

Article

# Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine

Ran Goldblatt <sup>1,\*</sup>, Wei You <sup>2</sup>, Gordon Hanson <sup>1</sup> and Amit K. Khandelwal <sup>3</sup>

<sup>1</sup> School of Global Policy and Strategy, University of California, San Diego, CA 92093, USA; gohanson@ucsd.edu

<sup>2</sup> Department of Economics, University of California, San Diego, CA 92093, USA; wyou@ucsd.edu

<sup>3</sup> Columbia Business School, Columbia University, New York, NY 10027, USA; ak2796@columbia.edu

\* Correspondence: rgoldblatt@ucsd.edu; Tel.: +1-858-822-5087

Academic Editors: Soe Myint and Prasad S. Thenkabail

Received: 26 April 2016; Accepted: 26 July 2016; Published: 1 August 2016

**Abstract:** Urbanization often occurs in an unplanned and uneven manner, resulting in profound changes in patterns of land cover and land use. Understanding these changes is fundamental for devising environmentally responsible approaches to economic development in the rapidly urbanizing countries of the emerging world. One indicator of urbanization is built-up land cover that can be detected and quantified at scale using satellite imagery and cloud-based computational platforms. This process requires reliable and comprehensive ground-truth data for supervised classification and for validation of classification products. We present a new dataset for India, consisting of 21,030 polygons from across the country that were manually classified as “built-up” or “not built-up,” which we use for supervised image classification and detection of urban areas. As a large and geographically diverse country that has been undergoing an urban transition, India represents an ideal context to develop and test approaches for the detection of features related to urbanization. We perform the analysis in Google Earth Engine (GEE) using three types of classifiers, based on imagery from Landsat 7 and Landsat 8 as inputs. The methodology produces high-quality maps of built-up areas across space and time. Although the dataset can facilitate supervised image classification in any platform, we highlight its potential use in GEE for temporal large-scale analysis of the urbanization process. Our methodology can easily be applied to other countries and regions.

**Keywords:** Google Earth Engine; Landsat; remote sensing; urbanization; built-up land cover; pixel-based image classification

## 1. Introduction

Over the past century, many countries, especially in the developing world, have experienced rapid urbanization [1,2]. Between 1950 and 2014, the share of the global population living in urban areas increased from 30% to 54%, and by 2050 it is projected to expand by an additional 2.5 billion urban dwellers, primarily in Asia and Africa [3]. Urbanization also entails an increase in the land area incorporated in cities, which over the next 15 years is projected to grow by 1.2 million km<sup>2</sup> [4]. The process of urbanization profoundly influences economic [5] and social development [1], and has direct consequences for biodiversity, resource conservation, and environmental degradation [4,6,7].

Previous literature measures the extent of urban areas using household-survey-based socio-economic data, nighttime lights, and mobile-phone records. With the increasing availability of satellite imagery at ever-improving spatial and temporal resolutions, urban research is shifting towards the use of digital, multispectral images and towards the development of remote-sensing image

classification designed to capture urban land features [8–10]. The availability of earth-observation data, acquired primarily by Landsat and MODIS satellites, has triggered the development of several classification maps of urban areas [11–13], including multi-class land-cover maps, binary maps that indicate the presence/absence of urban land cover, and maps of variables associated with urban areas, such as impervious surfaces and nighttime light generation [14].

In parallel, cloud-based computational platforms have become increasingly accessible and allow one to scale analysis across space and time. One such platform is Google Earth Engine (GEE). GEE leverages cloud-computational services for planetary-scale analysis and consists of petabytes of geospatial and tabular data, including a full archive of Landsat scenes, together with a JavaScript, Python based API (GEE API), and algorithms for supervised image classification.

By definition, supervised classification requires ground-truth labeled data. Several datasets have been proposed to serve as ground-truth for urban research. These include gazetteer datasets of city locations; datasets of sites and boundaries, which are digitized, rated, and assessed by expert analysts; medium-resolution Landsat-based urban maps [12]; and census-based population databases [15]. Crowd-sourced datasets, such as OpenStreetMap (OSM) can also be used to map urban areas [16,17], especially when they are combined with remotely-sensed settlement and land cover data [11]. OSM is a valuable source for ground-truth data, primarily because of its vast extent and free availability. However, the completeness of OSM and its suitability for urban research is subject to the number and reliability of OSM contributors [18]. The use of OSM for supervised image classification remains challenging due to the risk of imbalanced distribution of class labels (including their spatial coverage), the presence of errors or missing class assignments (“class-noise”), and inaccurate polygon boundary delineations [19].

Despite the significant progress in the field of machine learning and the increasing availability of satellite imagery, there is still a scarcity of ground-truth labeled datasets that have been developed specifically to detect urban areas [20]. In this study, we aim to fill the need for this valuable data, and to provide, for the first time, reliable and comprehensive open-source ground-truth data for supervised classification that delineates urban areas in one country. We present a new dataset consisting of 21,030 polygons in India that were manually labeled as “built-up” or “not built-up” and use these data for supervised image classification and detection of urban areas. As a large and geographically diverse country that has been undergoing an urban transition, India represents an ideal context to illustrate the applicability of our approach for mapping urbanization. The results demonstrate the potential for integrating high-resolution satellite imagery, cloud-based computational platforms and ground-truth data to measure and to analyze the urbanization process. Although our study focuses on India, the methodology we develop can easily be applied to other parts of the world.

This study differs from previous efforts to map urban areas in four respects. First, we construct a large-scale and comprehensive georeferenced dataset that is designed for the express purpose of mapping urban areas. We make it available and accessible for the use and validation of existing classification products. As noted above, validated ground truth datasets are in short supply and many of those that do exist are small in size or spatial extent. Second, we validate this dataset and demonstrate its applicability for mapping urban areas at the national level for India. We present, in one study, an assessment of alternative classifiers and examine the effect of various inputs and class combinations on the performance of the classifiers. Third, we propose a methodology that is designed to evaluate the spatial generalizability of the classifiers. We use a spatial  $k$ -fold cross-validation procedure, which enables us to evaluate the performance of the classifiers in a large and geographically heterogeneous context. Finally, we leverage the computational power of GEE and its full Landsat archive to introduce a practicable and adaptable procedure for temporal analysis of urban areas at scale.

To summarize, the objectives of this study are: (1) to present a large-scale dataset for supervised image classification of built-up areas; (2) to integrate this dataset into the GEE platform; and (3) to compare different types of classifiers and inputs in GEE. The dataset can be downloaded as a Google fusion or KML file format [21].

The remainder of this article is organized as follows. In Sub-Section 1, we discuss the literature on urbanization and remote-sensing methods for urban research. In Section 2, we describe the study area and the methodology used to construct and to assess the dataset. In Sections 3 and 4, we present and evaluate the results. In Section 5, we offer a concluding discussion.

### *Measuring Urbanization by Means of Remote Sensing*

Urbanization occurs as rural areas are incorporated into cities, typically through sprawl radiating out from the city center or linearly along major transportation corridors [22,23]. The growth of cities, which often occurs in unplanned and uneven patterns [24], changes the spatial distribution of population sub-groups [25,26], and affects land cover and land use (LC/LU) [10,27] through the construction of built-up structures and impervious surfaces [22,28,29].

Previous literature characterizes urbanization, alternatively, as an increase in the share of the population living in cities, the level of non-agricultural employment or production, the pace of resource consumption, or the presence of traffic congestion [30]. Spatial metrics of urbanization include urban land area, population density, spatial geometry, accessibility, and building types, as well as various features of land use [31]. However, the dichotomy between “urban” and “rural” is not universal [32]. Urban areas are often defined according to social or administrative indicators derived from census-based sources, which, by their nature, vary in their availability, consistency, and spatial and temporal resolutions.

Given the spatial dimensions of urbanization, remote-sensing analysis of satellite images is valuable for mapping urban areas, and analyzing and modeling urban growth and land-use change [33]. Many features associated with urbanization can be detected in satellite images and used to delineate the boundaries of urban areas, including nighttime lights, LC and LU. However, the delineation of urban areas often differs according to the nature of input data [13], which may capture different dimensions of urbanization, such as population distribution, national income levels, or the distribution of physical structures. For example, it is common to see disparities between the extent of lighted areas and other spatial measures of urban extent [34], due in considerable part to the relatively coarse spatial resolution of these datasets [35].

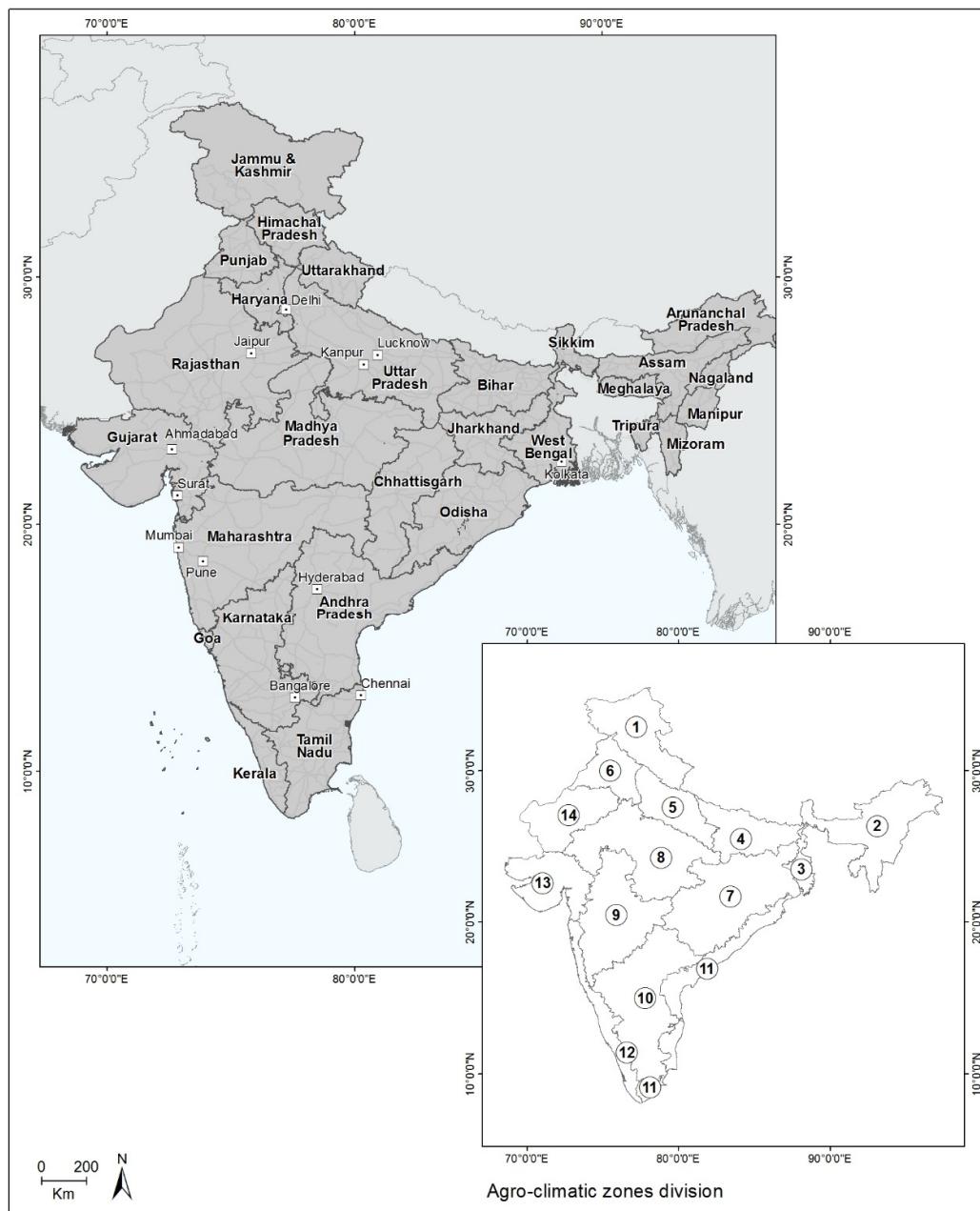
In this paper, we use satellite images to define urban extents according to built-up land cover, which can be observed in satellite images [22,29] and that is closely related to urbanization [22,28]. Detection of LC/LU using remote sensing can be performed at the level of a pixel (pixel-based), or at the level of an object (object-based), where pixels are grouped together to provide contextual information, such as image texture, pixel proximity, and salient geometric attributes of features. While several studies suggest that object-based classifiers outperform pixel-based classifiers in LC/LU classifications tasks [36–39], other studies suggest that pixel-based and object-based classifiers perform similarly when utilizing common machine-learning algorithms [40]. In addition, object-based classification requires significantly more computational power than pixel-based classification and there is no universally accepted method to determine an optimal scale level for image segmentation [37], especially when analyzing large-scale geographically diverse regions. Thus, object-based classification is typically conducted when the unit of analysis is relatively small, such as a city [37,39], or a region of a country [36,38,40,41].

In this study, we adopt a pixel-based classification approach to detect built-up areas in India that utilizes the full spectral imagery available in Landsat, as well as NDVI (Normalized Difference Vegetation Index) and NDBI (Normalized Difference Built-up Index) indices. We apply three types of classifiers that are integrated into GEE: Classification and Regression Tree (CART) [42], Random Forest [43], and Support Vector Machines (SVM) [44].

