	0.247***
				(0.00925)				(0.00940)	(0.00877)	(0.00883)	(0.0163)	(0.0490)
Lit-up share					0.0234***			0.0107***	0.0137***	0.0140***	0.0199*	0.0558***
					(0.00148)			(0.00177)	(0.00215)	(0.00218)	(0.0120)	(0.0179)
NDVI						-2.816***				0.612	-3.820	2.389
						(0.575)				(1.154)	(5.164)	(8.896)
NDWI							-0.415			-1.964	6.444**	7.253
							(0.639)			(1.217)	(2.906)	(7.259)
Population squared											-7.95e-06***	-9.92e-06
											(1.38e-06)	(7.15e-06)
Population density squared											-0.00250***	-0.00596***
											(0.000833)	(0.00164)
Built-up share squared											-0.00147***	-0.00216***
											(0.000196)	(0.000258)
Lit-up share squared											-9.32e-05	-0.000215*
											(0.000117)	(0.000126)
NDVI squared											6.534	4.058
											(8.223)	(14.25)
NDWI squared											-38.36***	-27.90*
											(10.68)	(16.44)
Constant	-1.562***	-1.235***	-1.794***	-1.519***	-1.603***	0.550***	-0.233***	-1.919***	-3.330***	-3.361***	-3.276***	-6.541***
	(0.0977)	(0.0887)	(0.109)	(0.0929)	(0.112)	(0.176)	(0.0817)	(0.122)	(0.201)	(0.370)	(0.768)	(1.584)
Interaction terms	No	No	No	No	No	No	No	No	No	No	No	Yes
Observations	1,276	1,276	1,276	1,276	1,276	1,276	1,276	1,276	1,276	1,276	1,276	1,276
Pseudo R2	0.282	0.148	0.306	0.317	0.170	0.0142	0.000242	0.338	0.506	0.507	0.553	0.597
Log likelihood	-626.5	-744	-606	-596.2	-724.3	-860.4	-872.6	-578.1	-431.4	-429.9	-390.5	-351.8
BIC	1267	1502	1233	1207	1463	1735	1760	1178	898.6	909.9	874	903.8

Note: The dependent variable is the urban status of the place according to the pooled classification. *Bechmark* refers to the four key indicators; *Extended* denotes the inclusion of NDVI and NDWI; *Quadratic* adds quadratic terms of individual indicators; and *Full* indicates that both quadratic terms and terms interacting two indicators at a time are included. Standard errors are reported in parentheses, and statistical significance is reported by *** if $p < 0.01$, ** if $p < 0.05$, and * if $p < 0.1$.

Figure 5. Prediction accuracy across approaches and specifications



Note: The solid blue line reports prediction accrucary and the dotted red lines presents its 95% confidency interval based on the 10-fold cross-validation. The x-axis presents the combination of covariates: *Population* denotes population size; *Density* denotes population density; *Built-up* denotes built-up share; *Lit-up* denotes lit-up share; *Tradional* represents population size and density; *Modern* represents built-up and lit-up shares; *Bechmark* represents the four key indicators; *Extended* denotes the inclusion of NDVI and NDWI; *Quadratic* adds quadratic terms of individual indicators; and *Full* indicates that both quadratic terms and terms interacting two indicators at a time are included.

Algorithmic modeling

Inferences from the sample can also be based on an algorithmic approach, which does not require any specific assumption on data distribution. This approach, known in computer science as machine learning, has gained prominence in the last the two decades thanks to the rapid increase in computing power and the availability of big data (Athey 2018; Mullainathan and Spiess 2017). In the case of urban economics, algorithmic modelling has been stimulated by the increasing availability of data measuring city characteristics at high frequencies and granular scale (Glaeser et al. 2016; Glaeser et al. 2018; Gorin et al. 2018; Varian 2014).

Following the literature, we also relied on machine learning to uncover generalizable patterns from the data and produce predictions on the outcome. To this effect we reviewed some of the tools most familiar to economists, including ridge regression, LASSO, nearest-neighbors, decision trees, random forests, neural networks and ensemble (Breiman 2001b; Hastie et al. 2017). The review led us to select the Least Absolute Shrinkage and Selection Operator (LASSO) as a parametric method and random forests as a non-parametric method.

The LASSO method applies a regularized regression to estimate parameter values in which an additional constraint is introduced to penalize large models and enhance prediction accuracy (Hastie et al. 2017; Mullainathan and Spiess 2017). Thus, parameters β are now approximated by the values β_{LASSO} that obtain from maximizing the log likelihood function $\ln L(U_i|X;\beta)$ subject to an additional constraint:

$$\beta_{LASSO} = \underset{\beta}{\operatorname{argmin}}\left\{-\sum_{i=1}^{N}\left[U_i\left(X_i^T\beta\right) - ln\left(1+exp\left(X_i^T\beta\right)\right)\right]\right\} \quad subject\ to \quad \sum_{j=1}^{K}|\beta_j| \le t$$

$$\beta_{LASSO} = \underset{\beta}{\operatorname{argmin}}\left\{-\sum_{i=1}^{N}\left[U_i\left(X_i^T\beta\right) - ln\left(1+exp\left(X_i^T\beta\right)\right)\right] + \lambda\sum_{j=1}^{K}|\beta_j|\right\} \tag{3}$$

As the penalty term λ in the LASSO method becomes smaller, the optimization problem in equation (3) converges to that in equation (2) and $\beta_{\mathrm{LASSO}} \to \beta_{Logit}$. But in the general case, the LASSO method identifies the covariates that have little predictive power and may contribute to overfitting and penalizes them by setting their coefficients to zero.

Given the ability of the LASSO method to significantly reduce the number of covariates, we considered rich initial specifications. The results were largely consistent with those of the Logit model as the four key indicators, together with some quadratic terms, emerged as the most important predictors of urban status. However, the coefficients associated with these four key indicators cannot be interpreted literally because the objective of the LASSO method is prediction, not explanation. Coefficients could change from sample to sample even if the out-of-sample prediction results remain similar. We applied the 10-fold cross-validation procedure to evaluate the performance of the LASSO method and found that its prediction accuracy was similar to that of the Logit model (figure 5b).

The random forests method, in turn, works as a combination of individual trees. In each tree, data are split into groups at various nodes, each defined by a threshold for a covariate. Unlike standard trees, the trees in random forests focus only on a randomly chosen subset of the covariates to develop their nodes. That way, the trees in the random forests method are less likely to correlate with each other and offer

different information on the underlying data generation process. The random forests method essentially averages over many noisy but approximately unbiased and un-correlated trees to reduce the variance of the predictor (Breiman 2001a; Hastie et al. 2017).

