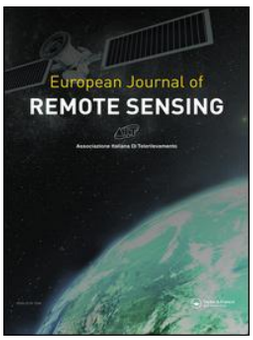

European Journal of Remote Sensing

ISSN: (Print) 2279-7254 (Online) Journal homepage: http://www.tandfonline.com/loi/tejr20

Image based classification of slums, built-up and non-built-up areas in Kalyan and Bangalore, India

Elena Ranguelova, Berend Weel, Debraj Roy, Monika Kuffer, Karin Pfeffer & Michael Lees

To cite this article: Elena Ranguelova, Berend Weel, Debraj Roy, Monika Kuffer, Karin Pfeffer & Michael Lees (2018): Image based classification of slums, built-up and non-built-up areas in Kalyan and Bangalore, India, European Journal of Remote Sensing, DOI: 10.1080/22797254.2018.1535838

To link to this article: https://doi.org/10.1080/22797254.2018.1535838

Published online: 03 Nov 2018.

EUROPEAN JOURNAL OF REMOTE SENSING
https://doi.org/10.1080/22797254.2018.1535838

Taylor & Francis
Taylor & Francis Group

Image based classification of slums, built-up and non-built-up areas in Kalyan and Bangalore, India

Elena Ranguelova [a], Berend Weel [a], Debraj Roy [b], Monika Kuffer [c], Karin Pfeffer [c,d] and Michael Lees [b,e]

[a]Netherlands eScience Center, Amsterdam, The Netherlands; [b]Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands; [c]Faculty ITC, University of Twente, Enschede, The Netherlands; [d]Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands; [e]Design and Urban Studies, ITMO University, St. Petersburg, Russia

ABSTRACT

Slums, characterized by sub-standard housing conditions, are a common in fast growing Asian cities. However, reliable and up-to-date information on their locations and development dynamics is scarce. Despite numerous studies, the task of delineating slum areas remains a challenge and no general agreement exists about the most suitable method for detecting or assessing detection performance. In this paper, standard computer vision methods – Bag of Visual Words framework and Speeded-Up Robust Features have been applied for image-based classification of slum and non-slum areas in Kalyan and Bangalore, India, using very high resolution RGB images. To delineate slum areas, image segmentation is performed as pixel-level classification for three classes: *Slums, Built-up* and *Non-Built-up*. For each of the three classes, image tiles were randomly selected using ground truth observations. A multi-class support vector machine classifier has been trained on 80% of the tiles and the remaining 20% were used for testing. The final image segmentation has been obtained by classification of every 10th pixel followed by a majority filtering assigning classes to all remaining pixels. The results demonstrate the ability of the method to map slums with very different visual characteristics in two very different Indian cities.

ARTICLE HISTORY
Received 12 January 2018
Revised 5 October 2018
Accepted 10 October 2018

KEYWORDS
Image segmentation; informal settlements; support vector machines; bag of visual words; speeded-up robust features

Introduction

Currently, about one-third of the urban population in Asia, home to half of the world's urban population, resides in deprived habitats – also referred to as informal settlements or slums (UN Habitat, 2016). With continued high urbanization rates in Asia (UN Habitat, 2016) and the low capacity of formal affordable housing, much of this growth will happen in slum areas. The Sustainable Development Goals of the United Nations Development Program monitors the proportion of urban population living in slums (or informal settlements) as a key indicator in assessing the outcome of goal 11 which relates to safe, resilient and sustainable cities. Global statistics on this indicator show a relative decrease, but in absolute terms an increase of slum inhabitants. The globally recognized definition of slums by UN-Habitat (UN Habitat, 2016) defines slums as being deprived from access to improved water, sanitation, lacking sufficient living area, durable housing and security of tenure. India adapted this slum definition and defines slums similarly, as areas lacking basic services, substandard, illegal or inadequate housing, overcrowding and high density, unhealthy living conditions and hazardous locations, tenure insecurity, poverty and social exclusion. However, India also uses a minimum size criterion of at least 300 inhabitants or 60 households.[1] Therefore, small and scattered slum pockets (often found at the outskirts of cities but also at central location) are often not included into official slum statistics. Such areas can be also of very temporary nature and transform quickly, for example, settlements of migrants working for the booming Indian construction industry. Further, in India, slum definitions vary across cities and depend on subjective parameters such as narrowness, decay, overpopulation, faulty design, lack of ventilation, lack of sanitation facilities and so on. In such a scenario, the separating line between these subjective parameters, such as "narrow" and "non-narrow" will be drawn differently by different agencies leading to different estimates.

In India, according to UN-Habitat statistics, the total slum population declined from 55% in 1990 to 24% in 2014.[2] Official slum maps often cover only the notified and recognized slums, while excluding a large amount of slum areas (Government of India, 2011). In general, the local policy of slum declarations

CONTACT Elena Ranguelova E.Ranguelova@esciencecenter.nl Netherlands eScience Center, Science park 140, 1098 XG Amsterdam, The Netherlands

[1]http://nbo.nic.in/Images/PDF/SLUMS_IN_INDIA_Slum_Compendium_2015_English.pdf.
[2]https://data.worldbank.org/indicator/EN.POP.SLUM.UR.ZS.

impacts the gap between notified and recognized slums and areas/pockets with slum-like conditions (areas that are not officially defined as slum but lack basic services as defined by UN-Habitat[3]) that exist on the ground. Therefore, the variations within data collection methods and political and operational decisions of including slum areas into these official statistics result in consistency problems. These problems exist within a city, country and even on a global scale. A possible alternative to address such consistency problems is offered by the increasing availability of very high resolution (VHR) remote sensing imagery that allows for a global coverage, and Earth Observation (EO)-based methods (Kuffer, Pfeffer, & Sliuzas, 2016) that can provide a consistent mapping approach of land cover/use. As these methods offer high frequency updates, they can potentially support the production of up-to-date urban base maps and provide possibilities to acquire information on the location, morphology and dynamics of slums (Kuffer et al., 2016) and to model and predict the emergence of slums (Roy, Lees, Palavalli, Pfeffer, & Sloot, 2014), (Roy, Lees, Pfeffer, & Sloot, 2017).