## 2. Materials and Methods

### 2.1. Study Area

India is one of the largest (3.287 million km<sup>2</sup> in size) and most populated countries in the world. In 2014, 1.295 billion people resided in the nation's 29 states which are distributed across 15 geographical regions (see Figure 1 [45]), 32.4% of whom lived in urban areas [46]. The country is urbanizing rapidly. In the last decade, the growth of its urban population outpaced the growth of its rural population by 31.80% to 12.18% [47], due primarily to natural urban population growth and secondarily to rural-to-urban migration [48]. This trend is expected to continue [49]. By 2050, half of India's population is projected to be urban [3].



**Figure 1.** India's states (**top**); and agro-climatic zones (without zone 15—the islands region) (**bottom**). The spatial extent of the agro-climatic zones was digitized according to the National Portal on Mechanization and Technology [45].

The urbanization of India is also reflected in the rapid expansion of built-up areas [22,50,51] and low-density sprawl [48], together with a decline of other types of land cover, including open land, agriculture land, and bodies of water [52–54]. By capturing the distinct spectral profile of built-up areas, by means of earth observation, it is possible to map and to quantify the extent of urbanization and the pace of urban growth.

India contains 15 distinct agro-climatic zones. These zones are geographical regions characterized by relatively homogenous environmental-physical characteristics, such as soil type, rainfall, temperature, and water resources [55]. India's unusually large number of climatic zones reflects the country's latitudinal expanse and widely varying elevation and rainfall. Previous studies have shown that these zones vary in their agriculture growth, rural poverty and population density [56]. By randomly sampling areas within India for our analysis, the country's geographic diversity allows us to create a training set that would incorporate agro-climatic zones found in the large majority of developing countries. This feature makes our training set of potential value for analysis throughout the tropics, as well as in sub-tropical regions.

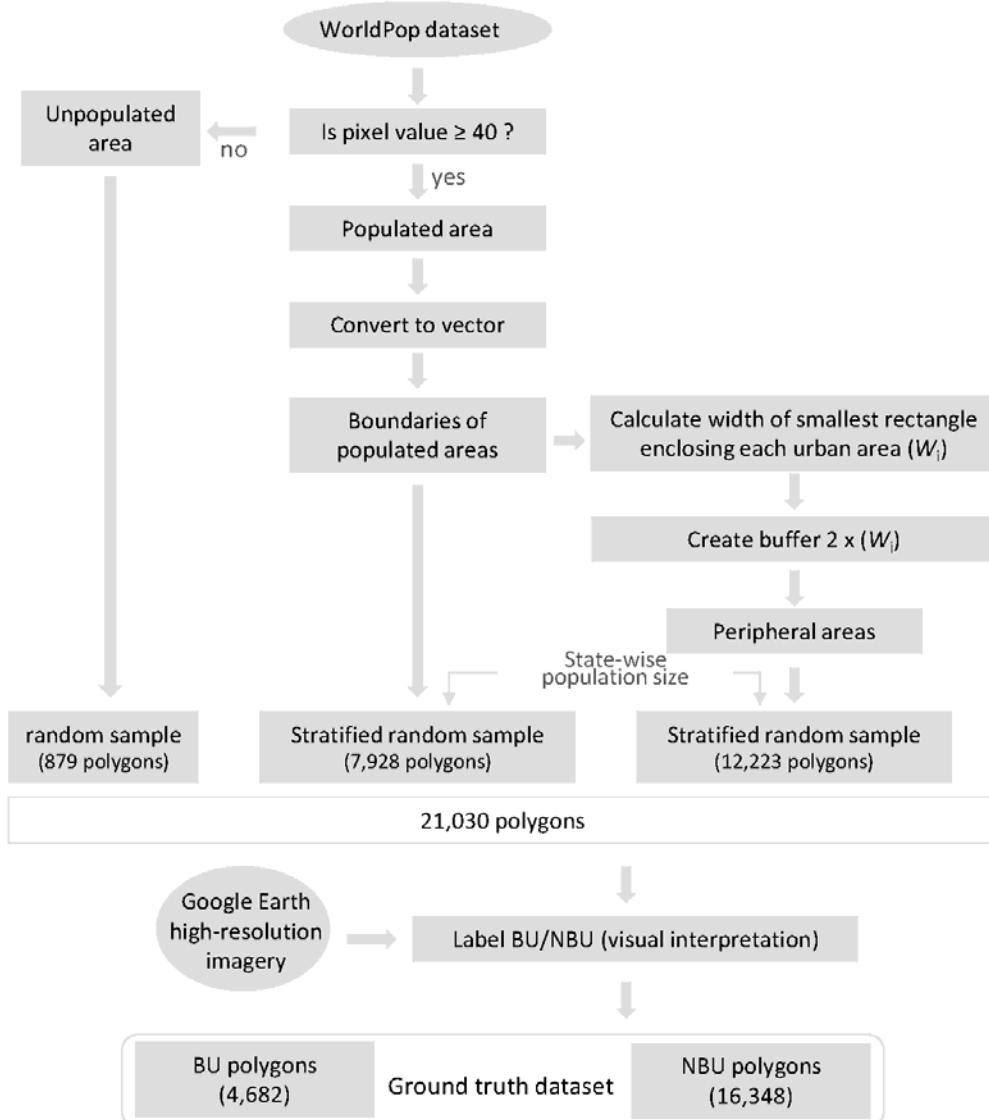
## 2.2. Dataset Construction

We define the boundaries of urban areas according to one property of urbanization: built-up land cover (i.e., the boundaries between built-up (BU) and not built-up (NBU) areas). We define BU areas as polygons where the majority of space (more than 50%) is paved or covered by human-made surfaces and used for residential, industrial, commercial, institutional, transportation, or other non-agricultural purposes. All other land cover is defined as NBU. Similar definitions for urban areas are proposed by [12,13] who characterize a pixel as "urban" when the built environment spans the majority (50% or greater) of the sub-pixel space.

Our classification utilizes a dataset consisting of 21,030 polygons,  $30\text{ m} \times 30\text{ m}$  in size, that are randomly distributed throughout India and manually labeled as BU or as NBU (the methodology is described in Figure 2). To construct this dataset, we begin with WorldPop, a per-pixel population estimation dataset [11,15], and create an initial random stratified sample of BU and NBU areas. WorldPop depicts a grid of per-pixel estimates of population densities, in a spatial resolution of approximately 100 m (we use India's population dataset for 2010, available at: [www.worldpop.org.uk](http://www.worldpop.org.uk)). The maximum value of a pixel is 1523 (i.e., 1523 people per hectare). A visual comparison between WorldPop dataset and Google Earth satellite imagery shows that a threshold of 40 persons per hectare (pixel) closely matches the extent of India's settlements and populated areas. We thus set a threshold of 40 persons per pixel as an initial indicator to identify highly populated areas. These areas constitute 0.41% of the country's land area and account for 19.2% of the country's population.

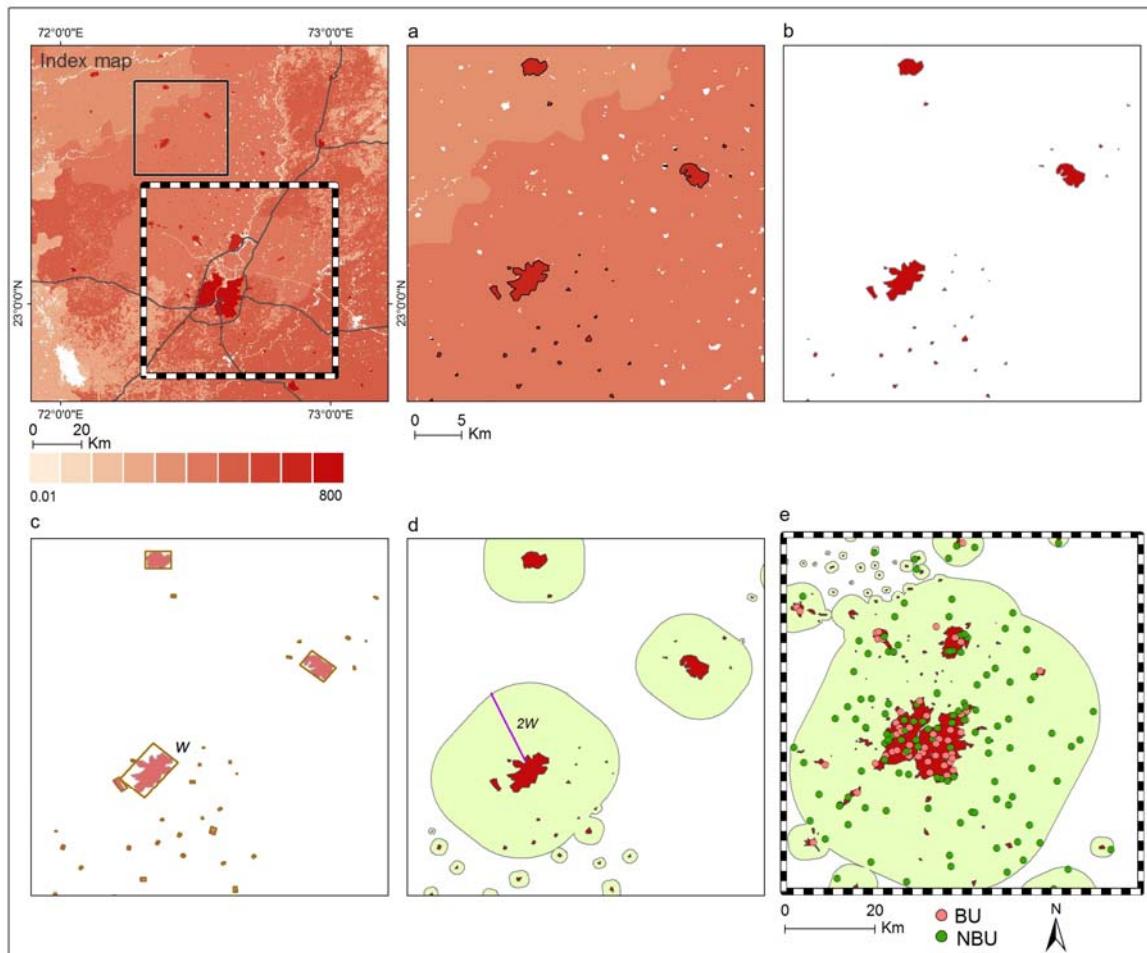
We define populated areas as clusters of neighboring pixels whose values are higher than, or equal to, 40 (i.e., 40 persons per pixel). We convert these clusters to polygons (a vector format), where the polygons represent the boundaries of highly populated areas. We define the adjacent periphery to these highly populated areas by calculating the width of each polygon's enclosing rectangle ( $W_i$ ), where  $W_i$  is the length of the shorter side of a given polygon's enclosing rectangle. To capture peripheral rural areas around cities, we create a buffer around each polygon, which is twice the size of its enclosing rectangle ( $2W_i$ ). We sample our NBU examples from these peripheral areas (Figure 3), such that in the classification we will consider pixels from established urban areas and immediate surrounding areas that have yet to experience urbanization. We focus on rural areas adjacent to urban areas because our BU/NBU classification targets the boundaries of cities and is therefore designed to characterize the process of urban sprawl. From this universe of high-population-density cores and surrounding peripheral rural areas in India, we randomly sample 20,151 polygons, 40% of which are from the core and 60% of which are from the periphery (7928 and 12,223 polygons, respectively), where the number of sampled polygons in each of Indian state is proportional to the state's total population. We oversample polygons from the periphery to account for heterogeneity in the types of land cover found in NBU regions. In order to have sample representation of rural areas that are distant from urban

zones, we randomly sample an additional 879 polygons from outside of the core and periphery areas of cities, for a total of 21,030 polygons.



**Figure 2.** The procedure to generate the ground-truth dataset.

We overlay the polygons with the Google Earth high-resolution base map and manually classify each polygon as BU or as NBU using a visual interpretation method. The polygons are manually labeled by two graduate students who were provided with extensive training and supervised by the researchers. We provided each student an equal proportion of samples. The students labeled each polygon either as BU or as NBU in Google Earth by a visual interpretation of the most recent available satellite image (typically from 2014 to 2015). The students were instructed to label polygons that have at least 50% of their area covered with built-up land cover (according to the definition above) as BU and otherwise as NBU. The manual labeling resulted in a dataset (a KML file) of 4682 polygons that were labeled as BU and 16,348 polygons that were labeled as NBU (some polygons from the urban core did not contain a majority of built-up pixels, leading us to label them as NBU, whereas some polygons from peripheral areas surrounding cities did have a majority of built-up pixels, leading us to label them as BU). The KML file is then converted to a Google fusion table, which is used for supervised classification in GEE.



**Figure 3.** The procedure to generate the stratified random sample. We begin with WorldPop dataset, a grid of per-pixel estimates of population densities (**a**). Then, we extract clusters of neighboring pixels whose values are greater than or equal to 40. These clusters represent highly populated areas and are converted into a vector format (polygons) (**b**). We calculate the width ( $W_i$ ) of the shorter side of each polygon's enclosing rectangle (**c**) and create a buffer around each polygon that is twice this width ( $2W_i$ ) (these buffers represent the periphery of the populated areas) (**d**). Finally, we randomly sample 7928 and 12,223 polygons from the highly populated and from their periphery, respectively (**e**).