Following the literature, we considered a large number (B) of binary classification trees for our urban classification problem, each of them (b) grown on an independently bootstrapped subsample from the 1,277 sample of places. The bootstrapped subsample thus served as training data. For each tree we randomly selected k out of the K covariates in the full set, with the rule of thumb being that k should be smaller than $\sqrt{K}$ to reduce the correlation between trees. We also set a predetermined number of nodes (n) for each tree. Parameters B, k and n characterize the forest.

At each node of each tree, we chose the cutoff point of the selected covariate that ensured the greatest similarity among places in each of the two resulting groups, and the greatest diversity between places in the two groups. Similarity and diversity were evaluated based on the pooled classification of places, using the Gini index to identify the optimal cutoff point. We then considered the next node, again involving one of k randomly selected covariates, and repeated the same steps. We continued the process recursively until the predetermined number of nodes n was reached.

Proceeding this way, each of the B trees led to an unequivocal assessment of all places in the training subsample as either urban or rural. But such assessment is not necessarily the same as that resulting from the pooled classification. We then used the B trees developed on each of the independently bootstrapped subsamples to the entire universe. By proceeding this way, all places in the universe were assessed B times. The random forest predictor of the urban status of a place is derived by applying majority rule to these B assessments.

We tuned our random forests method by varying the number of trees (B) and the number of covariates to be randomly selected to develop each node (k). We relied on the standard out-of-bag (OOB) error rate to evaluate the performance of different combinations of parameter values. We found that the OOB error rate is minimized as the number of trees considered increases above 300 and the number of covariates to be selected declines to two (appendix 4). In light of this, we choose 500 trees and two covariates at each node to implement the random forests model.

To facilitate the comparison with the results from the Logit model, we also considered multiple combinations of covariates. We focused on mean decrease accuracy and mean decrease Gini to evaluate their importance. Overall, built-up share and population size emerged as the most important predictors of urban status (appendix 5). This finding is consistent with the results from the Logit model. However, the relative importance of different covariates is only indicative. This is because the random forests method is a prediction tool, not an explanation tool, and the estimated relative importance of predictors is also sensitive to sample and process.

Again, we evaluated the prediction accuracy of the random forests method using the 10-fold cross-validation procedure and found that it outperforms the Logit model (figure 5c). The percentage of observations correctly predicted reached 88.2 percent, 1 percentage point higher than with the best-performing specification of the Logit model. And the confidence interval was comparable. However, the results from the random forests method are consistent with those from the Logit model in terms of the selection of covariates. Including all four key indicators significantly improves prediction accuracy over using indicators based exclusively on traditional or modern data. On the other hand, the addition of NDVI and NDWI indicators, and of higher-order terms, yields marginal gains.

Robustness to subjective bias

We applied both classical statistical modeling and algorithmic modeling to the classification of the sample based on the pooled classification. This exercise led to high consistency of predictions regardless of the specific approach used. Overall the random forests method is the best performer, followed closely by the Logit model. But a relevant question is whether the prediction performance would have been similar had the sample classifications based on individual rounds of assessments been used instead.

We showed above that assessor characteristics and the nature of the protocol followed only had a modest impact on the classification of places in the sample, but their impact could potentially become significant when predicting the urban status for out-of-sample units. To evaluate this potential bias, we first trained the model on each of the eight classifications, including the pooled classification and the seven classifications based on individual rounds of assessments. To this effect we used the random forests method, given its higher prediction accuracy. We then generated a predicted classification and computed the percentage of agreement between this predicted classification and each of the other classifications.

The prediction model performs well regardless of which subjective classification is used for training and regardless of the subjective classification used for validation (table 7). The average prediction accuracy is between 84.8 and 88.4. Standard deviations are small in all cases. This confirms our previous finding on the high consistence between classifications.

5. How urban is India?

Our proposed methodology can be used to shed light on the ongoing debate about India's "true" urbanization rate. This can be done by applying the trained models from the sample to predict the urban status of all administrative units in our universe of places in 2011 and then estimating the population living in those administrative units we deem urban. This prediction can be compared with official estimates, with the outcomes of other studies and with recent global products that estimate land use.

The urbanization rate in 2011

We chose as the benchmark to estimate India's urbanization rate the pooled classification of places in the sample. Also, because of its higher prediction accuracy, we relied on the random forests method with the full set of covariates (key indicators, NDVI, NDWI, quadratic terms and interactive terms) as our preferred approach. If the predicted likelihood of being urban was above 0.5, we classified the place as urban; otherwise we considered it rural.

According to the census, 31.2 percent of the population lived in urban areas in 2011. But this figure corresponds to the entire country, whereas our universe of places excludes small states as well as villages for which information is missing. If only the 564,052 places considered in our analyses were retained, the census would yield an urbanization rate of 31.6 percent. This is only slightly higher than the 29.9 percent rate predicted by the random forests method with the full set of covariates.[8]

[8] Predictions also fall within a close range of the official estimate when the Logit model or the LASSO method are used or when smaller sets of covariates based on both types of data are considered. The results are available upon request.

Table 7. Cross-validation of predictions across classifications

		Validation									
				GWU students		MTurk- India		MTurk- USA		Prediction accuracy	
		Pooled	World Bank analysts	Impromptu judgment	Structured protocol	Impromptu judgment	Structured protocol	Impromptu judgment	Structured protocol	Average	Std. Dev.
Train	**Pooled**	88.1	84.7	86.5	88.3	87.8	86.6	87.8	88.1	87.2	1.24
	World Bank analysts										
	Structured protocol	84.6	87.2	80.3	82.3	83.6	86.8	85.8	88.0	84.8	2.67
	GWU Students										
	Impromptu judgment	85.8	80.3	88.7	91.1	88.5	80.9	85.1	82.5	85.4	3.92
	Structured protocol	87.0	81.8	90.1	89.2	88.0	82.9	86.3	84.4	86.2	2.98
	MTurk- India										
	Impromptu judgment	88.5	85.1	89.5	90.4	88.3	86.4	88.6	88.2	88.1	1.68
	Structured protocol	86.0	86.3	81.0	82.3	85.3	88.9	86.7	88.3	85.6	2.72
	MTurk- USA										
	Impromptu judgment	89.2	87.1	87.5	88.9	88.7	88.1	87.6	90.2	88.4	1.04
	Structured protocol	88.8	87.5	83.8	85.5	86.4	89.5	89.2	88.6	87.4	2.00

Note: Train indicates the classification used for prediction model development and validation indicates the classification used for prediction accuracy evaluation. The random forests method is used.