Recent studies have employed texture and object-based image analysis (Graesser et al., 2012; Kohli, Sliuzas, Kerle, & Stein, 2012) coupled with machine learning techniques to classify slums (Duque, Patino, & Betancourt, 2017a; Wurm, Taubenböck, Weigand, & Schmitt, 2018; Kit, Ldeke, & Reckien, 2012; Kuffer, Pfeffer, Sliuzas, Baud, & van Maarseveen, 2017; Engstrom, Newhouse, Haldavanekar, Copenhaver, & Hersh, 2017). Kuffer et al. examined and compared various spectral, textural and spatial feature sets to obtain insight into the most significant slum indicators. Despite numerous studies proposing image-based methods for mapping general urban structure types and specifically slums (Wurm, Taubenböck, Weigand, & Schmitt, 2017), the task of delineating slum areas remains challenging and currently there is no general agreement about the most suitable method (Kuffer et al., 2016). The main uncertainties and challenges inherent in EO-based slum mapping approaches refer to (a) the diversity of spatial, spectral and textural characteristics of slums within and across cities (Kuffer et al., 2017); (b) the required variations in feature sets and methods to cover this diversity (Duque et al., 2017a); (c) transferability problems arising from variations in feature sets and methods (Kohli, Stein, & Sliuzas, 2016); (d) access to commercial and often expensive VHR imagery (Duque et al., 2017a; Wurm, Weigand, Schmitt, Gei, & Taubenböck, 2017); (e) sensor differences particularly in the context of change detection (Ranguelova, Kuffer, Pfeffer, Roy, & Lees, 2017), (f) scalability and computational issues when aiming at a city, urban region, national or global coverage (Kuffer, van Maarseveen, Sliuzas, & Pfeffer, 2017; Ranguelova et al., 2017) and (g) uncertainties in generation and usage of reference ground truth data to assess the quality of slum mapping outputs (Pratomo, Kuffer, Martinez, & Kohli, 2016; Kohli et al., 2016). Furthermore, employing the globally recognized definition of UN-Habitat for slum mapping with EO-based methods has to solve the dilemma that many of the indicators cannot be directly observed in an image (e.g. access to improved water or tenure security). Thus EO-based methods build on the knowledge that slums share specific morphological features that can be recognized in an image, that is, high built-up densities, irregularity of settlement patterns, relatively small building footprint areas and often specific location characteristics (e.g. hazardous areas). Therefore, Wurm et al. (Wurm & Taubenböck, 2018) refer to morphological slums that allow their mapping in images employing morphological characteristics (e.g. building density and size, patterns, and building materials). Though satellite imagery has been available for many years, its application has been limited due to costs and quality issues, specifically for analysing urban land use in developing countries. The advent of new and open source mapping technologies, such as Google Earth (GE), which offers free satellite imagery of most of the Earth's land surface, has led to increased acceptance of such technology for urban land use (Chang et al., 2009; Gunter, 2009; Taylor & Lovell, 2012). The quality and resolution of the maps offered on the mapping platforms by researcher vary greatly. A recent study (Duque et al., 2017a) demonstrated the scope and limitations of applying standard machine learning methods on GE imagery in the context of several South American cities. The approach faced consistency problems due to sensor variations that resulted in different illumination conditions and colour intensity of the GE imageries.

Nevertheless, in our study, the aim is to contribute another methodological approach using such remote sensing images. Specifically, in this article, the aim is to explore whether generic methods developed in computer vision have the ability to detect slum areas when applied on GE RGB images of different resolutions from cities with different morphologies. In particular, we analyse whether these methods offer a robust and computationally feasible approach for mapping slums at a city scale. The methods in question are a multi-class support vector machine (SVM) trained on histograms of visual words (VW) to classify small image windows (tiles). VW represent characteristic parts of an image and are used to compactly represent and index an image or collection of images. Usually the VW are used in the bag of visual words (BoVW) framework (Li & Perona, 2005). First, low-level Speeded-Up Robust Features (SURF) (Bay, Ess, Tuytelaars, & van Gool, 2008) are obtained from the imagery, then they are clustered and the centroids of these clusters are used as VW. Histograms of these

[3]https://data.worldbank.org/indicator/EN.POP.SLUM.UR.ZS.

words per class are then generated for each image (tile), serving as final classification features. These high-level features are used as input to an SVM to classify images into Slum, Built-up and Non-Built-up classes. This classification performed on pixel level results in the delineation of Slums, Built-up and Non-Built-up areas. While many publications on slum detection from satellite imagery have used SVM as a classifier (Kuffer et al., 2016), and BoVW have been used for land use classification (Chen and Tian (2015)) and for scene change detection (Du, Zhang, & Zhang, 2016) and classification (Jun, Y., Jiang, Y.-G., Hauptmann, A.G., & Ngo, C.-W.(2007).), to the best of our knowledge there are no other studies applying the combination for slum detection on easily obtainable RGB GE images (Mahabir, Croitoru, Crooks, Agouris, & Stefanidis, 2018).

In the research conducted for this paper, the focus was on three main questions:

(1) How well does the SURF + BoVW + SVM method work for delineating Slums, Built-up and Non-Built-up areas by classifying image pixels?
(2) How applicable is this method for images of different resolution?
(3) How well does this method perform for different cities?

To answer these questions, the methodology was applied to delineate boundaries of *Slums, Built-up* and *Non-Built-up* areas in two cities in India: Bangalore and Kalyan. These cities have rather different urban morphologies in particular with regards to slums. The use cases are discussed in detail in the Section Case studies and the methodology and experimental set-up are explained in the Section Methodology. The obtained results are presented in the Section Results and discussed in the Section Discussion. Our main conclusions and future plans are stated in the Section Conclusions and future work.

Case studies

In this section, the context of the two cases for testing the developed methodology, the data and the quantitative methods employed to detect the slums and non-slum areas in Kalyan in 2008 and in Bangalore in 2017 are presented.

Case 1 – Kalyan

Kalyan Dombivli (KD) is a fast-growing twin city in the Mumbai metropolitan region in the state Maharashtra, whose growth can be attributed to both being close to the megacity of Mumbai located on a peninsula, as well as to the transport connectivity through rail and road. It has a population of around 1.2 million according to the Census 2011[4] and occupies an area of 67 km^2. Since the 1970s, it received poorer migrants from Uttar Pradesh and Bihar and entrepreneurs from the neighbouring state Gujarat. The majority of its residents are Hindi (80.75%), followed by Buddhists and Muslims with 7.28% and 6.76%, respectively. It has a relatively high literacy rate of 91.37%. About 8% of the population reside in slums, while nearly 43% live in slum-like conditions (Kalyan-Dombivli Municipal Corporation (KDMC), 2007). Several anti-poverty schemes have been implemented in KD, such as the Jawaharlal Nehru National Urban Renewal Mission (JNNURM)[5] sub-program Basic Service Provision for the urban poor. The more recent Rajiv Awas Yojana housing program focuses on GIS-based mapping of slum settlements, where at least two areas were identified for a pilot study (Baud et al., 2013).

Case 2 – Bangalore

Bangalore is the capital of the state of Karnataka, and one of the fastest growing cities in India. Bangalore is the fifth largest city and third most populous city in India, located on the Deccan plateau in the south-east part of Karnataka. It is a multi-cultural city permeating class, religion and language. The city of Bangalore has 21.5% of the total slum population in the state of Karnataka, and every fourth person within the city limits lives in a slum (Roy et al., 2018). The population living in the slums of Bangalore has doubled in a decade (2001–2011) and this poses a serious challenge to urban planners and policy makers. This rapid increase in slum population in Bangalore has been attributed to the high rate of rural-urban migration in the past three decades coupled with a high fertility rate (Krishna, 2013; Krishna, Sriram, & Prakash, 2014; Roy, Lees, Pfeffer, and Sloot (2018); Schenk, 2001). According to the Karnataka Slum Development Board, the city has around 597 slums. However, the Association for Promoting Social Action estimates that the city has over 1500 slums, which are not counted by the government, illustrating the importance of the issue. Therefore, an automated method of delineating slums from satellite images can help in addressing such issues providing a basis for policy intervention based on more accurate data.

Data acquisition

The datasets for Kalyan and Bangalore consist of two types. First, a GE RGB image was downloaded from GE using maximum zoom level with sub-meter pixel resolution and second, a GIS layer representing

[4]www.census2011.co.in/census/city/369-kalyan-and-dombivali.html.
[5]https://www.niua.org/projects/appraisal-urban-reforms-agenda-under-jnnurm.

boundaries of slums. The slum locations were marked using GPS devices followed by on-screen digitizing to collect the geographic coordinates of slum boundaries. These coordinates were then compiled in the Map Puzzle software[6] and saved as a layer of points in DXF format, which was finally converted into a shapefile in ArcGIS (ESRI 2011) that could be overlaid over the satellite RGB image. The final base maps are geo-referenced raster maps with borders outlining the different slums in Kalyan and Bangalore. These maps were used as ground truth for the slum class.