### 2.3. Pre-Processing and Scene Selection

We use Landsat 7 and Landsat 8 as inputs for image classification (Table 1 presents a description of the spectral bands). Although the spectral resolution of Landsat 7 is lower than that of Landsat 8, the former satellite was launched in 1999 (Landsat 8 was launched in 2013) and thus allows for a longer time horizon over which to study urbanization. Since a composite of pre-processed scenes of Landsat 7 is available in GEE, we use a *Landsat 7 annual TOA percentile composites* (2014) (referred to as Landsat 7). This composite includes Top of Atmosphere (TOA) calibrated Landsat 7 (ETM+) images (filtered to 2014), excluding images with a negative sun elevation. The composite includes pixels with the lowest cloud cover, computed as per-band percentile values and scaled to 8 bits ([0,255]) (bands 1–5,7) or to units of Kelvin-100 (band 6). For Landsat 8, we apply a standard TOA calibration on *USGS Landsat 8 Raw Scenes* (filtered to 2014) and assign a cloud score to each pixel. We select the lowest possible range of cloud scores and compute per-band percentile values from the accepted pixels. We scale the values to 8 bits.

**Table 1.** The bands that were used as features for the classification.

	<b>Spectral Band</b>	<b>Wavelength (Micrometers)</b>	<b>Resolution (Meters)</b>
Landsat 7			
B1	Band 1—blue-green	0.45–0.52	30
B2	Band 2—green	0.52–0.61	30
B3	Band 3—red	0.63–0.69	30
B4	Band 4—reflected IR	0.76–0.90	30
B5	Band 5—reflected IR	1.55–1.75	30
B6	Band 6—thermal	10.40–12.50	120
B7	Band 7—reflected IR	2.08–2.35	30
NDVI	(B4 – B3)/(B4 + B3)		30
NDBI	(B5 – B4)/(B5 + B4)		30
Landsat 8			
B1	Band 1—Ultra blue	0.43–0.45	30
B2	Band 2—Blue	0.45–0.51	30
B3	Band 3—Green	0.53–0.59	30
B4	Band 4—Red	0.64–0.67	30
B5	Band 5—Near Infrared (NIR)	0.85–0.88	30
B6	Band 6—SWIR 1	1.57–1.65	30
B7	Band 7—SWIR 2	2.11–2.29	30
B8	Band 8—Panchromatic	0.50–0.68	15
B10	Band 10—Thermal Infrared (TIRS) 1	10.60–11.19	100 (resampled to 30)
B11	Band 11—Thermal Infrared (TIRS) 2	11.50–12.51	100 (resampled to 30)

To improve the classification when using Landsat 7 as the input, we add two additional indices: the Normalized Difference Vegetation Index (NDVI) [57] and the Normalized Difference Built-up Index (NDBI) [58].

- NDVI expresses the relation between red visible light (which is typically absorbed by a plant's chlorophyll) and near-infrared wavelength (which is scattered by the leaf's mesophyll structure). It is computed as:

$$(NIR - RED)/(NIR + RED) \quad (1)$$

where NIR is the near infra-red wavelength and RED is the red wavelength. The values of NDVI range between (−1) and (+1). An average NDVI value in 2014 was calculated for each pixel (with *Landsat 7 32-Day NDVI Composite*).

- NDBI expresses the relation between the medium infra-red and the near infra-red wavelengths. It is computed as:

$$(MIR - NIR)/(MIR + NIR) \quad (2)$$

where MIR is the medium infra-red and NIR is the near infra-red wavelength. The index assumes a higher reflectance of built-up areas in the medium infra-red wavelength range than in the near infra-red.

## 2.4. Detection of Built-Up Areas

We perform detection of built-up areas in GEE. First, we overlay the labeled polygons on the input. We collect all Landsat pixels within the regions of these polygons (a total of 5092 BU examples and 17,751 NBU examples), including the reflectance values (per band) and the index values of the examples. Note that the number of the sampled pixels (examples) differs from the number of polygons in the dataset because the polygons do not overlap entirely with Landsat's pixels; these variables are the input for the classifiers (the classifiers' feature space). In addition, each example included an

output: a binary class—BU or NBU. We use this set to train, test and evaluate the performance of the classifiers.

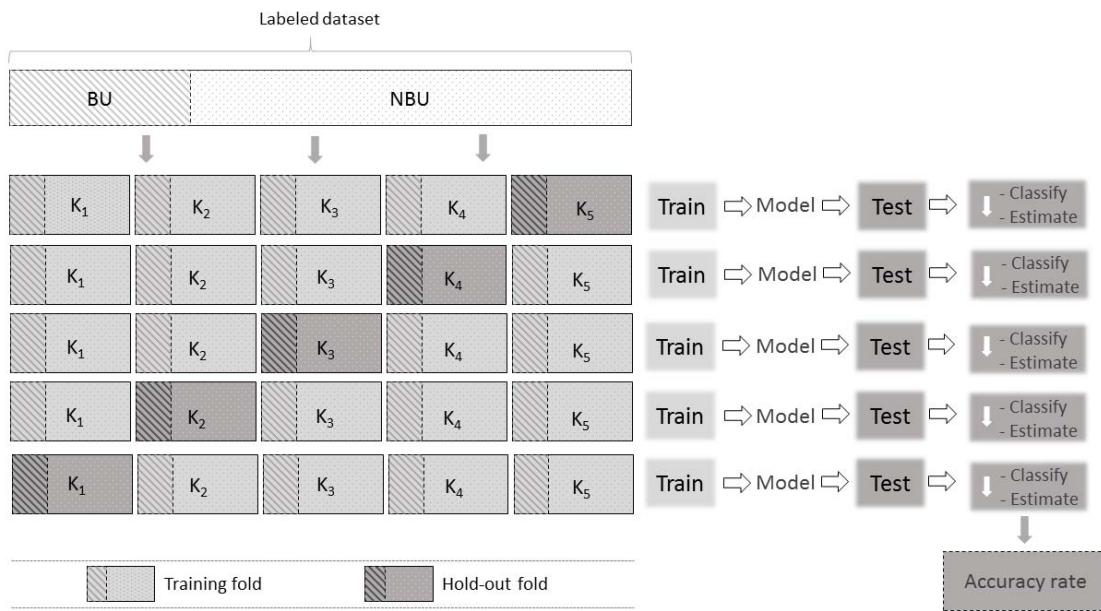
We perform pixel-based classification with three types of classifiers: (i) Classification and Regression Tree (CART)—a binary decision tree classifier; (ii) Support Vector Machines (SVM)—a classifier that identifies decision boundaries which optimally separate between classes (we use a basic linear SVM); and (iii) Random Forests—tree-based classifiers that include  $k$  decision trees ( $k$  predictors). We present a detailed description of the classifiers and their tuning parameters in Appendixes A and B, respectively.

## 2.5. Accuracy Assessment

The performance or the accuracy of a classifier refers to the probability that it will correctly classify a random set of examples [59]. To assure a “fair” assessment of a classifier’s generalization, the data used to train the classifier must be separated from the data that is used to assess its accuracy. Thus, labeled data is typically divided into a *training set* and a *test set* (a *validation set* may also be used to “tune” the classifier’s parameters). Different data splitting heuristics can be used to assure a separation between the training and test sets [59], including the *holdout method*, in which the data is divided into two mutually exclusive subsets: a training set and a test/holdout set; *bootstrapping*, in which the dataset is sampled uniformly from the data, with replacement; and *cross-validation*, also known as  $k$ -fold cross-validation, in which the data are divided into  $k$  subsets (optimally 5 or 10, to allow a less biased estimation [60]) with  $k$  “experiments”. The cross-validation procedure ensures that each example is included exactly once in the test fold and that each example in the test fold is not used to train the classifier. Averaging the overall accuracy across all  $k$  partitions yields  $k$  accuracy values, or  $k$  hold-out estimators, and a variance estimation of the classification error [61,62]. Though each of these methods can be used to assess the performance of a given classifier, cross-validation is a widely accepted procedure [63] that provides a robust estimate of a classifier’s generalization error [64]. When the instances are representative of the underlying population and when sufficient instances are available for training, this procedure results in an unbiased estimate of the accuracy of the classifier over the population [65].

In this study, we adopt a  $k$ -fold cross-validation procedure (with  $k$  “experiments”) to estimate the accuracy of the classifiers. In each experiment, the examples in one of the data folds is left out for testing and the examples in the remaining  $k-1$  fold are used to train the classifier. The performance quality of the trained classifier is tested on the left-out fold, and the overall performance measure is then averaged over the  $k$  folds (over the  $k$  experiments) (Figure 4).

We first conduct a 5-fold cross validation by dividing the data into five randomly stratified folds (while maintaining a constant proportion of BU and NBU examples per fold). Then, to evaluate the spatial generalization of the classifiers, we conduct a 14-fold cross validation by dividing the data into 14 distinct geographical regions according to India’s agro-climatic zones [55] (see Figure 1) (note: we exclude zone number 15, which is the islands region). Each zone includes between 558 and 2695 BU and NBU examples (see Table 2).



**Figure 4.** *k*-fold (5-fold) cross validation scheme. In each “experiment”, the examples in one of the data folds is left out for testing and the remaining examples in the  $k-1$  fold are used to train the classifier. The performance quality of the trained classifier is tested on the left-out fold (in each “experiment”), and the overall performance measure is then averaged over the  $k$  folds ( $k$  “experiments”) (This figure is adapted from [63]).

**Table 2.** Built-up (BU) and not built-up (NBU) examples per agro-climatic zone.

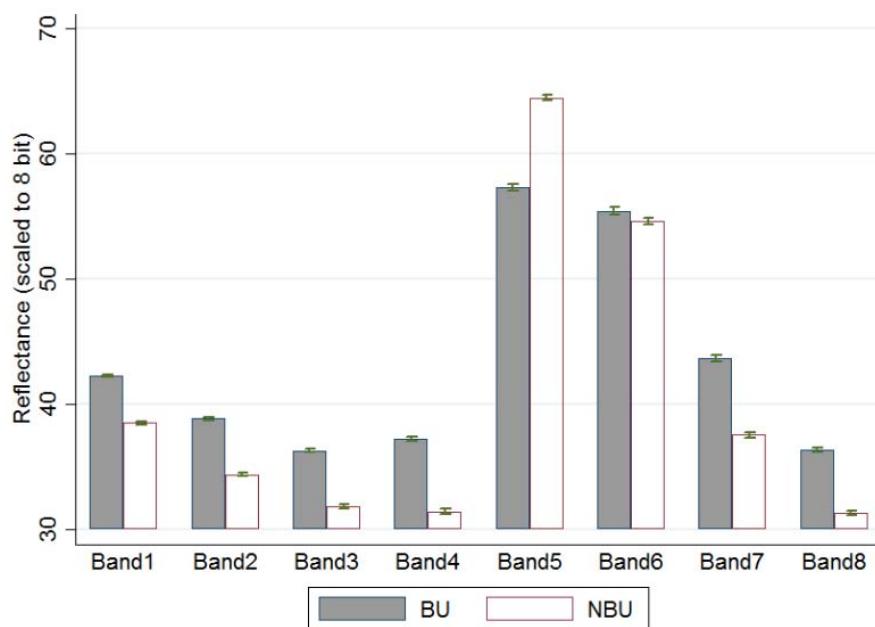
Zone Number	Number of Examples		BU/NBU Ratio	
	BU	NBU	BU	NBU
1	82	476	14.7%	85.3%
2	169	825	17.0%	83.0%
3	222	837	21.0%	79.0%
4	425	1816	19.0%	81.0%
5	671	2024	24.9%	75.1%
6	382	953	28.6%	71.4%
7	326	1545	17.4%	82.6%
8	333	1066	23.8%	76.2%
9	421	1464	22.3%	77.7%
10	645	1979	24.6%	75.4%
11	391	1197	24.6%	75.4%
12	250	894	21.9%	78.1%
13	262	805	24.6%	75.4%
14	103	467	18.1%	81.9%
Total	4682	16,348		

Note: The table indicates the number of built-up (BU) and not built-up (NBU) polygons per agro-climatic zone and the ratio between BU and NBU examples per zone.

### 3. Results

We now turn to describe the dataset and to present an evaluation of the classification of built-up areas in India using the three classifiers and different combinations of training-set examples and inputs. We assess the performance of the classifiers and map the classified built-up areas. As a preliminary step to validate our BU/NBU dataset’s examples, we examine the reflectance profile of the examples, calculated as the average reflectance value of the sampled regions/pixels per band, scaled to 8 bits (Figure 5). Consistent with built-up areas containing structures and impervious surfaces that are reflective relative to vegetation and undeveloped land of non-built-up areas, the reflectance of NBU

regions is lower than the reflectance of BU regions in all bands except band 5 (the near infra-red range). This anomaly in band 5 is likely due to higher reflectance of vegetation land cover in this wavelength range. A *t*-test of equal means and a Kolmogorov–Smirnov test show that BU and NBU regions are characterized by a significantly different ( $p < 0.001$ , for both tests) reflectance values in all bands (Table 3). The BU/NBU distinction is also expressed by significantly different ( $p < 0.001$ , for both tests) NDVI and NDBI values. As seen in Figure 6, the distribution of the NDVI values of NBU regions is to the right of that of BU regions, while the distribution of the NDBI values of NBU regions is a left-skewed and flatter than of BU regions. The standard error bounds of the average reflectance values within BU and NBU regions are relatively small in all bands.