However, when relying only on indicators from traditional data, the estimated urbanization rate becomes significantly higher (table 8). It reaches 35.2 percent of the population when applying the random forests model and 32.1 percent when applying the Logit model. On the other hand, when only relying on indicators from modern data, the estimate can be significantly smaller. For example, the urbanization rate falls to 26.4 percent according to the Logit model.

Table 8. The predicted urbanization rate of India in 2011

	Urbanization rate		Jaccard index					
	Area	Popu-lation	With census		With other approaches			
Approach			Area	Pop.	Area	St. dev.	Pop.	St. dev.
Logit (full)	3.4	30.2	0.56	0.84	0.83	(0.043)	0.94	(0.014)
LASSO (full)	3.3	29.7	0.53	0.83	0.83	(0.047)	0.94	(0.016)
Random forests (full)	3.2	29.9	0.54	0.85	0.83	(0.065)	0.94	(0.021)
Logit (traditional)	4.6	32.1	0.53	0.83	0.52	(0.054)	0.84	(0.017)
Random forests (traditional)	5.1	35.2	0.38	0.74	0.41	(0.018)	0.76	(0.012)
Logit (modern)	2.9	26.4	0.38	0.74	0.62	(0.049)	0.84	(0.016)
Random forests (modern)	5.7	30.3	0.33	0.73	0.41	(0.017)	0.79	(0.010)

Note: Other approaches include the Logit model (benchmark, quadratic and full), the LASSO method (quadratic and full) and the random forests method (benchmark, quadratic and full).

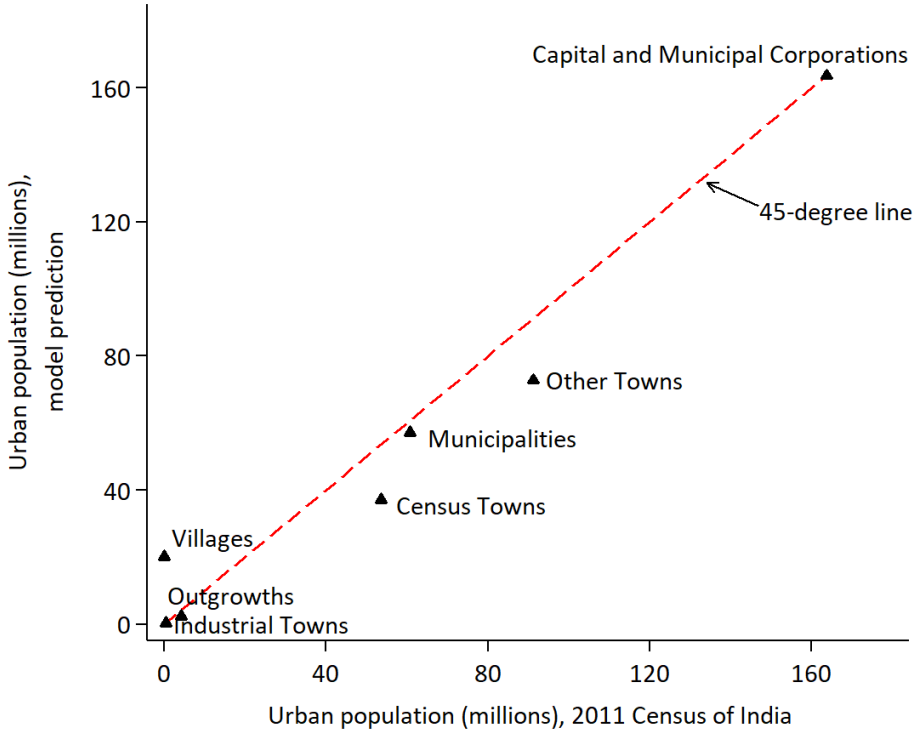
Estimations that come close at the national level may differ significantly at more disaggregated levels. Following de Bellefon et al. (2018), we use the Jaccard similarity index to measure the extent to which urban areas predicted by two different approaches overlap. A Jaccard index is the ratio between the size of the intersection of the urban areas predicted by two approaches and the size of the union of the two urban areas. When considering the full set of covariates, the average Jaccard index across approaches is 0.83 when size is measured in terms of surface, and 0.94 when it is measured in population terms. These results reveal a high coincidence in the delineation of the urban extent by our methodology even at disaggregated levels.

However, predictions are more volatile when relying exclusively on indicators from traditional or from modern data. For example, the average Jaccard index falls to 0.41 when measured by surface, and to 0.79 when measured by population, if the random forests prediction relies exclusively on modern data.

Urbanization by administrative category

The similarity between our predicted urbanization rate for 2011 and the census estimate could hide important differences at subnational levels. To explore this possibility more systematically, we first reran the comparison for places falling under each of the seven administrative categories considered by the 2011 Census of India (figure 6).

Figure 6. India's official and predicted urban population by administrative category



Note: The figure reports the official and predicted urban population for each of the seven administrative categories.

Our prediction suggests that many places deemed villages under the administrative classification have urban characteristics in practice. Based on the random forests method with the full set of covariates, about 7 percent of 555,292 places administratively classified as villages could well be urban. The gap may seem small in relative terms, but it has nontrivial implications when estimating the urban population. Overall, 20 million people resided in villages with urban characteristics in 2011.

Our results also reveal that many places administratively classified as towns could be deemed rural. Based on the random forests method with the full set of covariates, 48 percent of the 3,847 places labelled census towns, 48 percent of the 2,861 places administratively designated as other towns, and 16 percent of the 911 places deemed municipalities, could be considered rural. Taken together, this is the equivalent of about 22 million people whose classification as urban residents could be questioned.

Substantial gaps in urbanization rates relative to official figures are corroborated, in the case of municipalities, other towns, census towns and villages, when using other combinations of covariates for

the random forests method, as well as when relying on predictions based on the Logit model and the LASSO method.[9]

These gaps in urbanization rates by administrative category could be interpreted as the outcome of prediction error in both directions, around otherwise consistent urbanization rates at the aggregate level. However, it should be noted that the gap does not affect all administrative categories in the mid-range between rural and urban to the same extent. In relative terms, the gap is much larger among census towns and other towns than among municipalities and outgrowths.

These uneven gaps suggest that there could be more than prediction error at play, and the way places are administratively classified may matter as well. Different biases by administrative levels may indeed reflect an official classification of places that is not exclusively based on their urban characteristics but is also affected by fiscal incentives and statistical initiatives.

In India, the implementation of central government programs and the financing available for the development of settlements depend on the urban or rural status of beneficiary places. Statutory towns are declared by the state governments and governed by urban local authorities. These urban local governments have taxation authority, but they are also subject to a complex set of rules, regulations, building bylaws and development planning controls known as "urban laws". Rural local governments have no taxation powers, but they receive a host of transfers in support of farmers and the rural poor. As a result, there are numerous examples of statutory urban places being reclassified back and forth, especially other towns (Aijaz 2017; Mukhopadhyay et al. 2012; Ministry of Finance 2017).