Kalyan

The 2008 RGB image of KD was acquired within the NWO Integrated Program funded project "The role of spatial information infrastructures for tackling urban poverty in Indian cities". The image has dimensions 8193 × 8194 pixels and a spatial resolution of 0.6m per pixel, for example, it covers an area of 4.9 km^2. The GIS layer representing boundaries of 39 slums was created by a private sector consultant to the municipality of KD within the context of the national urban renewal program JNNURM, and modified based on ground truth collected in 2008.

Bangalore

Satellite imagery of the city of Bangalore, India, captured from GE was used to create a base-map of slums in ArcGIS 10.3. A base-map was created using the downloaded geo-referenced GE images as described in the following steps. First, a shapefile of Bangalore city was obtained from Bruhat Bengaluru Mahanagara Palike (BBMP). The total area of Bangalore covered by the shapefile is around 740 km^2. Second, the shapefile is converted into KML through the ArcGIS software and then loaded into GE. The entire Bangalore area is then divided into 455 grids of equal size, each grid covering an area of approximately 1.6 km^2. Each image is geo-referenced using the grids in the GE. Finally, every single image is downloaded from GE (June 2017) with maximum zoom level corresponding to 0.15m resolution. All images are then merged into a single large base-map using global mapper software. Around 1500 slums were identified over the entire Bangalore city for the year 2017 using the list provided by Karnataka Slum Development Board and Association for Promoting Social Action.

Regions of interest (ROIs)

For the purposes of the current study, representative ROIs have been selected from the image data. The main reasons for working on subregions from the data are computational considerations (on a commodity laptop) for this pilot study and mitigating the diversity of image sensors when using GE imagery.

Kalyan

From the original image, an ROI is chosen to be the bounding box of the intersection of the image with the ground truth shapefile. The resulting ROI has a size of 4992 × 6024 pixels or 2995 × 3614 m as shown in Figure 1. Since, there was one image of Kalyan available, only one relevant ROI had been used, which had a computationally permissible size for our experiments.

Bangalore

For Bangalore, there were data covering a much larger region, allowing us to use more ROIs. From the original data, five ROIs have been selected of each of different size covering an area similar to the GE grid size. Since the GE images are a mosaic of images obtained by different satellites, the ROIs have been selected with similar visual appearance indicating that they are probably from the same satellite. The deprived areas in Bangalore consist of many scattered pockets; hence the study areas were chosen to cover as many pockets as possible. The GE images mosaic (some part of the

Figure 1. Image data for Kalyan – selected region of interest (ROI). The ROI is the bounding box of the intersection of the original image data with the ground truth multi-polygon (in red).

[6]http://www.mappuzzle.se/.

city was missing), the slum boundaries and overviews of the chosen ROIs are shown in Figure 2.

The selected regions are shown in Figure 3 and their sizes in image pixels and meters are summarized in Table 1.

Methodology

In this study, we have defined three land-use types (classes) for delineating slum boundaries. The three classes are *Slum* (for the slum areas), *Built-up* (formal built-up areas called Built-up for simplicity) and *Non*-Built-up (usually vegetation and areas without buildings), where the last two represent the non-slum areas. Characteristic generic image features are extracted from image tiles of a certain size. An SVM classifier is then trained to assign each tile to one of the three classes. The trained classifier is used for final pixel-level segmentation of the satellite image ROIs.

The process can be summarized by few key steps. Firstly, the ground truth for the Non-Built-up (mostly vegetation) and Built-up classes is generated. The term "ground truth" is used to coin what is known to be the truth in a semantic classification sense. Secondly, datasets of image tiles are created from the satellite imagery using the ground truth for each class. Thirdly, the datasets are made balanced and partitioned into training and test subsets. Next, a visual vocabulary, or BoVW is created from extracted image feature descriptors from representative tiles of each class. Finally, a multi-class SVM is trained using the BoVW features to classify whether tiles belong to the Slums, Built-up and Non-Built-up class. The performance of our method is measured in terms of classification accuracy and F1 score (Olson & Delen, 2008). To obtain the final segmentation, we use the trained SVM model to classify each pixel of the satellite image: the trained model is used to classify a window (of size equivalent to the tile sizes for training) around every p^{th} of regularly spaced pixels of the satellite image, which results in initial sparse classification. The classes for the remaining unprocessed pixels are determined using majority filtering followed by a final smoothing via majority voting, used often as post-processing regularization steps (Lu & Weng, 2007), of the segmentation result (see Figure 9). The need for initial sparse classification stems from computational considerations, the value of p is determined by optimizing the trade-off between the times needed for the initial classification and the processing of the remaining pixels.

Step 1: ground truth masks

The ground truth for training the classifier was obtained through field surveys (see Section Data acquisition) for the slum regions and computer vision-based annotation for the *Built-up* and *Non-built-up* regions.

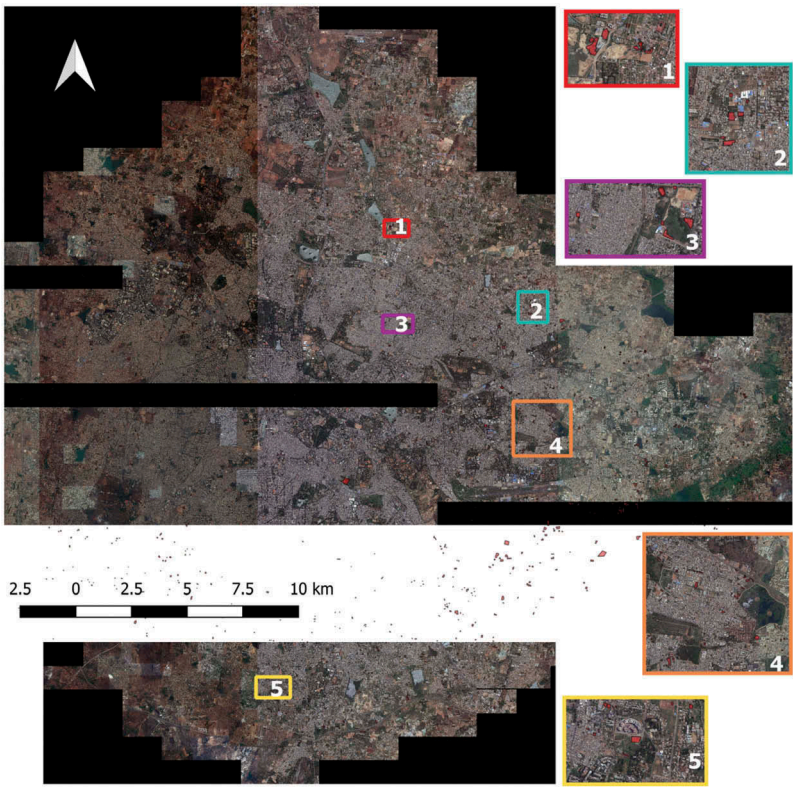

Figure 2. Google Earth images of Bangalore – selected five regions of interest (ROIs). They are numbered from East to West and from North to South. The slum ground truth is depicted by red polygons.