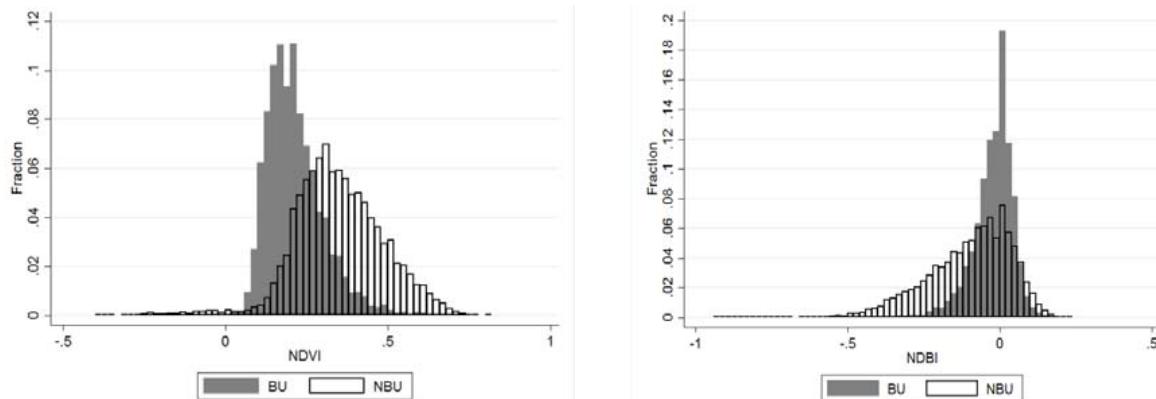


**Figure 5.** The mean and 95% confidence intervals of reflectance values of built-up (BU) and not built-up (NBU) regions (Landsat 8 bands). Note: Per-band percentile values were scaled to 8 bits.

**Table 3.** Average reflectance values of built-up (BU) and not built-up (NBU) regions (Landsat 8 bands). Note: Per-band percentile values were scaled to 8 bits.

		B1	B2	B3	B4	B5	B6	B7	B8	NDVI	NDBI
BU	(mean)	42.27	38.84	36.29	37.20	57.29	55.44	43.64	36.36	0.21	-0.02
	(st. err.)	0.048	0.058	0.073	0.099	0.126	0.151	0.137	0.085	0.001	0.001
NBU	(mean)	38.51	34.37	31.82	31.43	64.49	54.65	37.54	31.27	0.35	-0.10
	(st. err.)	0.069	0.075	0.081	0.097	0.104	0.132	0.120	0.087	0.001	0.001
<i>t</i> -tests of equal means	( <i>t</i> -stats)	44.91	47.17	40.95	41.50	-43.99	3.92	33.54	41.92	-85.98	57.02
	( <i>p</i> -value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
K-S tests of equal dist *	( <i>p</i> -value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: \* Kolmogorov–Smirnov tests of equality of distributions.

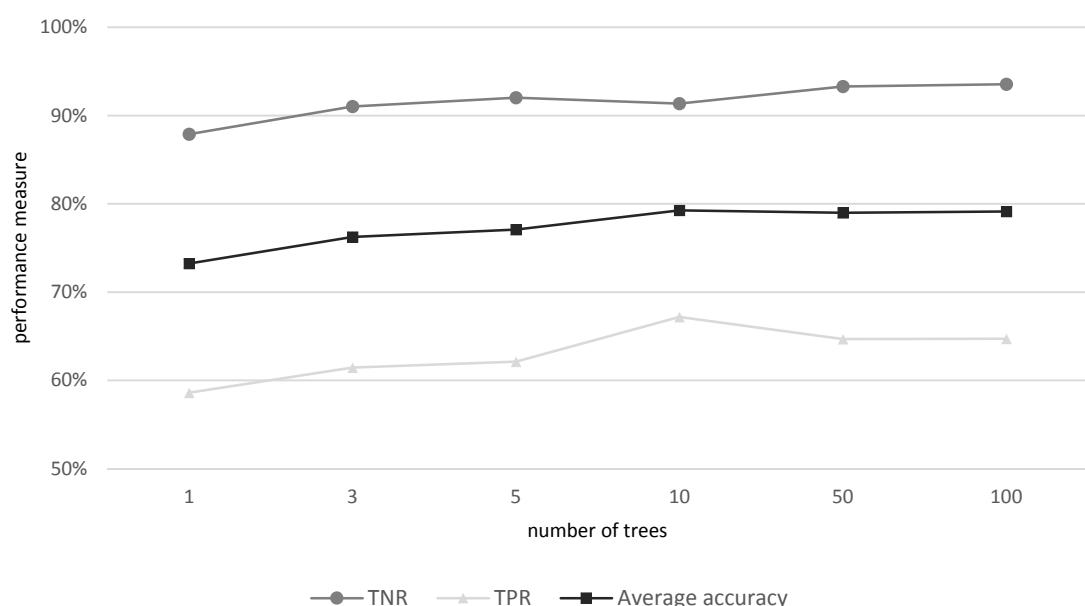


**Figure 6.** The figure shows the histogram of NDVI (Normalized Difference Vegetation Index) and NDBI (Normalized Difference Built-up Index) values by built-up (BU) and not built-up (NBU) examples.

### 3.1. Detection of Built-Up and Not Built-Up Areas

#### 3.1.1. Evaluation of the Classifiers

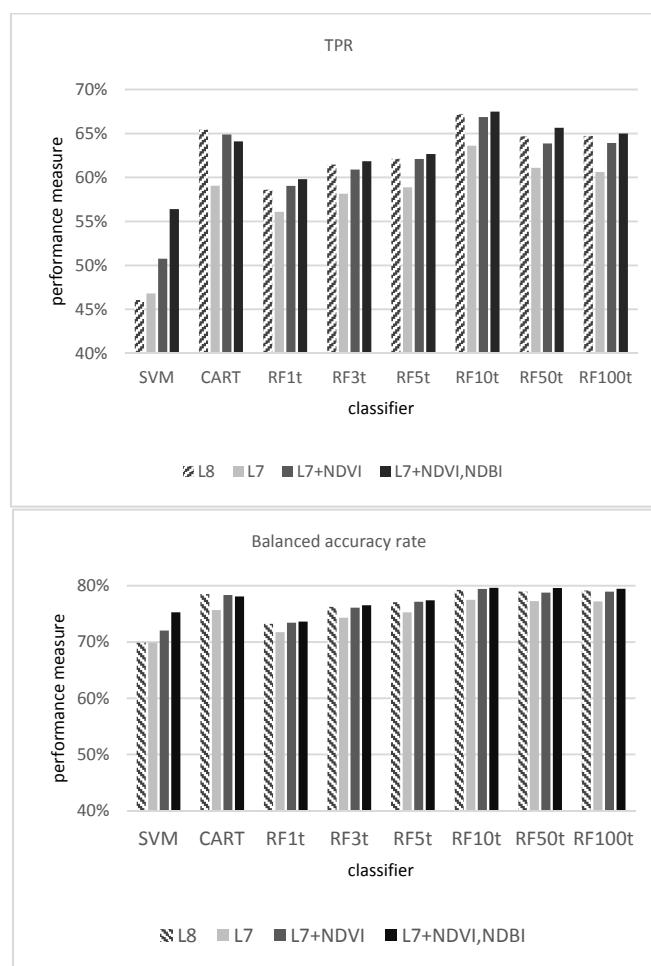
GEE includes several classifiers for pixel-based image classification. In this study we compare the performance of three prominent ones—SVM, CART and Random Forest—in detecting BU and NBU areas in India. A five-fold cross-validation test shows that Random Forest (with 100 decision trees) achieves the highest overall accuracy rate—defined as the percentage of examples that were classified correctly—while SVM achieves the lowest. With Landsat 8 as the input, these two classifiers predict correctly 87.1% and 83.1% of the examples, respectively. Previous studies have suggested that the number of decision trees of the Random Forest is generally proportional to the classifier's accuracy [66]. The results show that though the performance of Random Forest improves as the number of trees increase, this pattern holds only up to 10 trees (see Figure 7). The classifier's performance remains nearly the same with 50 and with 100 decision trees.



**Figure 7.** The effect of the number of trees of Random Forest on the True-positive rate (TPR), True-negative rate (TNR) and the balanced accuracy rate.

We refer to the class “Built Up” (BU) as positive and to the class “Not Built-Up” (NBU) as negative and evaluate the performance of the classifiers using three additional estimators: (1) *True-Positive Rate (TPR)* (the percentage of actual BU examples classified correctly as BU); (2) *True-Negative Rate (TNR)* (the percentage of actual NBU examples classified correctly as NBU); and (3) the average of TPR and TNR (referred to as the balanced accuracy rate).

Random Forest (with 10 decision trees) shows the highest balanced accuracy rate (79.7%) while SVM shows the lowest (around 69%) (see Figure 8). The classifiers’ TPR ranges between 46% (with SVM) and 67% (with Random Forest, 10 trees). As expected, performance with Landsat 8 exceeds that of Landsat 7 likely because of the former’s higher resolution relative to the latter. However, when NDVI and NDBI are added to Landsat 7’s bands, performance using this input improves. With the exception of SVM, Landsat 7 plus NDVI and NDBI as inputs performs similarly to Landsat 8 as the input. As seen in Figure 8, the addition of these two indices primarily improves the balanced accuracy rate of SVM; the classifier’s TPR increases from 47% to 56%, and, accordingly, its balanced accuracy rate increases from 70% to 75%. We relate this to the linear kernel that we use with SVM, which is unable to express nonlinear functions from the input variables to the predicted classes.



**Figure 8.** The True-positive rate (TPR) and the balanced accuracy rates of the examined classifiers: SVM, CART, and Random forest with 1 (RF1t), 3 (RF3t), 5 (RF5t), 10 (RF10t), 50 (RF50t), and 100 (RF100t) trees. Inputs: Landsat 8 (L8), Landsat 7 (L7), Landsat 7 with NDVI (L7 + NDVI) and Landsat 7 with NDVI and NDBI (L7 + NDVI + NDBI).

The performance of the classifiers can also be described in a confusion matrix, where the predicted classes of the examples in the test set are compared with their actual class (resulting in four possible

combinations: TP (True-positive), TN (True-negative), FP (False-positive) and FN (False-negative)). The confusion matrix of the five-fold cross validation (Table 4) describes the predicted and the actual class of the tested examples in the five experiments. As noted above, in each experiment, a different subset (fold) is used for the evaluation, and each example—and all examples—are tested exactly once. Several performance estimators can be calculated from this confusion matrix. We present three that are related to the classification of the positive (BU) class: (1) *Overall accuracy rate*: the portion of instances that were classified correctly (calculated as:  $(TP + TN)/(TP + TN + FP + FN)$ ); (2) *Precision rate*: the portion of instances that were correctly predicted as positive out of all instances that were predicted as positive (calculated as  $TP/(TP + FP)$ ); and (3) *Recall rate*: the portion of instances that were correctly predicted as positive out of all actual positive instances (calculated as  $TP/(TP + FN)$ ). Since the best performance is achieved with Landsat 8 as the input, we use Landsat 8 as the input in subsequent analysis.

**Table 4.** A confusion matrix of the five-fold cross validation tests.

		Predicted									
		L8			L7			L7 + NDVI + NDBI			
		BU	NBU	Total	BU	NBU	Total	BU	NBU	Total	
SVM	BU	2347	2750	5097	2376	2700	5076	2863	2213	5076	
	NBU	1122	16727	17849	1261	16659	17920	1044	16876	17920	
	Total	3469	19477	22946	3637	19359	22996	3907	19089	22996	
		Ac: 0.831 Re: 0.460 Pre: 0.677			Ac: 0.828 Re: 0.468 Pre: 0.653			Ac: 0.858 Re: 0.564 Pre: 0.733			
CART	BU	3335	1762	5097	2998	2078	5076	3253	1823	5076	
	NBU	1500	16349	17849	1380	16540	17920	1413	16507	17920	
	Total	4835	18111	22946	4378	18618	22996	4666	18330	22996	
		Ac: 0.858 Re: 0.654 Pre: 0.690			Ac: 0.850 Re: 0.591 Pre: 0.685			Ac: 0.859 Re: 0.641 Pre: 0.697			
RF3t	BU	3133	1964	5097	2951	2125	5076	3140	1936	5076	
	NBU	1601	16248	17849	1709	16211	17920	1576	16344	17920	
	Total	4734	18212	22946	4660	18336	22996	4716	18280	22996	
		Ac: 0.845 Re: 0.615 Pre: 0.662			Ac: 0.833 Re: 0.581 Pre: 0.633			Ac: 0.847 Re: 0.619 Pre: 0.666			
Actual	BU	3167	1930	5097	2989	2087	5076	3181	1895	5076	
	RF5t	NBU	1423	16426	17849	1494	16426	17920	1402	16518	17920
		Total	4590	18356	22946	4483	18513	22996	4583	18413	22996
			Ac: 0.854 Re: 0.621 Pre: 0.690			Ac: 0.844 Re: 0.589 Pre: 0.667			Ac: 0.857 Re: 0.627 Pre: 0.694		
RF10t	BU	3424	1673	5097	3229	1847	5076	3426	1650	5076	
	NBU	1543	16306	17849	1539	16381	17920	1471	16449	17920	
	Total	4967	17979	22946	4768	18228	22996	4897	18099	22996	
			Ac: 0.860 Re: 0.672 Pre: 0.689			Ac: 0.853 Re: 0.636 Pre: 0.677			Ac: 0.864 Re: 0.675 Pre: 0.700		
RF50t	BU	3297	1800	5097	3102	1974	5076	3332	1744	5076	
	NBU	1196	16653	17849	1180	16740	17920	1153	16767	17920	
	Total	4493	18453	22946	4282	18714	22996	4485	18511	22996	
			Ac: 0.869 Re: 0.647 Pre: 0.734			Ac: 0.863 Re: 0.611 Pre: 0.724			Ac: 0.874 Re: 0.656 Pre: 0.743		
RF100t	BU	3299	1798	5097	3078	1998	5076	3299	1777	5076	
	NBU	1151	16698	17849	1116	16804	17920	1088	16832	17920	
	Total	4450	18496	22946	4194	18802	22996	4387	18609	22996	
			Ac: 0.871 Re: 0.647 Pre: 0.741			Ac: 0.865 Re: 0.606 Pre: 0.734			Ac: 0.875 Re: 0.650 Pre: 0.752		