Census towns are not even the result of an explicit reclassification of places by local authorities, but rather a statistical attempt by the Census of India to generate a more accurate estimation of the urban population. Census towns started being defined consistently in 1971, as an addition to the statutory towns declared by the state governments. They correspond to places that are administratively rural, host more than 5,000 inhabitants, have a population density exceeding 400 persons per square kilometer, and where over 75 percent of their male working population engaged in non-agricultural activities (2011 Census of India). These places are deemed urban by the Census of India, but not by local and state authorities.

Urbanization by state

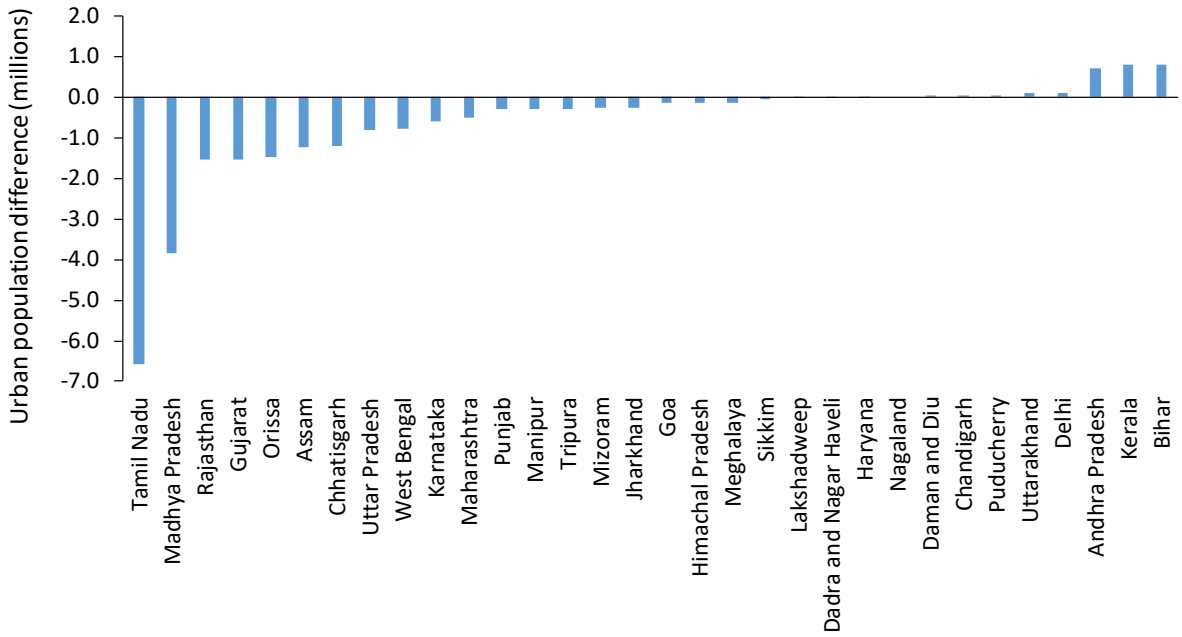
There are also significant differences between our predicted urbanization rates and official figures at the state level. Again, using the random forests method with the full set of covariates, our estimate of the urban population falls short of the official estimate by roughly 1.5 million people in each of the states of Orissa, Gujarat and Rajasthan. The gap is even bigger in Madhya Pradesh, where it reaches 3.9 million people, and especially in Tamil Nadu, where it approaches 6.6 million people. Conversely, our estimates of the urban population exceed the official figures by 0.8 million people in both Bihar and Kerala, and by 0.7 million in Andhra Pradesh (figure 7).

While these gaps in the estimated urban population could reflect prediction error, there are grounds to believe that they result, at least in part, from prevailing institutional arrangements. In India, urban development is a subject under the purview of state governments. By default, all settlements are rural

[9] The conclusions also hold when using classifications based on different rounds of assessments. The results are available upon request.

and they become urban only after the state government converts them, following a well-specified legal process. Converted settlements are labelled statutory towns.

Figure 7. Gap between India's predicted and official urban populations by state



Note: The figure presents the difference between the urban population by state based on our prediction and that based on the official census. The unit of measurement is million people.

There are guidelines for classifying a settlement as a statutory town, but they are vague and not binding. As a result, state governments exert large discretion in their choices, and the criteria to define a statutory town vary considerably across states. For example, in the southern coastal states the threshold population of statutory towns varies from around 2,000 in Tamil Nadu to over 20,000 in Kerala and above 30,000 in Andhra Pradesh (Aijaz 2018; Denis and Marius-Gnanou 2011). The first state has the least strict criteria in the country while these last two states have among the strictest, which is consistent with the sizeable gap between our predicted urbanization rates and the corresponding official figures.

A comparison with other estimates

Our results are in contrast with those reported in several studies using traditional data. Measures based on different thresholds for population size yield urbanization rates ranging from 47.2 to 64.9 percent for 2011 and those based on different cutoffs for population density produce urbanization rates at 55.0 and 78.0 percent (IDFC Institute 2015; Ministry of Finance 2017). In a World Bank report, a measure based on an agglomeration index that combines population density, the population of a "large" center and the estimated travel time to that center yields an urbanization rate of 55.3 percent in 2011 (Ellis and Roberts 2016).

The discrepancies between these measures and our predicted urbanization rates are substantial. But such discrepancies should not come as a surprise, as these other measures are more parsimonious in their definition of what makes a place urban. In light of these differences, we see these studies as offering interesting perspectives, but not fundamentally challenging our results.

The discrepancy is also substantial relative to some recent urban layer products based on satellite imagery that report high urbanization rates. The Geopolis project proposes a measure based on built-up cover, cell contiguity and a population threshold, and predicts an urbanization rate of 42.0 percent for 2011 (Denis and Marius-Gnanou 2011). The GHSL project applies thresholds on share of built-up area, cell contiguity, population size and population density to define urban cores. Based on this approach, GHSL estimates that 53.6 percent of India's population was urban in 2015 (Dijkstra et al. 2018; Pesaresi et al. 2019).[10] These figures exceed our predicted urbanization rate, sometimes by a wide margin.

Meanwhile, other recent studies based on satellite imagery suggest the urbanization rate could be close to or even lower than our prediction. Baragwanath et al. (2019) apply an algorithmic approach with different distance buffers to define urban markets for India. Their estimates on the share of India's population residing in urban markets in 2011 have a median value of 29.1 percent and a mean value of 27.6 percent. As another example, the Metropolitan Areas Extension Database or BEAM project uses carefully cleaned nighttime lights to identify urban areas (Ch et al. 2018). Based on its results, 23.9 percent of India's population lived in urban areas in 2010.