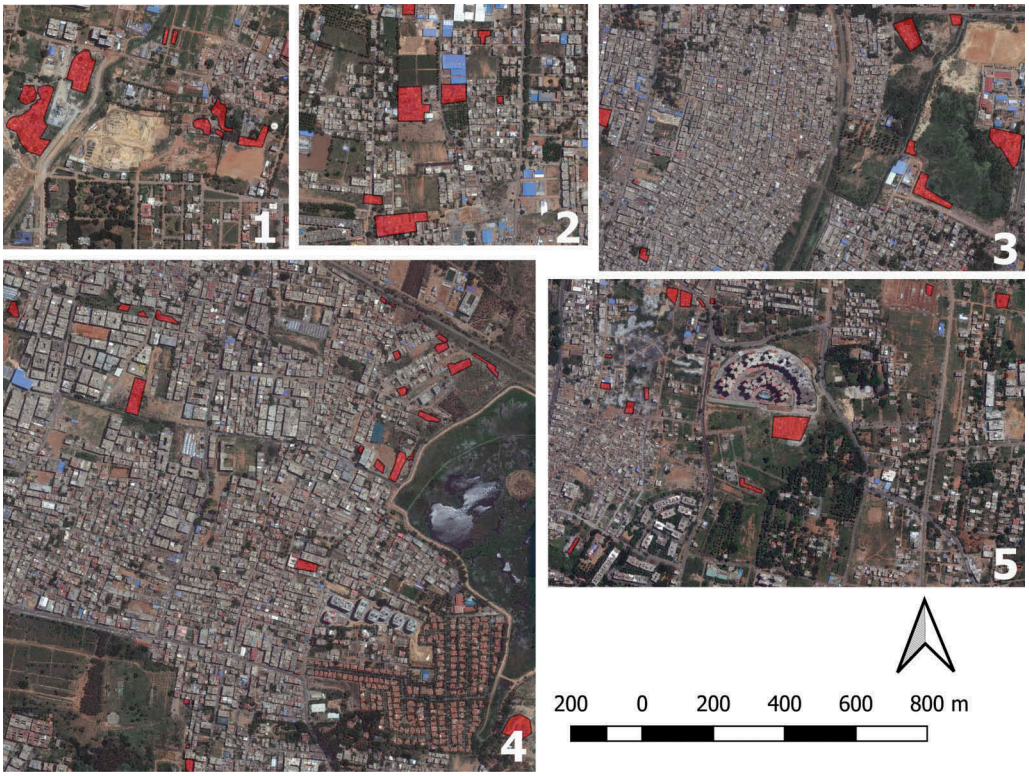

Figure 3. The five selected regions of interest for Bangalore with overlaid slum ground truth in red polygon.

Table 1. Bangalore ROIs resolutions in pixels and meters.

ROI	Resolution [pixels]	Resolution [m]
1	5476 × 5103	821 × 765
2	4983 × 4212	747 × 632
3	8947 × 5939	1342 × 891
4	10,041 × 9960	1506 × 1494
5	8143 × 5174	1221 × 776

To annotate the non-slum area, we took the ROI and used standard vegetation indices to detect the Non-Built-up (indicating mostly vegetation) areas. Since the data are purely RGB and do not contain the near infrared spectral bands, a vegetation index that uses only the red, green and blue channels was needed. After comparison of several such indices, including TGI (Hunt et al., 2013), VARI (Gitelson, Kaufman, Stark, & Rundquist, 2002), rgNDVI (Motohka, Nasahara, Oguma, & Tsuchida, 2010), rbNDVI (Tanaka et al., 2007), VDVI (Xue & Su, 2017) and Visible Vegetation Index (VVI) (Laboratory, n.d.), the VVI was chosen to detect vegetation. The VVI has produced the most compact and connected vegetation segments. For the lack of actual ground truth, visually inspecting the result of this index showed that the high values of the index correlate with the vegetation areas in the images, as can be seen in Figure 4. A threshold value of 25 followed by mathematical morphological processing[7] was determined in order to obtain good delineation of vegetation areas (see also Figure 6.).

The remaining pixels of the ROI, for example, those that were not labelled as *Slums* nor as *Non-Built-up* (by delineating vegetation as its approximation) were labelled as built-up. This results in an image segmentation, as illustrated in Figure 5 for Kalyan and in Figure 6 and A1 (in Appendix 2) for the two of the Bangalore ROIs. These segmentations serve as the ground truth (reference) for the *Built-up* and *Non-Built-up* classes.[8]

As can be seen in Figures 5 and 6, the ground truth for slums is the most precise compared to the other classes because the slum classes were collected using field surveys while the *Built-up* and *Non-Built-up* classes were generated using computer algorithms.

Step 2: tile dataset generation

From all ROIs, image tiles for each of the three classes have been generated using the ground truth masks. Each ROI is scanned from left to right and from top to bottom with a square sliding window of size $N \times N$ pixels and stride (overlap) of $n = N/2$ pixels for only the pixels labelled as belonging to each ground truth class. A tile is selected to represent that class if its central pixel belongs to the class and at least 80% of the pixels of the tile have the desired class label. We

[7]https://en.wikipedia.org/wiki/Mathematical_morphology.
[8]All through the paper we use the following colour-coding for the class labels: red for slums, blue for built-up and green for non built-up.

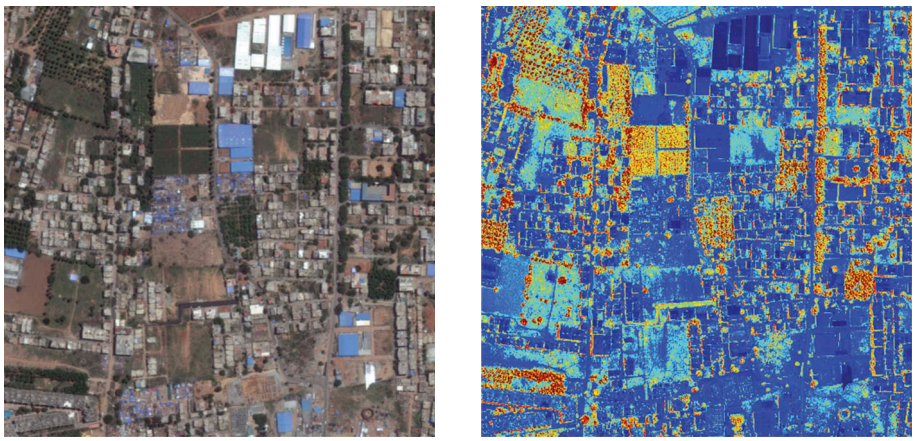

Figure 4. The result of applying VVI to ROI 2 of Bangalore. Left: RGB image and Right: the output of the Visible Vegetation Index, VVI . The "hotter" the colour, the higher the value-probability of vegetation pixel.

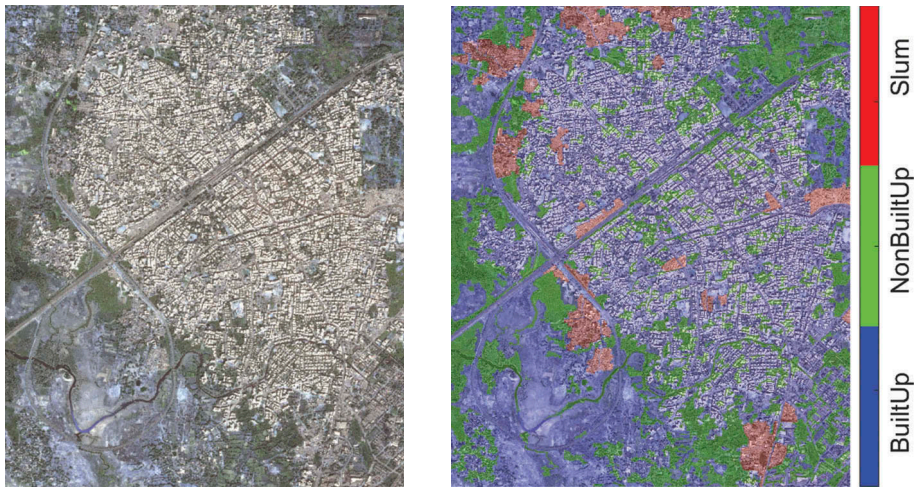

Figure 5. The Kalyan ROI. Left: original image and Right: overlaid ground truth for slums, non built-up (mostly vegetation) and the remaining areas (built-up).