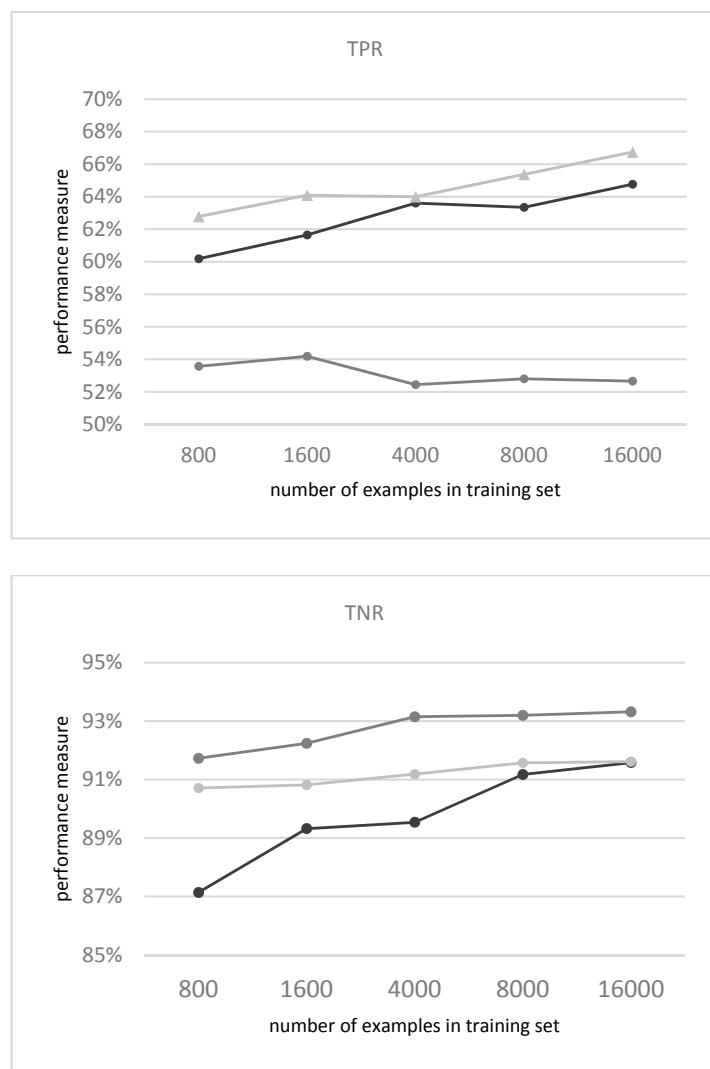
Note: The confusion matrix is calculated for the five experiments. In each experiment a different fold is used as the test fold and each example is tested exactly once. Key: TP: True-positive; TN: True-negative; FP: False-positive; FN: False-negative; Accuracy rate (Ac): the portion of instances that were classified correctly (calculated as:  $(TP + TN)/(TP + TN + FP + FN)$ ); Recall (Re): the portion of instances correctly predicted as positive out of all actual positive instances (calculated as:  $TP/(TP + FN)$ ); Precision rate (Pre): the portion of instances that were correctly predicted as positive out of all instances predicted as positive (calculated as  $TP/(TP + FP)$ ).

We also evaluate the classifiers at the geographical level of agro-climatic zones. A  $k$ -fold cross-validation test was conducted by dividing the examples into 14 folds according to their geographical location (zone). Similar to the results shown above (where the examples were divided into five random folds), Random Forest (with 10 trees) shows the best performance while SVM shows the worst. When Landsat 8 is used as the input, the TPR and the balanced accuracy rate of Random

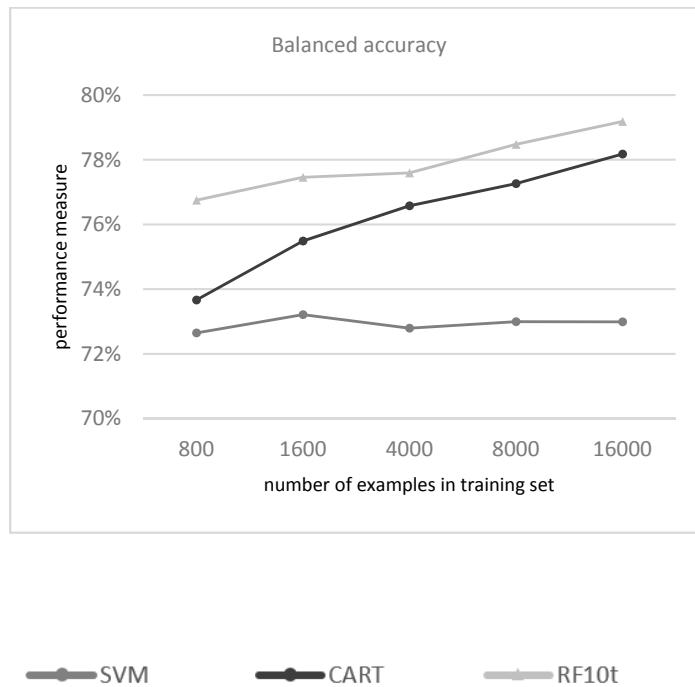
Forest (with 10 trees) are 66% and 78.7%, respectively, while only 54% and 74%, respectively, with CART. This result provides an additional dimension to the assessment of the classifiers' accuracy, and confirms their generalization as predictors under varying geographical conditions.

Our analysis is based on a large training set relative to past work in remote sensing, with over 20,000 hand-labeled polygons. In many settings, constructing a training set of this magnitude may be infeasible. To provide insight into the importance of training-set size for the analysis, we next examine how the prediction rate of the alternative classifiers compares for randomly drawn training sets of different dimensions. We conduct experiments with 800, 1600, 4000, 8000 and 16,000 randomly drawn examples and evaluate each experiment using a five-fold cross-validation test. In each experiment, we use the same test sets (approximately 4500 examples) and a similar proportion between BU and NBU examples (equal to the proportion in the full sample).

The results show strongly improved performance as the size of the training set increases, both, in terms of the TPR and balanced accuracy (see Figure 9). CART shows the largest improvement; for example, as the training set size increases from 800 to 16,000 examples, CART's balanced accuracy increases from 74% to 78%. On the other hand, SVM does not show a significant improvement as the training set size increases; the balanced accuracy remains around 73%.



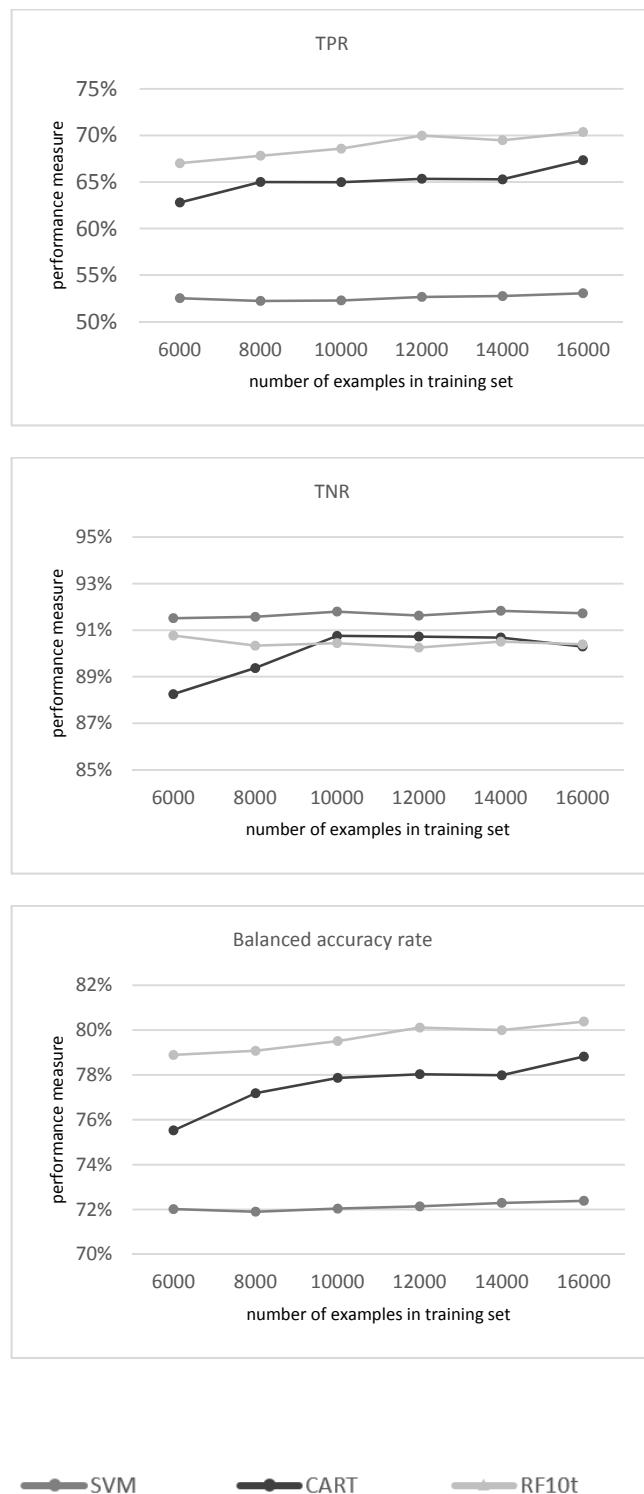
**Figure 9. Cont.**



**Figure 9.** The figure reports the effect of the training set size on the True-positive rate (TPR), True-negative rate (TNR) and balanced accuracy (with Landsat 8 as the input).

In the experiments described above, we maintain a constant proportion between the BU and the NBU examples in the training set (similar to the proportion in the full dataset). In an additional experiment, we examine the effect of varying the proportion between BU and NBU training examples on the classifiers' performance. In each experiment, we use all BU examples in the dataset as training examples and increase the size of the training set by adding NBU training examples. This allows us to evaluate whether performance improves as the size of the training set increases despite an increased imbalance between BU and NBU examples. We conduct a five-fold cross validation test by increasing the number of NBU examples in the training set and maintaining a constant number of BU examples (4000) to a total of 6000, 8000, 10,000, 14,000 and 16,000 BU and NBU examples in the training set. The size of the test set and the proportion between the BU and NBU examples remain constant (Landsat 8 is used as the input).

Results show a moderate improvement in the classifiers' performance as the training set's size increases, primarily with CART. Although the size of the training set is increased only by adding NBU examples, the TPR also improves (in addition to TNR). As the size of the training set increases from 6000 to 16,000 examples, CART's TPR increases from 63% to 67% and its TNR increases from 88% to 90% (see Figure 10). Thus, although we increase the disproportion between BU and NBU examples, the addition of NBU examples improves the overall performance of the classifiers.

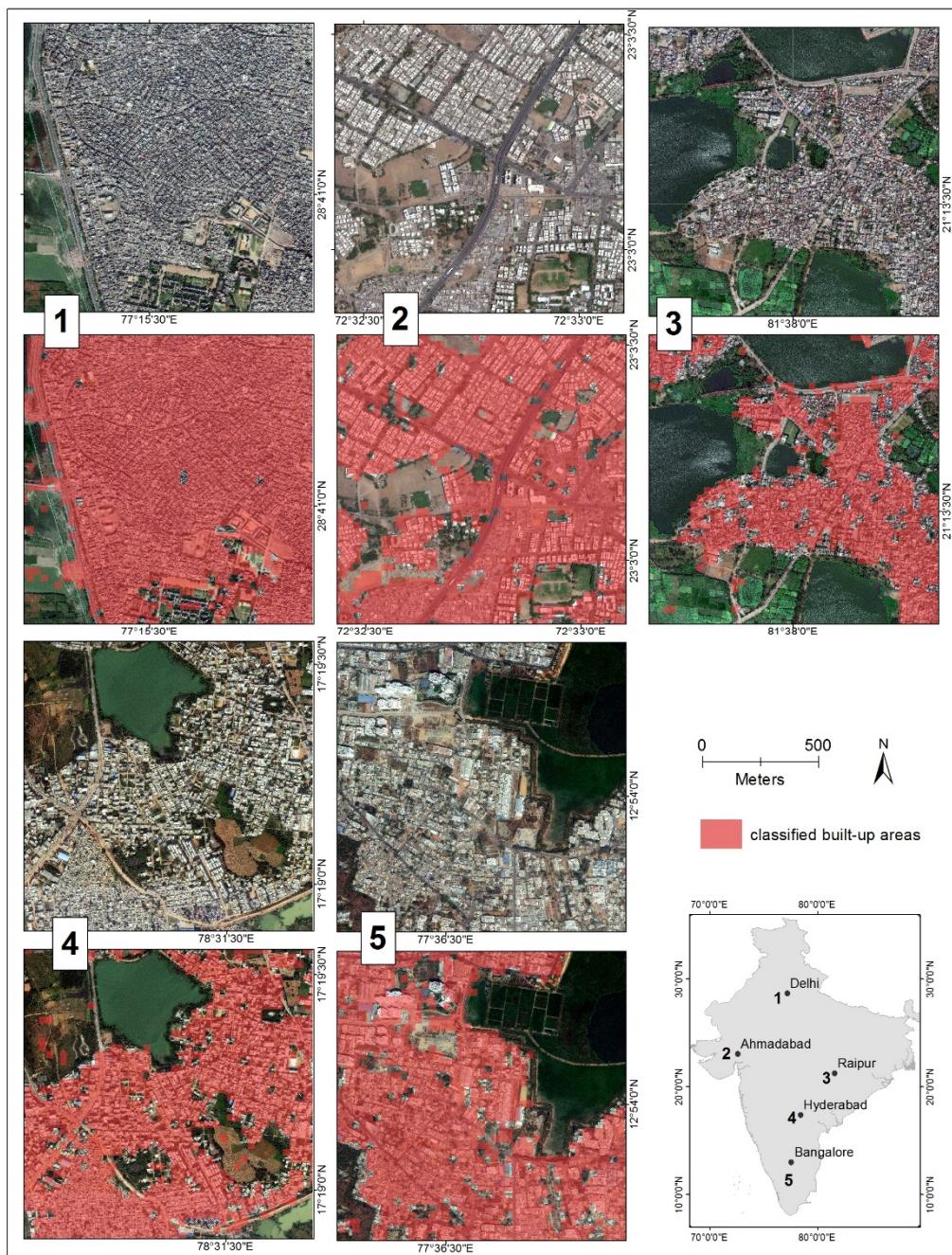


**Figure 10.** The effect of the training set size on the True-positive rate (TPR), True-negative rate (TNR) and the balanced accuracy. The x-axis represents the total number of examples in the training set (each training set includes 4000 BU examples) (with Landsat 8 as the input).