On closer inspection, however, the difference between our estimates and those in other recent studies does not arise from the underlying land classification, but rather from the population data used. This can be seen by comparing our estimates to two global urban layer products relying on satellite imagery and available as open source data, namely BEAM and GHSL. As mentioned above, a recent study based on BEAM leads to a lower urbanization rate than our methodology, while a study based on GHSL delivers a much higher urbanization rate.

For the Indian places considered in our universe, the total area we classify as urban based on our methodology is quite close to that of the GHSL urban core and higher than the BEAM urban area (table 9). However, these urban areas translate into very different estimates of the urbanization rate depending on which gridded population data are used. We consider three products in this respect: GHSL for 2015, LandScan for 2011 and WorldPop for 2010.[11]

Our predicted urbanization rate for India would have declined only marginally had we used population data from LandScan, instead of data from the 2011 Census of India. But it would have been much smaller if we had used population data from WorldPop and much higher had we used population data from GHSL. The similarity of results, relative to the 2011 Census of India, makes the use of Landscan data more appealing than the alternatives.

[10] GHSL country statistics can be generated from the Community pre-Release of GHS Data Package https://ghsl.jrc.ec.europa.eu/CFS.php.

[11] Information on gridded population is provided by Schiavina et al. (2019) for GHSL, by LandScan™ (2011) for LandScan and by WorldPop (2017) for WorldPop.

Table 9. India's urban area and urban population by global urban layer products

	Urban area (percent of land area)	Urban population (percent of population)			
		Census	GHSL	LandScan	WorldPop
Prediction	3.2%	29.9%	39.6%	28.8%	22.7%
GHSL	2.9%	-	54.7%	29.5%	23.1%
BEAM	1.9%	-	29.0%	26.9%	20.1%

Combining the GHSL's classification of urban cores with population data from LandScan yields an urbanization rate that is less than half a percentage point apart from our preferred estimate (29.9 percent). Similarly, applying the same approach to BEAM urban areas leads to an urbanization rate that is 3 percentage points lower than our preferred estimate. Therefore, when relying on the same population data the urbanization rates predicted by other recent studies are quite similar to the one we obtain with our methodology.

Discrepancies between our proposed methodology and global urban layer products become much wider when using other population data. All predicted urbanization rates decline when relying on WorldPop data and increase when using GHSL data instead. But regardless of the population data used our predicted urbanization rate falls in between the BEAM-based and GHSL-based estimates. We interpret this as further evidence that our predicted urbanization rate is not an outlier.

6. Conclusion

When assessing the urban extent there is value in relying on what one "sees," especially in countries where urbanization is messy in nature. 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 methodology for the subjective assessment of urban status across very diverse places.

Human judgment could be dismissed as a tool on the grounds that much rests on the eyes of the beholder. However, our methodology convincingly addresses this potential shortcoming. It shows indeed that the classification of places only changes at the margin when involving assessors with different backgrounds or following different protocols.

The increased availability of satellite imageries, of crowdsourcing tools and of computing capability also makes human judgment a scalable methodology. Physically visiting thousands of places to assess their status would be costly and time-consuming. But Google images make a remote assessment possible in a growing number of countries. Crowdsourcing makes it efficient to collect judgments for a representative sample of places. And the availability of remote sensing data, especially of built-up cover and nighttime lights, makes it viable to predict the urban status of thousands of other places.

Building on the burgeoning literature that uses machine learning to study economic issues, we show that the outcome of the prediction is not significantly affected by the prediction approach used—statistical or algorithmic. It is not substantially affected by the number of indicators considered either: as long as traditional and modern data are combined, predictions are very similar.

In passing, our analyses shed light to the ongoing debate on India's urbanization rate. In recent years, studies using different population thresholds, or combining traditional indicators into more complex indices, have claimed that the actual urbanization rate is higher than official figures suggest. Analyses based on global urban layer products seemingly confirm this conclusion. However, we find relatively minor discrepancies with official figures, and identify the institutional and statistical mechanisms that could underlie the observed gaps. Importantly, we show that the discrepancy with estimates based on global urban layer products is mainly due to the population data used, not to the resulting land classification.

The methodology we propose could also serve as the basis for further research. In this paper we apply it to the classification of places at one point in time, but it would be interesting to explore how to use it to assess changes in urbanization. Our results also suggest that the discrepancies in the assessment of the urban areas vary across states. A relevant question is whether accuracy would increase if prediction approaches were fine-tuned for places operating under different institutional settings or facing diverse geographical conditions.

Finally, our methodology uses administrative boundaries as the unit of analysis, whereas global products are built on grid cells of standard size. Digitized administrative boundaries—the polygons delimiting the jurisdiction of cities, towns, and villages—are an important anchor for data integration. They can be used as the basis for combining traditional and modern data, and therefore to extend prediction beyond the places virtually visited by assessors. A more systematic evaluation of the tradeoffs faced when relying on administrative boundaries, instead of grid cells, is therefore warranted.

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.

Aijaz, R., 2017. Measuring Urbanisation in India. ORF Issue Brief, 218. Observer Research Foudation.

Alonso, W. 1964. "Place and Land Use: Toward a General Theory of Land Rent." Harvard University, Cambridge, MA.

Athey, S., 2018. The impact of machine learning on economics. In The economics of artificial intelligence: An agenda. University of Chicago Press.

Beller, A., Borjas, G., Tienda, M., Bloom, D. and Grenier, G., 1994. "Beauty and the Labor Market." *The American Economic Review* 84(5): 1174-1194.

Bosker, Maarten, Jane Park and Mark Roberts, 2018. "Definition Matters: Metropolitan Areas and Agglomeration Economies in a Large Developing Country," *CEPR Discussion Papers 13359*, C.E.P.R. Discussion Papers.

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.

Briant, A., Combes, P.P. and Lafourcade, M., 2010. Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?. Journal of Urban Economics, 67(3), pp.287-302.

Brueckner, Jan K. 1987. The structure of urban equilibria: A unified treatment of the Muth-Mills model. In Edwin S. Mills (ed.) Handbook of Regional and Urban Economics, volume 2. Amsterdam: North-Holland, 821–845.

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

Campbell, J.B. and Wynne, R.H., 2011. *Introduction to remote sensing*. Guilford Press.

Ch, Rafael, Diego Martin, and Juan F. Vargas. 2018. "Measuring cities with nighttime light data." Processed, CAF Development Bank of Latin America.

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.

de Bellefon, Marie-Pierre, Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon. 2018. "Delineating Urban Areas using Building Density." Working Paper No. 811, The Wharton School, University of Pennsylvania.