Figure 6. The second of the selected Bangalore ROIs. Left: original image and Right: overlaid ground truth for slums, non built-up (vegetation) and the remaining areas (built-up).

have considered several different tile sizes to study the best tile resolution for our task. For Kalyan, N = 50, 100, 150, 200 and 250 m, while for Bangalore, the tile sizes are N = 10, 20, 30, 40, 50 and 60 m. This difference is due to the different image resolutions for the two cities and a difference in slum area sizes explained before; slums in Bangalore (average size 1.48×10^{-3} km^2) are in general smaller compared to slums in Kalyan (average size 33.11×10^{-3} km^2).

Figures 7 and 8 illustrate some of the generated tiles with different sizes and from the three different classes for Kalyan and Bangalore.

The generated tiles are combined into tile datasets containing representative image tiles for each class along with their labels. The parameters for the tile generation and the number of generated tiles are presented in the Experimental setup section.

Step 3: balancing, training and test sets

Since the slum areas are the smallest (in terms of area) in comparison to the other two classes, there is less data on which to train a classification model, which makes the task more challenging. To overcome these limitations and to make the best use of the available data, two approaches were used: (1) dataset balancing and (2) dataset combination. To balance the dataset, the same number of image tiles are selected from each class. The smallest number of tiles in all cases come from the Slum class, hence all slum tiles are used, while the same number of tiles are selected randomly from the other two classes generating a balanced dataset. For Kalyan, we do have only one ROI, but for Bangalore all generated tiles from all ROIs are combined to create six image datasets corresponding to the six tile sizes. After the construction of the datasets, each dataset was divided

Figure 7. Random selection of 4 tiles from the Kalyan ROI per class. Each column corresponds to a different class, from left to right: Built-Up, Non Built-Up and Slum. The rows from top to bottom correspond to tiles of sizes 50, 100 and 200 m, respectively.

Figure 8. Random selection of 4 tiles from Bangalore ROIs per class. Each column corresponds to a different class, from left to right: Built-Up, Non Built-Up and Slum. The rows from top to bottom correspond to tiles of sizes 20, 40 and 60 m, respectively.

into a training (80%) and a test set for validation (20%) based on the Pareto principle.[9] For the exact sizes of the datasets (in numbers of generated tiles) refer to Tables 2 and 3 in the Experimental setup section.

Step 4: bag of visual words

The concept of a visual vocabulary or BoVW (Li & Perona, 2005) is a classic computer vision method which has been inspired by the text retrieval community. The analogy is that the distribution of words in a text document can be used to compactly summarize and represent the document by its word counts (known as a bag-of-words) and index the document for efficient retrieval. Similarly, an image can be represented by the most characteristic image patches generalized from several images representing the same semantic class.

The standard procedure to create the visual vocabulary or BoVW consists of (1) collecting a large sample of low-level features from a representative dataset of images, and (2) quantizing the feature space according to their statistics.

To obtain the low-level features (e.g. histogram-of-gradients) standard low-level image feature detector and descriptor are used. Often Scale Invariant Feature Transform (SIFT) (Lowe, 1999) and SURF (Bay et al., 2008) are used for these tasks. Although SIFT is usually the best performing technique, often SURF is chosen because of its equally high discriminating ability, but smaller computational cost in comparison to SIFT (Panchal, Panchal, & Shah, 2013).

To generate the VW from the low-level features, often simple k-means clustering is used for the quantization of the features space; the size of the vocabulary $V = k$ is a user-supplied parameter. The VW are then

[9] https://en.wikipedia.org/wiki/Pareto_principle.

Table 2. Number of selected image tiles from the Kalyan ROI per class and in total for different tile sizes.

	Tile size, *N*				
ROI	50 m 84 px	100 m 167 px	150 m 250 px	200 m 334 px	250 m 417 px
1 (per class)	707	115	33	10	2
All (for all classes)	2121	345	99	30	6

Table 3. Number of selected image tiles per Bangalore ROI and in total for different tile sizes.

	Tile size, *N*					
ROI	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px	50 m 334 px	60 m 400 px
1 (per class)	761	147	50	19	11	2
2 (per class)	729	132	45	19	10	6
3 (per class)	387	58	20	6	3	2
4 (per class)	614	88	21	6	2	0
5 (per class)	539	100	28	10	4	1
All (per class)	3030	525	164	60	30	11
All (for all classes)	9090	1575	492	180	90	33

the cluster centres. Features from a new image can be translated into words by determining which visual word they are nearest to in the feature space. Therefore, in general, the image patches assigned to the same visual word should have similar low-level appearance.

Using this technique, the empirical distribution of VW for an image is captured with a histogram counting how many times each word in the visual vocabulary occurs within it. This representation is very convenient as it transforms a set of high-dimensional local image descriptors into a single sparse vector of fixed dimensionality across all images.

Step 5: multi-class SVM training

After the BoVW representation, a model can be trained to classify an image tile into one of the three classes. We have chosen multi-class SVM for both theoretical and empirical reasons.

Previous studies have shown that SVM, originally designed for binary classification, can be effectively extended for multi-class classification tasks (Hsu & Lin, 2002). Currently there are two approaches for implementing multi-class SVM. One is by constructing and combining several binary classifiers, while the other is by directly considering all data in one optimization formulation (Hsu & Lin, 2002). Further, recent studies have shown that linear SVM (compared to non-linear SVM) obtained a better classification score and did not show any signals of over fitting (Duque, Patino, & Betancourt, 2017b). We have tested 22 different classifiers (e.g. Random forest, KNN, PCA, SVM with several different kernels, etc.) on our tile datasets. The best-performing classifier from this experiment was the multi-class SVM, which achieved 16.6% more accuracy than the least performing classifier (coarse KNN) and 0.6% more than the second-best classifier (Fine KNN). Therefore, in this paper, we have used a linear multi-class SVM to delineate *Slums, Built-up* and *Non-Built-up* areas from satellite images of Bangalore and Kalyan.

To measure the performance of the proposed classification method, following standard measures were calculated: accuracy (defined as the ratio of number of correct predictions over the total number of predictions) per class and *F1* score (Olson & Delen, 2008) were used. Accuracy measures how often the method correctly classifies a tile. The F1 score combines precision and recall and gives a more balanced view of the method's capabilities.

Step 6: image pixel segmentation

After the multi-class SVM classifiers with all combinations of different parameters (such as tile size, N and SURF vocabulary size, V) have been trained on the BoVW from SURF features extracted from all "pure" tiles from the selected ROIs, the best performing model with corresponding parameters was chosen for the pixel-level classification. To classify each pixel, it takes on average 0.03s on a laptop (Inter Core i7-6560 CPU @ 2.21 GHz, 19 GB RAM, Windows 10 Pro) and given the ROI sizes in pixels (see Table 1), the computational time is a challenge. Therefore, each p-th pixel in both dimensions was assigned an initial label via direct classification of the tile cantered around the pixel. Each of the remaining unprocessed pixels are given the most frequent initial class label from a window of size $P \times P$, $P \geq 2p + 1$, cantered around the pixel. Determining the value of an unprocessed pixel takes 3 orders of magnitude less time compared to assigning initial label, but there are usually many more unprocessed pixels. Therefore, for the selection of p this trade-off should be considered. Please, note that if the implementation is parallelized and run on multiple compute nodes, it is possible to skip this step and directly process each pixel. The final segmentation is obtained by assigning final class labels to all pixels using majority filtering via a sliding window of size $M \times M$ for each pixel leading to a less noisy and smoother result. Figure 9 illustrates the pixel segmentation steps with example parameter values.