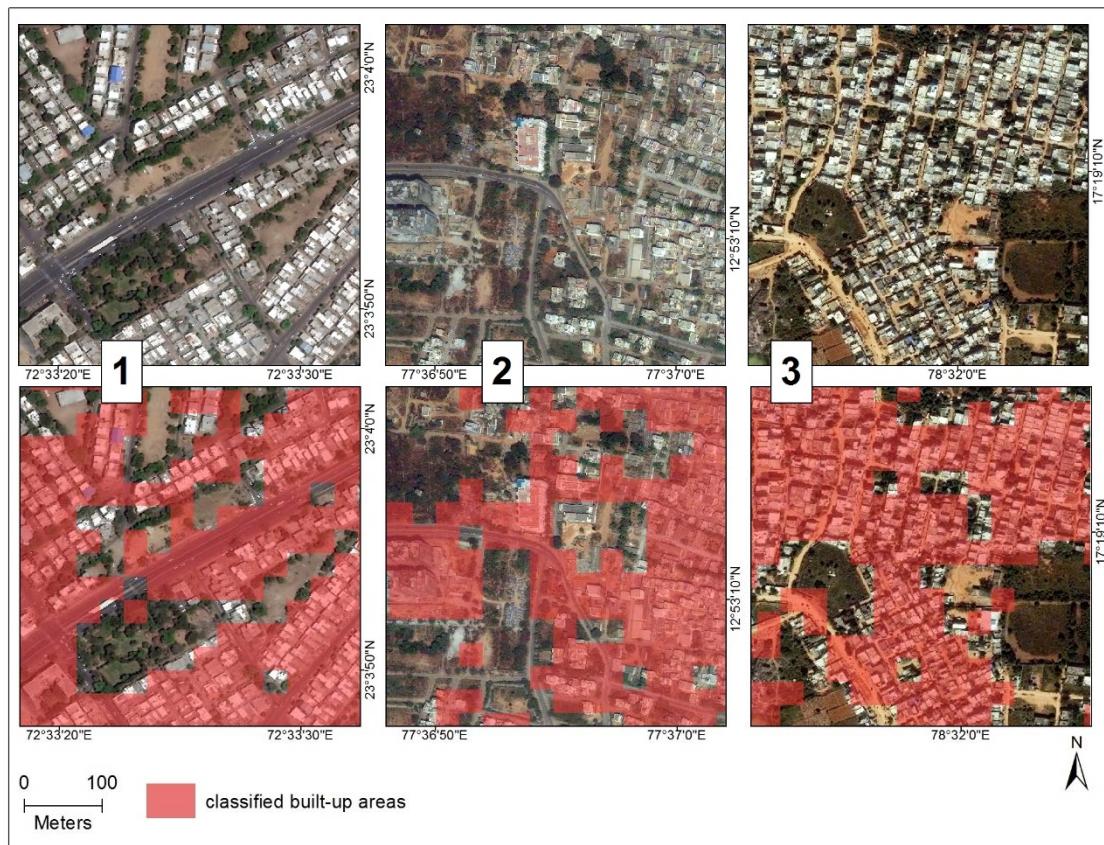
### 3.1.2. Mapping the Classification

In the final classification process, we use the trained classifier to map built-up and not built-up areas over new examples/pixels. The classified image (in a spatial resolution of 30 m) was post-processed to discard isolated pixels and improve the homogeneity of the classified image.

Figure 11 presents, as an illustration, a classified image of built-up areas in five regions across India and Figure 12 presents examples of this classified image in a finer (higher) resolution (we present a classified image of each site below its corresponding high-resolution satellite image). The classified image captures the fabric of built-up urban areas, as well as the fine boundaries between built-up areas and various types of land cover (e.g., vegetation, water bodies and open spaces).

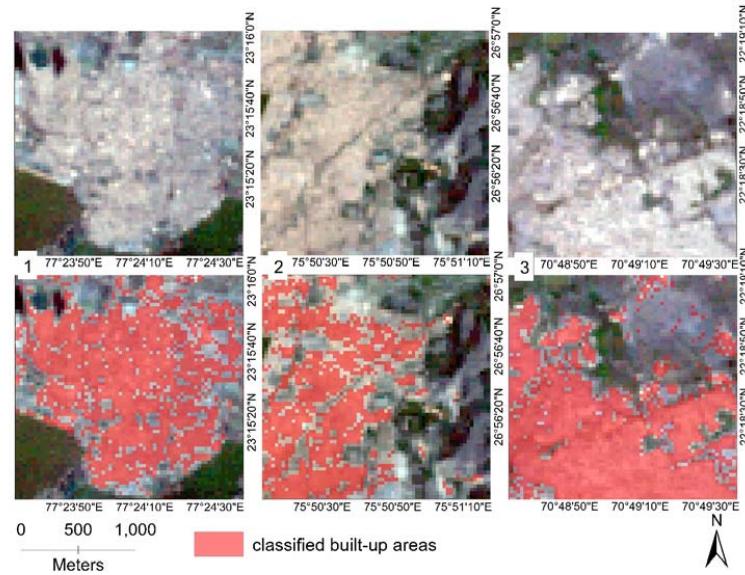


**Figure 11.** Classification of built-up areas (visualized in red) compared to raw satellite images in five regions in India (classifier: Random forest with 10 trees, input: Landsat 8). Satellite images from DigitalGlobe. Includes copyrighted material of DigitalGlobe, Inc. (Westminster, CO, Canada), All Rights Reserved.

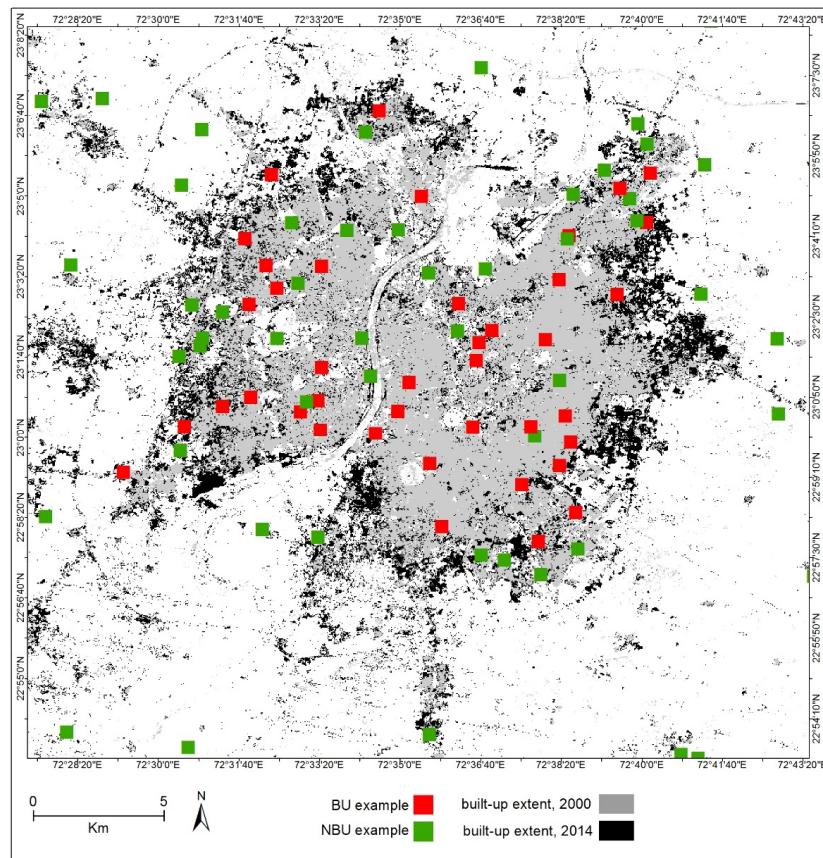


**Figure 12.** A detailed examination of the classification of built-up areas (visualized in red) compared to raw satellite images in five regions in India (classifier: Random forest with 10 trees, input: Landsat 8). Satellite images from DigitalGlobe. Includes copyrighted material of DigitalGlobe, Inc. All Rights Reserved.

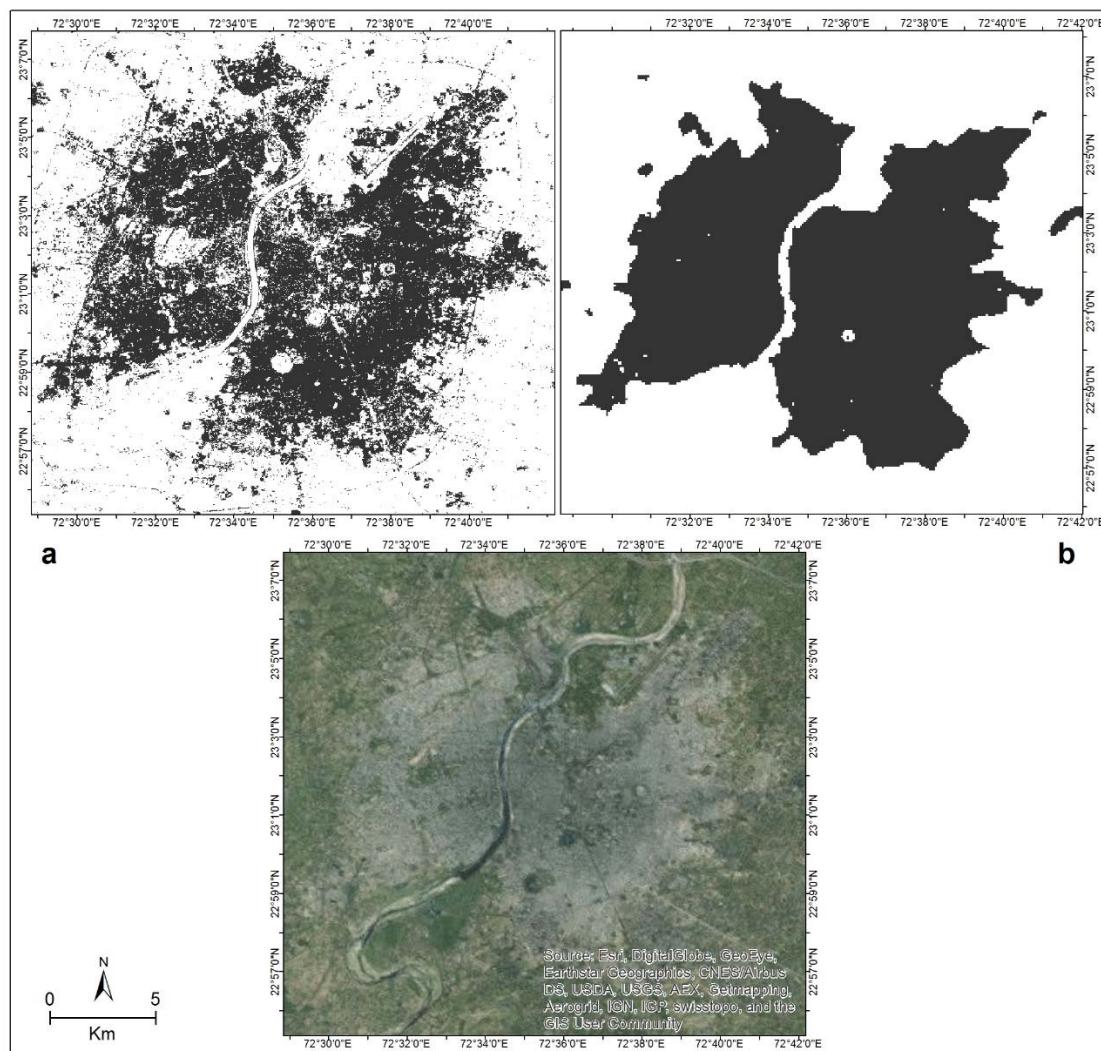
The classifier can also be used to map urban areas across time. Since our ground-truth dataset is collected based on 2014 imagery, using it as a training set with the 2000 imagery may result in “class-noise” [19] due to mislabeled examples. Thus, we first train the classifier (using Random forest, with 10 trees) with Landsat 7, filtered to 2014 as the input (in addition to per-pixel NDVI and NDBI values). Then, we use the trained classifier to map the extent of urban areas in 2000 (using the same feature space, based on Landsat 7, filtered to 2000) (Figure 13 presents examples of the classified image in 2000, and Figure 14 presents a comparison between the extent of urban areas in 2000 and in 2014 in the city of Ahmedabad). As an assessment of this classification, we choose 200 random polygons from our dataset, visually examine them against 2000 Landsat 7 imagery and assign each polygon with a class (BU or NBU). Then, we compare the classified image with this ground-truth dataset. The examination reveals an overall accuracy of 86% and a TPR of 58.6% (Table 5 presents a confusion matrix of this test). Finally, Figure 15 shows the advantage of using our methodology to map urbanization relative to the WorldPop dataset. The classified image is able to capture various types of LC/LU (e.g., built-up areas, parks and open spaces) that is not possible using estimates of local area populations.