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*.

Diegel, Jonathan, Antonio Miscio and Donald R. Davis. 2019. "Cities, Lights and Skills in Developing Economies." *Journal of Urban Economics*. In Press, Corrected Proof, Available online 23 May.

Dijkstra, Lewis, Aneta Florczyk, Sergio Freire, Thomas Kemper, and Martino Pesaresi. 2018. “Applying the degree of urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation.” Processed, European Commission.

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

Dubey, A., Naik, N., Parikh, D., Raskar, R. and Hidalgo, C.A., 2016, October. Deep learning the city: Quantifying urban perception at a global scale. *In European conference on computer vision* (pp. 196-212). Springer, Cham.

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.

Eeckhout, J., 2004. Gibrat's law for (all) cities. American Economic Review, 94(5), pp.1429-1451.

Ellis, Peter and Roberts, Mark. 2016. “Leveraging Urbanization in South Asia: Managing Spatial Transformation for Prosperity and Livability”. Washington, DC: World Bank.

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.

Galdo,Virgilio, Yue Li and Martin G. Rama. 2018. "Identifying Urban Areas by Combining Data from the Ground and from Outer Space : An Application to India," Policy Research Working Paper Series 8628, The World Bank.

Glaeser, E.L., Hillis, A., Kominers, S.D. and Luca, M., 2016. Crowdsourcing city government: Using tournaments to improve inspection accuracy. American Economic Review, 106(5), pp.114-18.

Glaeser, E.L., Kominers, S.D., Luca, M. and Naik, N., 2018. Big data and big cities: The promises and limitations of improved measures of urban life. Economic Inquiry, 56(1), pp.114-137.

Gorin, Clément, Pierre-Philippe Combes, Gilles Duranton, and Laurent Gobillon 2018. A random forest approach to mining historical map data: Land use in 19th-century France. (Processed)

Hamermesh, D.S. and Biddle, J.E., 1994. "Beauty and the Labor Market". *American Economic Review* 84(5): 1174-1194.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2017. "The Elements of Statistical Learning: Data Mining, Inference and Prediction." Second Edition (Corrected 12th printing), Springer, New York.

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.

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

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.

LandScan™ 2011. "High Resolution global Population Data Set copyrighted by UT-Battelle", LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy.

Levy, M., 2009. Gibrat's law for (all) cities: Comment. American Economic Review, 99(4), pp.1672-75.

Li, X. and Zhou, Y., 2017. Urban mapping using DMSP/OLS stable night-time light: a review. International Journal of Remote Sensing, 38(21), pp.6030-6046.

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

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 Alplace in a Metropolitan Area." *American Economic Review* 57 (2): 197–210.

Minsitry of Finance. 2017, *Economic Survey 2016-17*, Government of India, Ministry of Finance, Department of Economic Affairs.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2): 87-106.Muth, R. F. 1969. *Cities and Housing: The Spatial Patterns of Urban Residential Land Use.* Chicago: University of Chicago Press.

Mukhopadhyay, P., Zerah, M.H. and Denis, E., 2012. Subaltern urbanisation in India. *Economic and Political Weekly*, XLVII, 30, pp.52-62.

Naik, N., Kominers, S.D., Raskar, R., Glaeser, E.L. and Hidalgo, C.A., 2015. Do people shape cities, or do cities shape people? The co-evolution of physical, social, and economic change in five major US cities (No. w21620). National Bureau of Economic Research.

Naik, N., Raskar, R. and Hidalgo, C.A., 2016. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *American Economic Review*, 106(5), pp.128-32.

NOAA (National Oceanic and Atmospheric Administration). 2014. Global Radiance Calibrated Night Lights, National Geophysical Data Center, n

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.

ORGI (Office of the Registrar General and Census Commissioner). 2011a. Census of India—*2011.* New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

_____. 2011b. *Administrative Atlas—2011.* New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011census/maps/administrative_maps/admmaps2011.html.

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; Florczyk, Aneta; Schiavina, Marcello; Melchiorri, Michele; Maffenini, Luca. 2019. "GHS settlement grid, updated and refined REGIO model 2014, multitemporal (1975-1990-2000-2015)", R2019A. European Commission, Joint Research Centre (JRC).

Postma, E. 2014. "A Relationship between Attractiveness and Performance in Professional Cyclists". *Biology Letters* 10(2): 20130966.

Ross, J., Irani, I., Silberman, M. Six, Zaldivar, A., and Tomlinson, B., 2010. Who are the Crowdworkers?: Shifting Demographics in Amazon Mechanical Turk. In: CHI EA 2010. (2863-2872).

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 (5): 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.

Schiavina, Marcello; Freire, Sergio; MacManus, Kytt 2019. "GHS population grid multitemporal (1975, 1990, 2000, 2015) R2019A". European Commission, Joint Research Centre (JRC)

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

Straszheim, M. 1987. "The Theory of Urban Residential Place." In *Handbook of Regional and Urban Economics* 2: 717–57. Amsterdam: Elsevier.

UN (United Nations). 2005. "Designing Household Survey Samples: Practical Guidelines." Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

US Office of Management and Budget, 2010. 2010 standards for delineating metropolitan and micropolitan statistical areas; Notice. Washington, D. C.: Federal Register.

Varian, H.R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), pp.3-28.

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.

WorldPop. 2017. “India 100m Population”, Version 2. University of Southampton. DOI: 10.5258/SOTON/WP00532.

Appendices

Appendix 1. The formula for sampling size determination

We use the following formula to determine the sample size for each administrative category:

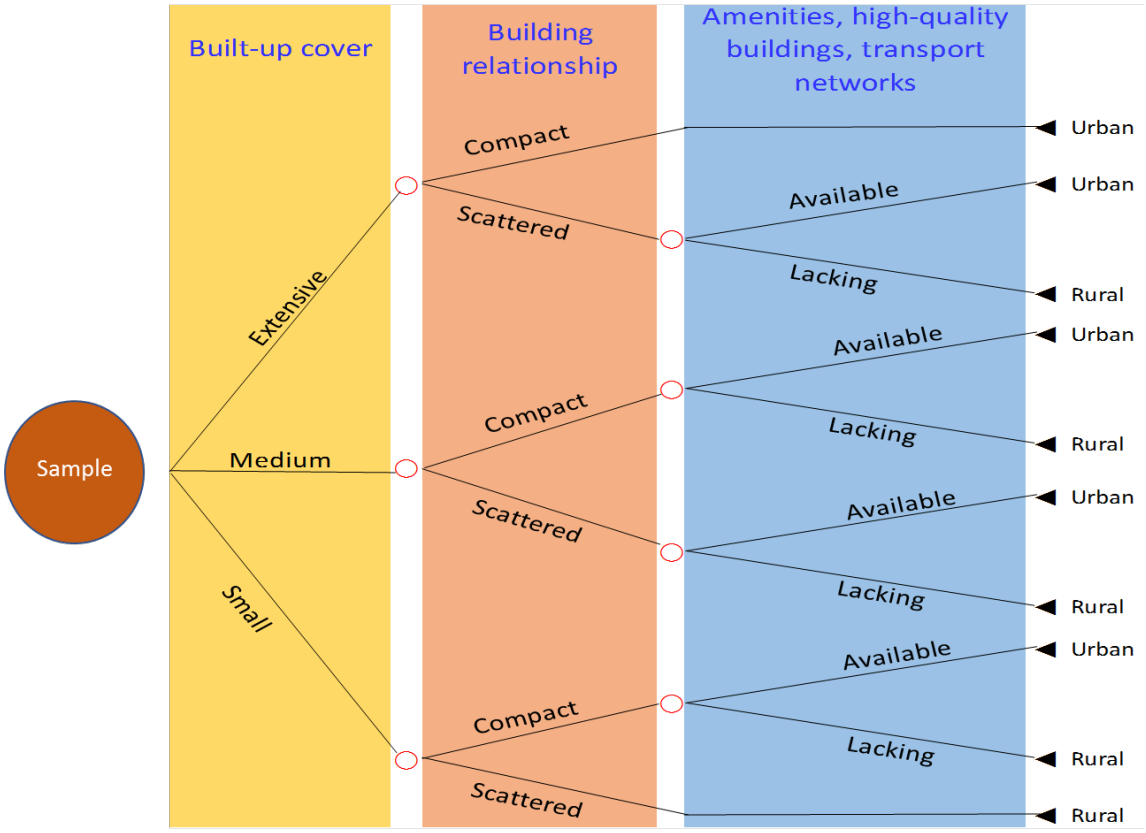
$$n_i^* = \frac{\frac{z^2 p_i(1-p_i)}{e^2}}{1+\left(\frac{z^2 p_i(1-p_i)}{N_i {e_i}^2}\right)} \quad (1)$$

where n_i^* is the sample size for administrative category i; N_i is the total number of places belonging to the category; and p_i is the prior probability of the place being urban in practice. In the absence of other information, the prior probability p_i should be set equal to 0.5. However, administrative categories provide information on the likelihood that a place is urban or rural in practice, so that the values of p_i should be correlated with the order of these administrative categories in the rural-urban gradation.

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. Based on the visual analysis of a random sample of 250 administrative units, 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.

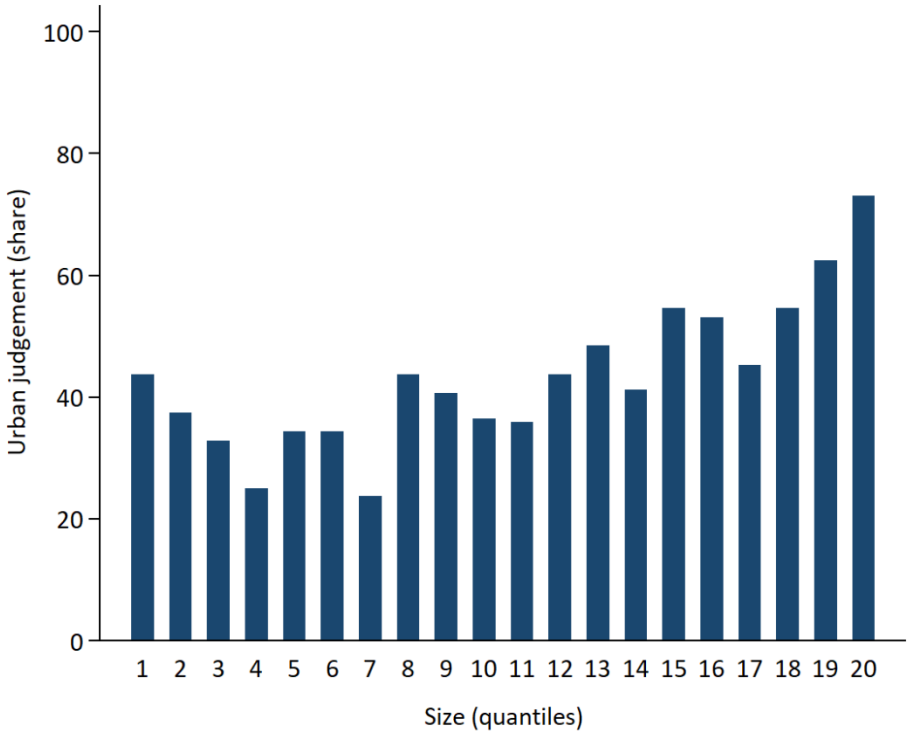
Appendix 2. The structured assessment protocol

The structured protocol to assess the urban status of places in the sample is based on a decision tree involving three steps. First, the assessor focuses on land cover. If the built-up area appears to be extensive as a share of the overall area, the place is likely to be urban; it is likely to be rural if the built-up area is small. Between these two extremes, the assessment is inconclusive. Next, the assessor adjusts this preliminary judgment based on the relationship among buildings. Compactness or clustering of buildings increases the likelihood that the place is urban, whereas a scattered pattern of buildings suggests that the place is rural. Finally, the assessor zooms in (and pulls out street views if available) to check the availability of amenities, high-quality buildings, and transportation networks. The availability of some of these structures confirms that the place is urban in practice.



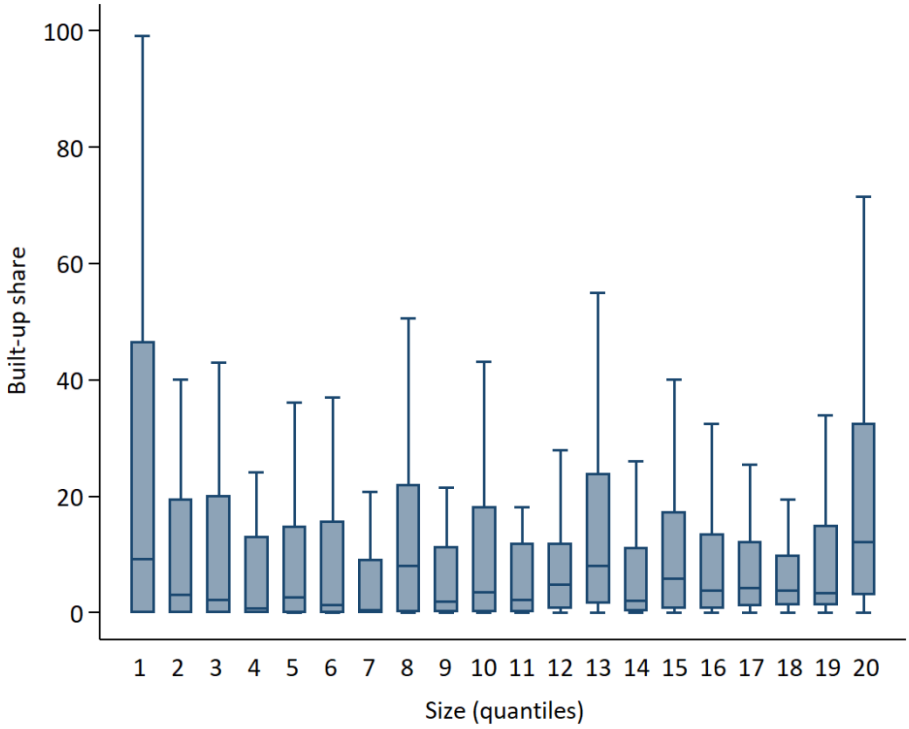
Appendix 3. The impact of the size of places on judgment outcomes