Figure A3 in Appendix 2 illustrates the result of each segmentation step for a zoomed area of Bangalore ROI1 with parameters $p = 10$, $P = 22$ and $M = 30$. These parameters were selected for computational considerations, as well as to ensure enough data for smooth interpolation from the sparse to the final full segmentation.

Experimental set-up

To test our method, we have performed several experiments. Here we describe the set-up and the chosen parameters.

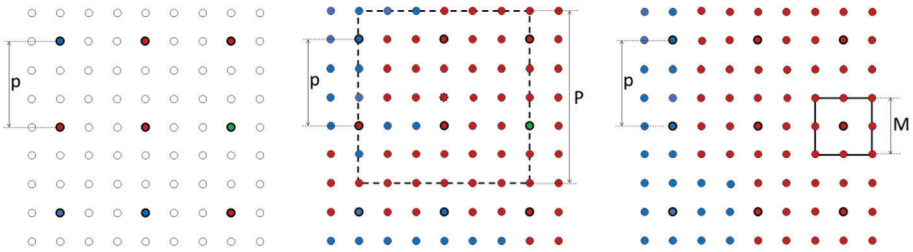

Figure 9. Pixel-level segmentation. Left: Every p = 4th pixel gets initial class label via classification of a tile with size N centred around the pixel; Middle: Assignment of labels to all unprocessed pixels using sliding windows of size P = 7 (this value is only for illustration purposes. For this example, P should be bigger than 9.). The centre of the dashed window gets the majority red label (4 red, 1 blue and 1 green initial labels within the window); Right: Final segmentation – all pixels within each window of size M = 3 get the majority pixel label. For the displayed window the majority label is red (8 red and 1 green removed as noise). Note the smooth boundaries between the red and blue areas and that some initial labels change.

Tile generation

To get a good estimate of the best spatial scale for feature extraction we have generated and tested on image tiles of different sizes. We have chosen these tile sizes guided by the spatial image pixel resolution and mostly by the size of the slums for both cities. As discussed in Section Step 2: Tile dataset generation, the slum class is the smallest class and for some ROIs there are very few slum tiles that are more than 80% pure. We created a balanced dataset by first selecting as many slum tiles with 80% purity or higher. We then selected an equal number of tiles from the other classes. The exact tile sizes in meters for Kalyan are *N*: 50.4, 100, 150, 200 and 250 m, but for convenience they are rounded to the nearest lower integer number. All the different tile sizes and the resulting number of tiles per ROI for Kalyan are summarized in Table A1 in Appendix 1. The number of images per class for the balanced datasets is summarized in Table 2.

There are only six tiles for the resolution of 250 m, which is too few for training, hence, we do not consider this resolution further. The exact tile sizes for Bangalore in meters are *N*: 10.05, 20.1, 30, 40.2, 50.1 and 60 m, for convenience they also rounded to the nearest lower integer number. All the different tile sizes and the resulting number of tiles per ROI for Bangalore are summarized in Table A4 in Appendix 2. The number of images per class for the balanced datasets are summarized in Table 3.

Tile classification

After the construction of the tile datasets, each dataset has been divided into training data using 80% of the tiles and testing data using the remaining 20%. The BoVW model was implemented using different vocabulary sizes (V = 10, 20 and 50) and the strongest 80% SURF features were used for classification. In this experiment, a three-class SVM classifier was trained on the training subsets and the performance evaluated on the test subsets, hence the level of granularity for classification is the image tiles. This experiment was performed for each of the different tile sizes datasets in order to determine the optimal tile size.

Pixel-level segmentation

After determining the best performing tile resolution from the tile classification experiments, the following parameter values for obtaining the pixel-level segmentation were used, Table 4.

Results

This section presents the results of our experiments. They are examined in two ways: based on the tile classification of both cities and using the pixel level classification to make a segmentation. The segmented results are compared qualitatively with the ground truth (reference) segmentation.

Tile classification

It is important to avoid over-fitting while performing classifier training. A set of five performance indicators were measured both during the training and during testing phases: accuracy, precision, sensitivity (recall), specificity and F1 score. In this paper, only accuracy and F1 score (as a composite measure) are reported, but it is very important to consider which statistical measures need to be optimized for a given application. The results for the chosen performance measures during training are given in Appendices 1 and 2, while the results on the test sets of tiles are given below.

While we are mostly interested in the performance of Slum classification, we give the results for the other two classes (*Built-up* and *Non-Built-up*) for completeness and to illustrate the suitability of the method to tackle semantic classification of different semantic classes.

Table 4. Parameters used for the pixel-level segmentation.

	Parameters				
City	Best tile size, *N*	Vocabulary size, *V*	Initial pixel labelling, *p*	De-noising filter size, *P*	Majority filter size, *M*
Kalyan	150 m | 250 px	50	10	22	30
Bangalore	40 m | 268 px	50	10	22	50

Kalyan

The performance metrics have been computed for all tile datasets, but due to the small number of tiles with size greater than 150 m (see Table 2), the metrics are noisy and unreliable. Here, we present the results for the datasets with tile sizes up to 150 m. The performance measures for Kalyan during training are given in Appendix 1, while Tables 5 and 6 below summarize the performance on the test sets.

Bangalore

The performance indicators were measured for all tile datasets, but due to the small number of tiles with size greater than 40 m (see Table 3), they are noisy and unreliable. Here, the results for the datasets with tile sizes up to 40 m are presented. The performance metrics during training are given in Appendix 2, while those on the test set of tiles are given below. Tables 7 and 8 summarize the tile classification performance on the test set: accuracy and F1 score, respectively.

Table 5. Accuracy for Kalyan tile classification during testing, [%].

		Tile size, *N*		
Vocabulary size	Class	50 m 87 px	100 m 167 px	150 m 250 px
10	Slum	73.0	75.4	66.7
	Built-up	75.2	86.9	85.7
	Non-built-up	72.3	68.1	80.9
20	Slum	74.2	86.9	80.9
	Built-up	76.4	88.4	85.7
	Non-built-up	72.8	78.3	76.2
50	Slum	78.9	82.6	90.5
	Built-up	78.7	82.6	95.2
	Non-built-up	74.7	76.8	95.2

Table 6. F1 score for Kalyan tile classification during testing.

		Tile size, *N*		
Vocabulary size	Class	50 m 87 px	100 m 167 px	150 m 250 px
10	Slum	0.62	0.45	0.36
	Built-up	0.61	0.80	0.82
	Non-built-up	0.57	0.63	0.71
20	Slum	0.65	0.82	0.75
	Built-up	0.64	0.80	0.82
	Non-built-up	0.56	0.67	0.44
50	Slum	0.69	0.77	0.86
	Built-up	0.67	0.70	0.93
	Non-built-up	0.63	0.65	0.92

Table 7. Accuracy for Bangalore tile classification during testing, [%].