**Figure 13.** Detection of built-up areas in three Indian cities in 2000—Erode, Visakhapatnam and Nagpur (bottom)—compared to the raw Landsat 7 filtered to 2000 (top). Classifier: Random Forest (10 trees). Input: Landsat 7 (plus NDVI and NDBI). The classifier was trained with Landsat 7 filtered to 2014, and the trained classifier was used to classify Landsat 7 filtered to 2000.



**Figure 14.** Detection of the boundaries of Ahmedabad, India, in 2000 and in 2014, together with built-up (BU) and not built-up (NBU) examples used for the training. Classifier: Random Forest (10 trees). Input: Landsat 7 (plus NDVI and NDBI). The classifier was trained with Landsat 7 filtered to 2014; the trained classifier was used to classify Landsat 7 filtered to 2000.



**Figure 15.** Estimation of the boundaries of Ahmedabad, India (in 2010) according to: (a) classification of built-up areas; and (b) population density (from WorldPop: [www.worldpop.org.uk](http://www.worldpop.org.uk)). Note: Detection of built-up areas in 2010 was done with Random Forest (10 trees) using Landsat 7 (plus NDVI and NDBI) as the input.

**Table 5.** Confusion matrix describing the classifier's performance (detection of urban areas in 2000).

		Predicted		
		BU	NBU	Total
Actual	BU	34	24	58
	NBU	4	138	142
Total		38	162	200

Note: The classifier is trained with Landsat 7 filtered to 2014. The trained classifier is used to map urban areas with Landsat 7 filtered to 2000 as the input. Classifier: Random forest with 10 trees. Input: Landsat 7 (plus NDVI and NDBI).

#### 4. Discussion

In recent decades, there have been substantial research investments in attempting to understand the social and physical dynamics related to urbanization. Though urbanization is one of the major potential threats to the global environment [22,29], its rate and magnitude have not been quantified

with sharp precision at global scale. For many low-income countries, the last significant mapping efforts occurred in the 1960s and 1970s [67]. Urban extent can be measured by different means, including population counts, nighttime illumination intensity, and detecting the unique LC/LU characteristics and physical attributes associated with urban areas [8–10,13,68]. With the increasing availability of satellite data at ever-improving spatial and temporal resolutions, urban research is rapidly shifting towards the use of image-classification methods designed to extract the “urbanized land” that can be observed and captured in multispectral imagery [13].

Several datasets of urban extent have now been developed to map urban areas at global scale [11,12,15,69]. However, these datasets show considerable disagreement on the location and extent of urban land [13,20] and the majority of the existing information provides classified raster images that have limitations across space and time [10,70–73] or that use ground-truth data that are limited in size, with no more than several thousand examples [20].

With the availability of cloud-based platforms such as GEE, it is now feasible to monitor urbanization in multi-spatial and temporal resolutions and to understand urban dynamics globally. High-resolution ground-truth data are fundamental for any supervised image classification, including classification of built-up land cover. Training data remain scarce, making it difficult to apply modern remote-sensing techniques [74]. At the current time, ground-truth data lag far behind the ever-growing supplies of satellite imagery and analytical tools for image classification. Though ground-truth labeled data for urban areas can be extracted from several existing datasets—e.g., Landsat-based urban maps and crowd-source-based datasets such as OpenStreetMap [16]—validated and processed datasets that are designed specifically for mapping urban areas are in scarce supply. This paper aims to fill this gap.

Ground-truth data can be used, in conjunction with high-resolution satellite imagery and cloud-based computational platforms to detect built-up land cover. GEE is a platform with tremendous potential for urban research at scale. With appropriate ground-truth data, GEE can serve as an accessible and feasible platform for image classification and analysis of large and geographically diverse regions. Though GEE has been used in previous studies for various applications, including population [70,75] and forest cover [76] mapping, ours is the first to provide comprehensive open-source ground-truth data that can serve as a training set for supervised classification of built-up land cover and for evaluation/validation of existing classifiers and classification products.

Of the three types of classifiers that we examine in GEE (SVM, CART and Random Forest), Random Forest achieves the best performance (a balanced accuracy rate of 80%). This classifier produces high-quality maps of built-up areas across space and time in India. Although the performance of CART and Random Forest are better when Landsat 8 is used as the input than when Landsat 7 is used (due perhaps to the higher spectral resolution of Landsat 8), performance improves substantially when NDVI and NDBI are added to Landsat 7, especially with SVM. Similar to the findings of [73,77], performance also improves as the size of the training set increases. Importantly, we find that increasing the size of the training set by expanding the number of NBU examples and holding BU examples fixed leads to marked improvements in accuracy.

We note several limitations of the analysis. First, the dataset was labeled according to 2014–2015 imagery using a visual-interpretation method, which, by its nature, may be subject to idiosyncratic variation across individuals performing the manual classification. As noted in previous studies [19], “class-noise” may impact the accuracy of the classification. Second, our analysis is limited to India. Creating manually labeled ground-truth data is expensive and time consuming. However, crowd-sourcing platforms may allow researchers to scale—at low cost—the labeling method and to construct larger and more comprehensive ground-truth datasets. Although various methods have been suggested for combining census-based data with satellite imagery [11] and to extract training data from different sources, such as from nighttime lights [74], validation by means of visual interpretation remains inescapable for the maintenance of accurate ground-truth training data. Third, the sampling method was designed to detect the boundaries between built-up areas and their periphery; we primarily sampled examples from highly populated areas and from their adjacent,

low-population environs. This approach may create a risk of false-positive detections when classifying distant/remote areas.

## 5. Conclusions

During the past century, many countries, especially in the developing world, have been experiencing rapid urbanization and complex changes in patterns of land cover and land use. Understanding the various ecological, environmental, social and economic impacts of these processes is essential for the preservation of a sustainable human society.

The increasing availability of satellite imagery at different spatial and temporal resolutions has shifted urban research towards the use of digital, multispectral images and the development of remote-sensing image classification methods designed to capture urban land features, such as non-vegetative, human-constructed elements. Though numerous low and medium-scale urban maps have been developed to capture urban land features, these maps are generally limited in their temporal or spatial resolution and cannot be used for analysis of continuous urbanization processes. Moreover, previous studies have generally analyzed urbanization processes over small regions, due in part to computational limitations and the lack of ground-truth data for supervised classification. As parallel computational platforms with much larger storage and capacity become accessible to researchers, it is possible to expand the spatial and temporal units of analysis and to investigate urbanization processes over larger areas and over longer periods of time. Expanding this research frontier creates an urgent need for ground-truth data that can facilitate the development of supervised machine-learning algorithms and enable reliable evaluation and validation.

This paper contributes to this domain by providing ground-truth data that will further efforts to understand the urbanization processes at scale. The dataset we present consists of 21,030 polygons in India that were manually labeled as “built-up” or “not built-up” through a visual interpretation method. Though existing datasets, such as OSM and others, can facilitate supervised classification of urban areas, the majority of these were not developed for this purpose and therefore require further processing and validation. Our large-scale georeferenced dataset was developed to facilitate the detection of urban areas at a national level and to provide a handy and reliable tool for temporal analysis of urban zones and their rural peripheries. Although GEE is steadily evolving as a platform for remote-sensing research at scale, its potential for urban research has not been fully explored. In this study, we highlight the use of GEE for urban research and demonstrate the applicability of our dataset for detection of urban areas in a country with a large population and a diverse land cover. We validate the dataset and show that when used with traditional classifiers available in GEE, the classifiers achieve an overall accuracy rate of around 87%. Our methodology, which is designed to evaluate the spatial generalizability of classifiers, shows that classifier performance is similar when the examples in the training and test sets are sampled from areas with heterogeneous land-cover characteristics. This evaluation procedure is thus suitable for studies that analyze large-scale regions.

Extensions to our approach may improve the classification of urban land cover by modifying the inputs to the classifiers or their dimensions, and by adding additional features to the input's feature space. Incorporating nighttime-light data, socio-economic variables, and physical/geographical characteristics to satellite imagery may offer opportunities to improve the accuracy rate of classifiers. Further extensions to our approach may also include the application of learning algorithms and evaluation with various tuning parameters of the classifiers.

**Author Contributions:** Ran Goldblatt designed the experiment together with Wei You, Gordon Hanson and Amit Khandelwal, implemented the experiment and wrote the manuscript. Wei You helped to carry out and implement the experiment and improve the manuscript. Gordon Hanson and Amit Khandelwal supervised the research, designed and implemented the experiment and improved the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

GEE	Google Earth Engine
BU	built up
NBU	not built up
RF	Random forest
SVM	support vector machines
TPR	True Positive Rate
TNR	True Negative Rate

## Appendix A

Description of the classifiers used in this study:

- *CART* (Classification and Regression Tree) [35] is a binary decision tree. The classifier recursively examines each example's variables with logical *if-then* questions in a binary tree structure. Questions are asked at each node of the tree, and each question typically looks at a single input variable. The variables are compared with a predetermined threshold, so that the examples are optimally split into "purer" subsets [35]. The examples are split to an overly large tree until reaching a terminal node (when the nodes have less than a defined number of examples or when further split will result in almost the same outcome). The tree is then pruned back through the creation of a nested sequence of less complex trees. The class is predicted at the terminal node according to the proportion of the classes in the training examples that reached that node.
- *SVM* (Support Vector Machines) identifies decision boundaries that optimally separate between classes. First, the  $n$  input vectors (examples)  $S = \{X_1, X_2, \dots, X_n\}$  are mapped to the output classes by a linear decision function on a (possibly) high-dimensional feature space  $F = \{\varphi(X_1, X_2, \dots, X_n)\}$ . SVM then optimizes the hyperplane that separates the classes by maximizing the margin between the support vectors of the classes (these are the examples that are closest to the decision surface) [37]. In this study we used a basic linear SVM.
- *Random Forests* are tree-based classifiers that include  $k$  decision trees ( $k$  predictors). When classifying an example, the example variables are run through each of the  $k$  tree predictors, and the  $k$  predictions are averaged to get a less noisy prediction (by voting on the most popular class). The learning process of the forest involves some level of randomness; each tree is trained over an independently random sample of examples from the training set and each node's binary question in a tree is selected from a randomly sampled subset of the input variables [67].

## Appendix B

**Table B1.** The parameters that were used for training (per classifier).

CART	
The cross-validation factor used for pruning	10
Maximal depth level of initial tree	10
Minimal number of training set points in node to allow node creation	1
Minimal number of points at node to allow its further split	1
The minimal cost of training set to allow split	1e-10
Whether to impose stopping criteria while growing the tree	false
The standard error threshold to use in determining the simplest tree whose accuracy is comparable to the minimum cost-complexity tree	0.5

**Table B1.** Cont.

CART	
The quantization resolution for numerical features	100
The margin reserved by quantizer to avoid overload, as a fraction of the range observed in the training data	0.1
The randomization seed	0
SVM	
The decision procedure to use	Voting
The SVM type	C_SVC
The kernel type	Linear
Whether to use shrinking heuristics	True
The cost (C) parameter	1
Random Forest	
The number of Rifle decision trees to create per class	1,3,5,10,50,100
The number of variables per split. If set to 0 (default), defaults to the square root of the number of variables	0
The minimum size of a terminal node	1
The fraction of input to bag per tree	0.5
Whether the classifier should run in out-of-bag mode	false