a. The relationship between size and judgments Size



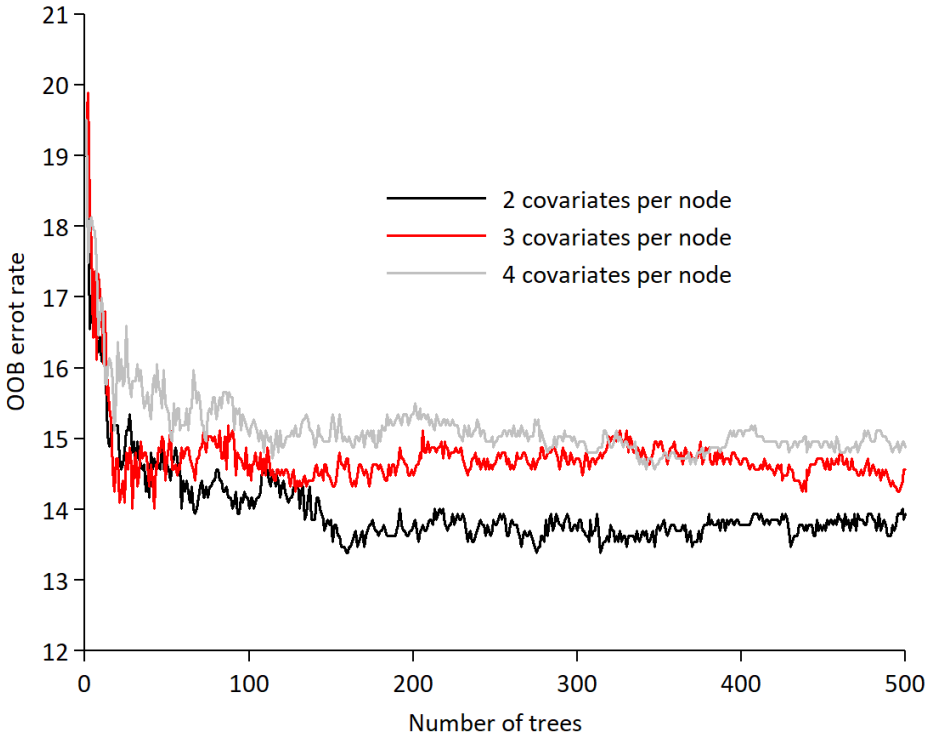
Note: The blue bars report the share of places judged to be urban for a size quantile.

b. The relationship between size and built-up share



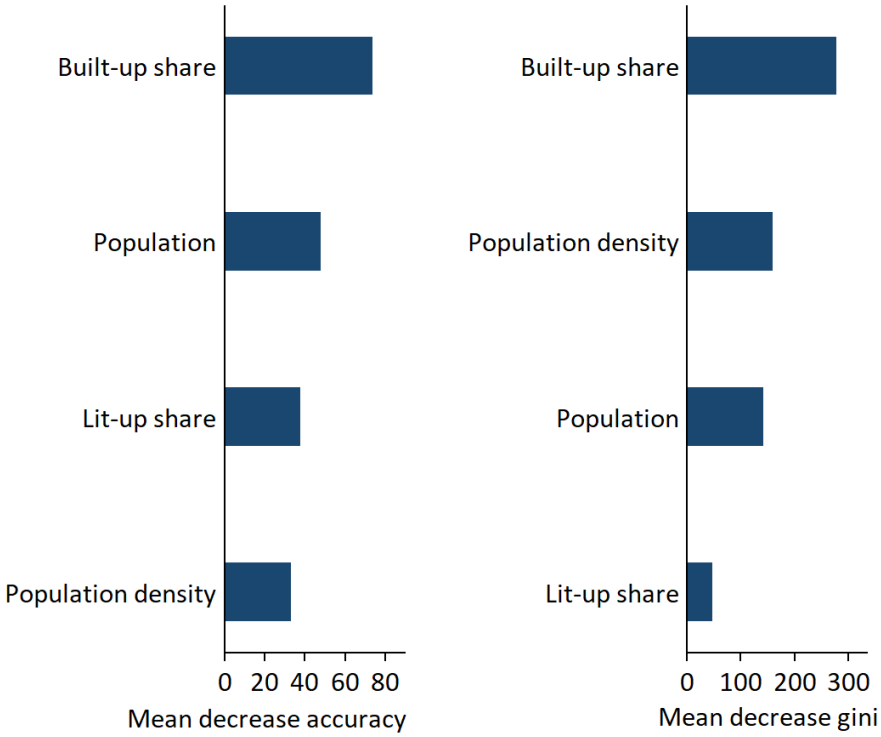
Note: The blue bars report the range of the built-up share defined by the 25 and 75 percentiles for a size quantile. The blue lines within the bars report the median value of the built-up share for a size quantile.

Appendix 4. Tuning of random forests



Note: The figure reports out-of-bag (OOB) error rate for different combinations of the number of trees and the number of covariates to be sampled to develop each node when the four key indicators, NDVI, NDWI, the quadratic terms of individual indicators and terms interacting two indicators at a time are included as covariates in the random forests analysis.

Appendix 5. The importance of covariates in the random forests analysis



Note: The figures report the Mean Decrease Accuracy and Mean Decrease Gini for the four key indicators when all of them are included as covariates in the random forests analysis. Mean Decrease Accuracy gives a rough estimate of the loss in prediction performance when a covariate is omitted from the training set. Mean Decrease Gini estimates the important of a covariate is to split the data correctly. It does so by assessing the decrease in Gini, which is a measure of node purity, when a covariate is omitted.