		Tile size, *N*			
Vocabulary size	Class	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px
10	Slum	70.4	78.7	90.9	83.33
	Built-up	67.8	75.6	83.8	80.6
	Non-built-up	68.4	80.9	84.9	80.6
20	Slum	70.9	81.3	87.9	91.7
	Built-up	67.2	73.7	78.8	88.9
	Non-built-up	69.1	81.6	84.8	91.7
50	Slum	70.1	80.9	86.9	97.2
	Built-up	67.8	76.5	82.8	91.7
	Non-built-up	68.5	80.3	83.8	94.4

The performance evaluation for both Kalyan and Bangalore indicate that the larger the tile size (while still ensuring the semantic class "purity") and the larger the BoVW vocabulary, the higher the performance. This trend is illustrated, for example, for the F1 score values from Table 8 in respect to the tile size and vocabulary size on Figure A2 in Appendix 2.

Therefore, for the pixel-level segmentation of the ROIs, the optimal parameter values were selected: tile sizes $N = 150$ m (250 px) for Kaylan, $N = 40$ m (268 px) for Bangalore and the best visual vocabulary size for both cities $V = 50$. We observed also that the Slum class has usually the highest performance due to the most accurate ground truth used for generating the tiles.

Pixel-level segmentation

The purpose of the segmentation is to delineate the spatial areas occupied by *Slums, Built-up* and *Non-Built-up* areas. After successful training of the 3-class SVM classifier (Section Tile classification) and experiments with all combinations of parameters, the parameter values corresponding to the highest classification accuracy were chosen for the segmentation (see Table 4). With these settings the three steps of the segmentation algorithm as described in Section Step 6: Image pixel segmentation was performed.

Kalyan

The final pixel segmentation result for Kalyan in comparison to the ground truth is shown on Figure 10.

Bangalore

The final pixel segmentation result in comparison to the ground truth is shown in Figures 11 and 12 (A4-A6 in Appendix 2). Figure 11 illustrates a relatively good segmentation result, while Figure 12 shows a poorer segmentation.

Discussion

Tile classification

The accuracy of our classification is presented in Tables 5 and 7 for Kalyan and Bangalore, respec-

Table 8. F1 score for Bangalore tile classification during testing.

		Tile size, *N*			
Vocabulary size	Class	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px
10	Slum	0.59	0.70	0.86	0.75
	Built-up	0.27	0.60	0.75	0.77
	Non-built-up	0.62	0.71	0.76	0.58
20	Slum	0.61	0.72	0.83	0.87
	Built-up	0.30	0.57	0.67	0.87
	Non-built-up	0.60	0.73	0.76	0.85
50	Slum	0.59	0.72	0.81	0.96
	Built-up	0.27	0.60	0.70	0.89
	Non-built-up	0.61	0.72	0.76	0.90

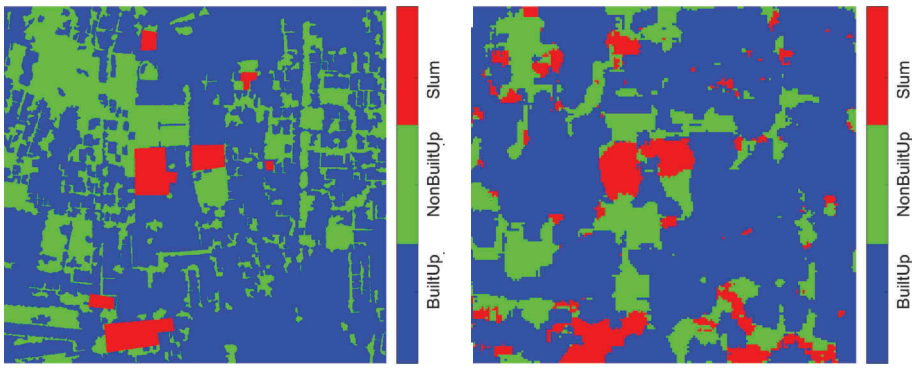

Figure 10. Kalyan ROI: Left: ground truth and Right: the result from pixel-level classification (segmentation) for all classes.

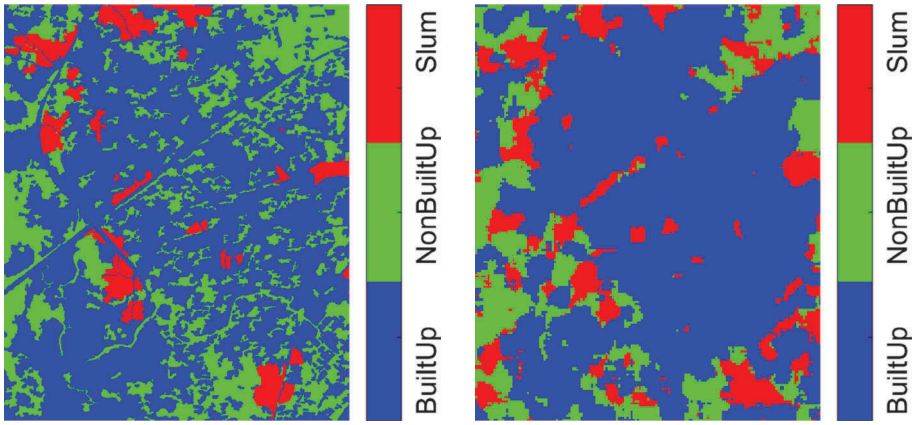

Figure 11. Bangalore ROI 2: Left: ground truth and Right: result from pixel-level classification (segmentation) for all classes.

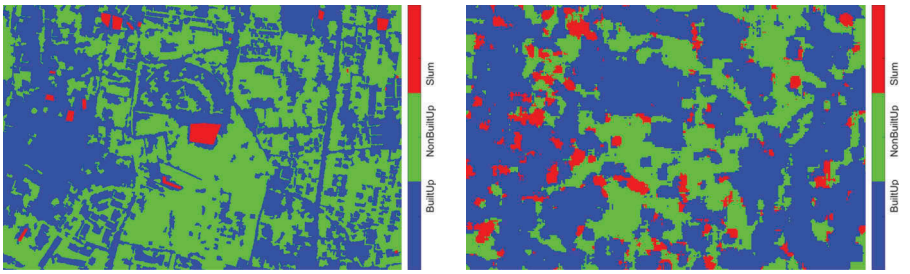

Figure 12. Bangalore ROI 5: Left: ground truth (and Right: result from pixel-level classification (segmentation).

tively. For slums, the performance is quite good with a minimum of 73.0% for Kalyan and 70.4% for Bangalore using a vocabulary size of 10 and a tile size of 50 and 10 m, respectively. The best accuracy for both cities, 90.5% for Kalyan and 97.2% for Bangalore, is achieved using the largest tile size (150 and 40 m, respectively) and with the largest vocabulary of size 50.

The accuracy for the other classes is a little lower than the accuracy for slums; however, it is still quite high. It also follows the same pattern as for slums, the best accuracy is achieved with the largest tile size and the largest vocabulary. We also calculated the F1 scores for our method which are presented in Tables 6 and 8. These results show that the combination of largest tile and the largest vocabulary performs best. The highest F1 values (around 0.9 or higher) demonstrate that the classification is not only accurate, but also has high precision and recall.

The correlation between performance and visual vocabulary size can be explained by the ability of a larger vocabulary to better capture the characteristics and variance of the various classes. To train a larger vocabulary well, more SURF features are required, which corresponds to larger tiles. As we mention in Section Tile generation, tiles larger than 150 m for Kalyan and 40 m for Bangalore result in tile sets, which are not of sufficient size to train a good classifier. In other words, a large vocabulary trained on tiles that are as large as possible, given the data and physical extent of the slums, seems to give the best results. In addition, to classify slum areas, a relatively large neighbourhood (context) is essential as slum and formal built-up areas often have similar spectral,

but different contextual characteristics (e.g. different object sizes, layout patterns). However, as slums vary in size, the definition of an optimal tile size needs a thorough investigation and also depends on the urban morphology (e.g. Bangalore having much smaller slum areas compared to Kalyan). The use of larger tile sizes, which leads to better performance, has limitations due to the availability of training data. Reliable training data are limited both due to their nature (e.g. small area slum pockets) and the complexity of performing ground checks.