## References and Note

- Buhaug, H.; Urdal, H. An urbanization bomb? Population growth and social disorder in cities. *Glob. Environ. Chang.* **2013**, *23*, 1–10. [[CrossRef](#)]
- Glaeser, E.L. A world of cities: The causes and consequences of urbanization in poorer countries. *J. Eur. Econ. Assoc.* **2014**, *12*, 1154–1199. [[CrossRef](#)]
- Department of Economic and Social Affairs, Population Division, United Nations. *World Urbanization Prospects: The 2014 Revision*; United Nations: New York, NY, USA, 2015.
- Seto, K.C.; Güneralp, B.; Hutyra, L.R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16083–16088. [[CrossRef](#)] [[PubMed](#)]
- Wu, K.Y.; Ye, X.Y.; Qi, Z.F.; Zhang, H. Impacts of land use/land cover change and socioeconomic development on regional ecosystem services: The case of fast-growing Hangzhou metropolitan area, China. *Cities* **2013**, *31*, 276–284. [[CrossRef](#)]
- McKinney, M.L. Urbanization, Biodiversity, and Conservation: The impacts of urbanization on native species are poorly studied, but educating a highly urbanized human population about these impacts can greatly improve species conservation in all ecosystems. *Bioscience* **2002**, *52*, 883–890. [[CrossRef](#)]
- Pugh, C. *Sustainability the Environment and Urbanisation*; Earthscan: New York, NY, USA, 1996.
- Taubenböck, H.; Esch, T.; Fellbier, A.; Wiesner, M.; Roth, A.; Dech, S. Monitoring urbanization in mega cities from space. *Remote Sens. Environ.* **2012**, *117*, 162–176. [[CrossRef](#)]
- Dewan, A.M.; Yamaguchi, Y. Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Appl. Geogr.* **2009**, *29*, 390–401. [[CrossRef](#)]
- Bhatta, B. Analysis of urban growth pattern using remote sensing and GIS: A case study of Kolkata, India. *Int. J. Remote Sens.* **2009**, *30*, 4733–4746. [[CrossRef](#)]
- Gaughan, A.E.; Stevens, F.R.; Linard, C.; Jia, P.; Tatem, A.J. High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **2013**, *8*, e55882. [[CrossRef](#)] [[PubMed](#)]
- Potere, D.; Schneider, A.; Angel, S.; Civco, D.L. Mapping urban areas on a global scale: Which of the eight maps now available is more accurate? *Int. J. Remote Sens.* **2009**, *30*, 6531–6558. [[CrossRef](#)]

13. Schneider, A.; Friedl, M.A.; Potere, D. Mapping global urban areas using MODIS 500-m data: New methods and datasets based on ‘urban ecoregions’. *Remote Sens. Environ.* **2010**, *114*, 1733–1746. [[CrossRef](#)]
14. Potere, D.; Schneider, A. Comparison of global urban maps. In *Global Mapping of Human Settlement: Experiences, Datasets, and Prospects*; Gamba, P., Herold, M., Eds.; CRC Press, Taylor and Francis Group: Boca Raton, FL, USA, 2009; pp. 269–309.
15. Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE* **2015**, *10*, e0107042. [[CrossRef](#)] [[PubMed](#)]
16. Belgiu, M.; Drăguț, L. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 67–75. [[CrossRef](#)] [[PubMed](#)]
17. Estima, J.; Painho, M. Investigating the potential of OpenStreetMap for land use/land cover production: A case study for Continental Portugal. In *OpenStreetMap in GIScience*; Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 273–293.
18. Schlesinger, J. Using crowd-sourced data to quantify the complex urban fabric—OpenStreetMap and the urban–rural index. In *OpenStreetMap in GIScience*; Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 295–315.
19. Johnson, B.A.; Iizuka, K. Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. *Appl. Geogr.* **2016**, *67*, 140–149. [[CrossRef](#)]
20. Miyazaki, H.; Iwao, K.; Shibasaki, R. Development of a new ground truth database for global urban area mapping from a gazetteer. *Remote Sens.* **2011**, *3*, 1177–1187.
21. The dataset can be accessed online as a Google Fusion Table at: <https://www.google.com/fusiontables/DataSource?docid=1fWY4IyYiV-BA5HsAKi2V9LdoQgsbFtKK2BoQiHb0#rows:id=1> (Note: class “1” = “BU”, class “2” = “NBU”).
22. Sudhira, H.S.; Ramachandra, T.V.; Jagadish, K.S. Urban sprawl: Metrics, dynamics and modelling using GIS. *Int. J. Appl. Earth Obs. Geoinf.* **2004**, *5*, 29–39. [[CrossRef](#)]
23. Baum-Snow, N. Did highways cause suburbanization? *Q. J. Econ.* **2007**, *122*, 775–805. [[CrossRef](#)]
24. Sudhira, H.S.; Ramachandra, T.V. Characterizing urban sprawl from remote sensing data and using landscape metrics. In Proceedings of the 10th International Conference on Computers in Urban Planning and Urban Management, Iguassu Falls, Brazil, 11–13 July 2007.
25. Rahman, A.; Aggarwal, S.P.; Netzband, M.; Fazal, S. Monitoring urban sprawl using remote sensing and GIS techniques of a fast growing urban centre, India. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 56–64. [[CrossRef](#)]
26. Barnes, K.B.; Morgan, J.M., III; Roberge, M.C.; Lowe, S. *Sprawl Development: Its Patterns, Consequences, and Measurement*; Towson University: Towson, MD, USA, 2001; pp. 1–24.
27. Schneider, A. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. *Remote Sens. Environ.* **2012**, *124*, 689–704. [[CrossRef](#)]
28. Bhatta, B.; Saraswati, S.; Bandyopadhyay, D. Urban sprawl measurement from remote sensing data. *Appl. Geogr.* **2010**, *30*, 731–740. [[CrossRef](#)]
29. Jat, M.K.; Garg, P.K.; Khare, D. Monitoring and modelling of urban sprawl using remote sensing and GIS techniques. *Int. J. Appl. Earth Obs. Geoinf.* **2008**, *10*, 26–43. [[CrossRef](#)]
30. Frenkel, A.; Ashkenazi, M. Measuring urban sprawl: How can we deal with it? *Environ. Plan. B Plan. Des.* **2008**, *35*, 56–79. [[CrossRef](#)]
31. Yue, W.; Liu, Y.; Fan, P. Measuring urban sprawl and its drivers in large Chinese cities: The case of Hangzhou. *Land Use Policy* **2013**, *31*, 358–370. [[CrossRef](#)]
32. Dahly, D.L.; Adair, L.S. Quantifying the urban environment: A scale measure of urbanicity outperforms the urban–rural dichotomy. *Soc. Sci. Med.* **2007**, *64*, 1407–1419. [[CrossRef](#)] [[PubMed](#)]
33. Herold, M.; Goldstein, N.C.; Clarke, K.C. The spatiotemporal form of urban growth: Measurement, analysis and modeling. *Remote Sens. Environ.* **2003**, *86*, 286–302. [[CrossRef](#)]
34. Small, C.; Pozzi, F.; Elvidge, C.D. Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sens. Environ.* **2005**, *96*, 277–291. [[CrossRef](#)]

35. Elvidge, C.D.; Safran, J.; Nelson, I.L.; Tuttle, B.T.; Hobson, V.R.; Baugh, K.E.; Dietz, J.; Erwin, W. Area and position accuracy of DMSP nighttime lights data. In *Remote Sensing and GIS Accuracy Assessment*; Lunetta, R.S., Lyon, J.G., Eds.; CRC Press: Boca Raton, FL, USA, 2004; pp. 281–292.
36. Whiteside, T.; Ahmad, W. A comparison of object-oriented and pixel-based classification methods for mapping land cover in northern Australia. In Proceedings of the SSC2005 Spatial Intelligence, Innovation and Praxis: The National Biennial Conference of the Spatial Sciences Institute, Melbourne, Australia, 12–16 September 2005; pp. 1225–1231.
37. Myint, S.W.; Gober, P.; Brazel, A.; Grossman-Clarke, S.; Weng, Q. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sens. Environ.* **2011**, *115*, 1145–1161. [[CrossRef](#)]
38. Whiteside, T.G.; Boggs, G.S.; Maier, S.W. Comparing object-based and pixel-based classifications for mapping savannas. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 884–893. [[CrossRef](#)]
39. Bhaskaran, S.; Paramananda, S.; Ramnarayan, M. Per-pixel and object-oriented classification methods for mapping urban features using IKONOS satellite data. *Appl. Geogr.* **2010**, *30*, 650–665. [[CrossRef](#)]
40. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272. [[CrossRef](#)]
41. Robertson, L.D.; King, D.J. Comparison of pixel-and object-based classification in land cover change mapping. *Int. J. Remote Sens.* **2011**, *32*, 1505–1529. [[CrossRef](#)]
42. Breiman, L.; Friedman, J.H.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, 1984.
43. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
44. Cortes, C.; Vapnik, V. Support vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
45. IASRI. *Study Relating to Formulating Long Term Mechanization Strategy for Each Agro-Climatic Zone/State in India*; Final report; Indian Agriculture Statistics Research Institute (IASRI): New Delhi, India, 2006.
46. World Bank. Available online: <http://data.worldbank.org/country/india> (accessed on 20 March 2016).
47. Census of India. Office of Registrar General & Census commissioner, Ministry of Home Affairs, Government of India. 2011. Available online: <http://censusindia.gov.in/> (accessed on 15 May 2016).
48. Chen, M.; Raveendran, G. *Urban India 2011: Evidence*; Indian Institute for Human Settlements Publications: Bangalore, India, 2011.
49. Sudhira, H.S.; Gururaja, K.V. Population crunch in India: Is it urban or still rural? *Curr. Sci.* **2012**, *103*, 37–40.
50. Prakasam, C. Land use and land cover change detection through remote sensing approach: A case study of Kodaikanal taluk, Tamil Nadu. *Int. J. Geomat. Geosci.* **2010**, *1*, 150–158.
51. Moghadam, H.S.; Helbich, M. Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model. *Appl. Geogr.* **2013**, *40*, 140–149. [[CrossRef](#)]
52. Ramachandra, T.V.; Aithal, B.H.; Sanna, D.D. Insights to urban dynamics through landscape spatial pattern analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 329–343.
53. Chadchan, J.; Shankar, R. An analysis of urban growth trends in the post-economic reforms period in India. *Int. J. Sustain. Built Environ.* **2012**, *1*, 36–49. [[CrossRef](#)]
54. Sharma, R.; Joshi, P.K. Monitoring urban landscape dynamics over Delhi (India) using remote sensing (1998–2011) inputs. *J. Indian Soc. Remote Sens.* **2013**, *41*, 641–650. [[CrossRef](#)]
55. Singh, P. *Agro-Climatic Zonal Planning Including Agriculture Development in North Eastern India, for XI FIVE YEAR PLAN (2007–12)*; Planning Commission, Government of India: New Delhi, India, 2006.
56. Palmer-Jones, R.; Sen, K. What has luck got to do with it? A regional analysis of poverty and agricultural growth in rural India. *J. Dev. Stud.* **2003**, *40*, 1–31. [[CrossRef](#)]
57. Pettorelli, N.; Vik, J.O.; Mysterud, A.; Gaillard, J.-M.; Tucker, C.J.; Stenseth, N.C. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends Ecol. Evol.* **2005**, *20*, 503–510. [[CrossRef](#)] [[PubMed](#)]
58. Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **2003**, *24*, 583–594. [[CrossRef](#)]
59. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Montreal, QC, Canada, 20–25 August 1995; Volume 2, pp. 1137–1143.

60. Rodriguez, J.D.; Pérez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [[CrossRef](#)] [[PubMed](#)]
61. Salzberg, S.L. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Min. Knowl. Discov.* **1997**, *1*, 317–328. [[CrossRef](#)]
62. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [[CrossRef](#)]
63. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-validation. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer US: New York, NY, USA, 2009; pp. 532–538.
64. Blum, A.; Kalai, A.; Langford, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In Proceedings of the 12th Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 7–9 July 1999; pp. 203–208.
65. Bradford, J.P.; Brodley, C.E. The effect of instance-space partition on significance. *Mach. Learn.* **2001**, *42*, 269–286. [[CrossRef](#)]
66. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sánchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [[CrossRef](#)]
67. Tatem, A.J.; Noor, A.M.; Von Hagen, C.; Di Gregorio, A.; Hay, S.I. High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLoS ONE* **2007**, *2*, e1298. [[CrossRef](#)] [[PubMed](#)]
68. Orenstein, D.; Bradley, B.; Albert, J.; Mustard, J.; Hamburg, S. How much is built? Quantifying and interpreting patterns of built space from different data sources. *Int. J. Remote Sens.* **2011**, *32*, 2621–2644. [[CrossRef](#)]
69. CIESIN, Columbia University. Gridded Population of the World, Version 3 (GPWv3) Data Collection. Available online: <http://sedac.ciesin.columbia.edu/data/collection/gpw-v3> (accessed on 2 April 2016).
70. Patel, N.N.; Angiuli, E.; Gamba, P.; Gaughan, A.; Lisini, G.; Stevens, F.R.; Tatem, A.J.; Trianni, G. Multitemporal settlement and population mapping from Landsat using Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *35*, 199–208. [[CrossRef](#)]
71. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [[CrossRef](#)]
72. Shao, Y.; Lunetta, R.S. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS J. Photogramm. Remote Sens.* **2012**, *70*, 78–87. [[CrossRef](#)]
73. Wieland, M.; Pittore, M. Performance evaluation of machine learning algorithms for urban pattern recognition from multi-spectral satellite images. *Remote Sens.* **2014**, *6*, 2912–2939. [[CrossRef](#)]
74. Xie, M.; Jean, N.; Burke, M.; Lobell, D.; Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv Preprint*, 2015. arXiv:1510.00098.
75. Trianni, G.; Lisini, G.; Angiuli, E.; Moreno, E.A.; Dondi, P.; Gaggia, A.; Gamba, P. Scaling up to national/regional urban extent mapping using Landsat data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3710–3719. [[CrossRef](#)]
76. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; Kommareddy, A. High-resolution global maps of 21st-century forest cover change. *Science* **2013**, *342*, 850–853. [[CrossRef](#)] [[PubMed](#)]
77. Qian, Y.; Zhou, W.; Yan, J.; Li, W.; Han, L. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sens.* **2015**, *7*, 153–168. [[CrossRef](#)]