Comparing the results on the test sets with those on the training sets, presented in Appendices 1 and 2, we can observe that the accuracy and F1 scores are similar. Therefore, the classifier has not been overfitted during training and can generalize well.

The classification of slums using this technique achieved comparable accuracy levels as reported in other recent studies on slum mapping (Kuffer, Pfeffer, Sliuzas, & Baud, 2016). Furthermore, the method uses only one type of low-level feature (SURF) on RGB images compared to the usually large sets of low-level features or multispectral images to achieve such performance. For example, a recent paper used a set of 30 (9 Spectral, 11 Texture and 10 Structure) features (Duque et al., 2017a). Also, our results show that this generic method from computer vision works well independently of the semantics of the classes, the tiles for non-slum areas have been classifier equally well.

Pixel-level segmentation

The results for the pixel level segmentation are shown on Figures 10–12. In most cases where the ground truth indicates slum, there is also a blob for a slum in the segmented image. In general, we notice that there are more slums in the segmentation than in the ground truth and the segmentation for *Non-Built-up* and *Built-up* areas roughly corresponds to the ground truth. The segmentation results indicate that the majority of false-negatives (missed slums) were tiny pockets of slums. This can be explained by the fact that the classifier has been trained on large tiles, compared to the size of these slums. An important observation is that for Bangalore, we have trained our classifier from tiles from all five ROIs; therefore, for segmenting each ROI training data of other parts of the city were used. This indicates robustness in respect to the available training data.

To get a better understanding of the performance of the proposed segmentation method, we focus on areas where the result differs significantly from the given or generated ground truth. A difference between the segmentation result and the ground truth, however, does not always mean that the segmentation is wrong. Unfortunately, the ground truth generation, as described in Section Step 1: Ground truth masks, is not perfect and does not always reflect the reality. In some cases, the segmentation method makes mistakes (that is wrong segmentation compared to the observed truth), while in others it classifies a region as belonging to the true observable class, while the generated ground truth is erroneous. The dilemma of generating ground-truth data on slums was highlighted in a recent study, showing that even slum experts do not necessarily agree on the location of slums in a complex city (Pratomo, Kuffer, Martinez, & Kohli, 2017). Therefore, we focused on some of those image areas (c.f. Figures 13, 14 and 15) which illustrate the strengths and weaknesses of the proposed segmentation method, as well as the weaknesses of the ground truth generation. The labels given or generated by the ground truth methods or obtained by the segmentation method are overlaid on the image data. This allows us to see the "observable truth" - image regions which based only on the RGB image information could be identified as *Slums, Built-up* or *Non-Built-up* by the human eye.

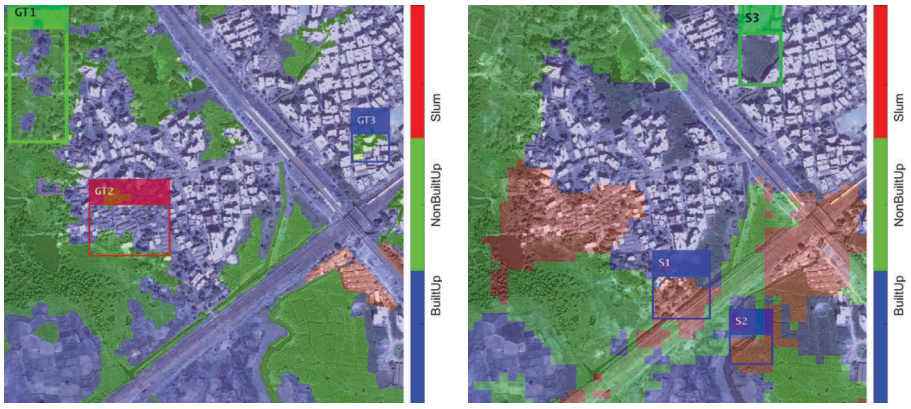

Figure 13. A subregion of the Kalyan ROI: Left: ground truth and Right: the result from pixel-level segmentation overlaid on the image. Some areas are delineated roughly as annotated rectangles indicating the largest errors in relation to the observable truth (not always the same as the generated ground truth). GT stands for Ground Truth and S for Segmentation.

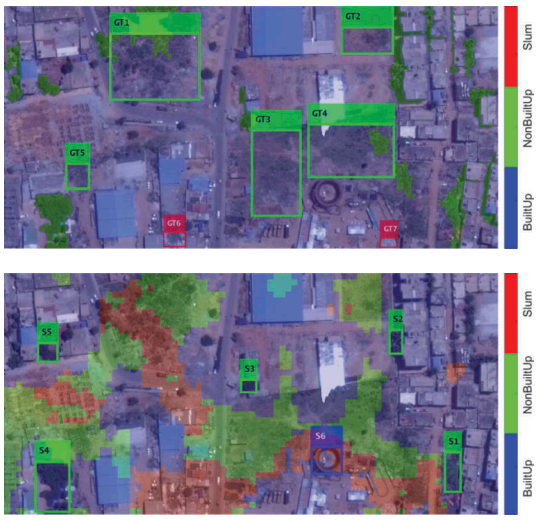

Figure 14. A subregion of Bangalore ROI 2: Top: ground truth and Bottom: the result from pixel-level segmentation overlaid on the image. Some areas are delineated roughly as annotated rectangles indicating the largest errors in relation to the observable truth (not always the same as the generated ground truth). GT stands for Ground Truth and S for Segmentation.

Capturing slums with EO imagery faces the challenge that densely formal and slum areas often share similar morphological characteristics, while slum areas can be very diverse across different cities but also within cities different slum typologies (Kuffer et al., 2017) exist (e.g. long established slums compared to very recent temporary shelters). This makes the generation of ground-truth data on slums and the classification of slums extremely challenging. For example, false positives often relate to areas that have similar morphological characteristics as slums (also called morphological slums (Friesen, Taubenböck, Wurm, & Pelz, 2018; Wurm & Taubenböck, 2018)) but are not included into the ground truth data as they do not have slum like conditions on the ground. This leads to the question, whether a false negative (an omission) or false positive (a commission) error in slum mapping is more problematic. As EO-based slum mapping only can indicate possible slum locations, where for a final decision ground-truth checks are essential (as only on the ground the living conditions and deprivation levels faced by inhabitants can be assessed), the omission of a slum location by an image classification approach seems more problematic.

In the following images several rectangular areas indicate some of the mistakes made by either obtaining the ground truth or the segmentation. The colours of the rectangles indicate the "observable truth" (e.g. the truth as we see it in the image), but they are located where the respective method differs from it, that is, makes a mistake. The abbreviation "GT" stands for Ground Truth, while "S" stands for Segmentation in the rectangle headings. Hence, the GT regions indicate where the ground truth is wrong, and the S regions where the segmentation method has made an error.

Kalyan

For the Kalyan ROI segmentation (see Section Step 6: Image pixel segmentation and Figure 10), we observed that most of the slum areas are detected successfully, along with many false positives. Figure 13 focuses on such an area in Kalyan.

What we can see is that both the ground truth and the segmentation have made mistakes with regards to Non-Built-up areas, GT1 shows an area where our method correctly identifies it as Non-Built-up (while it is missed by the ground truth generation

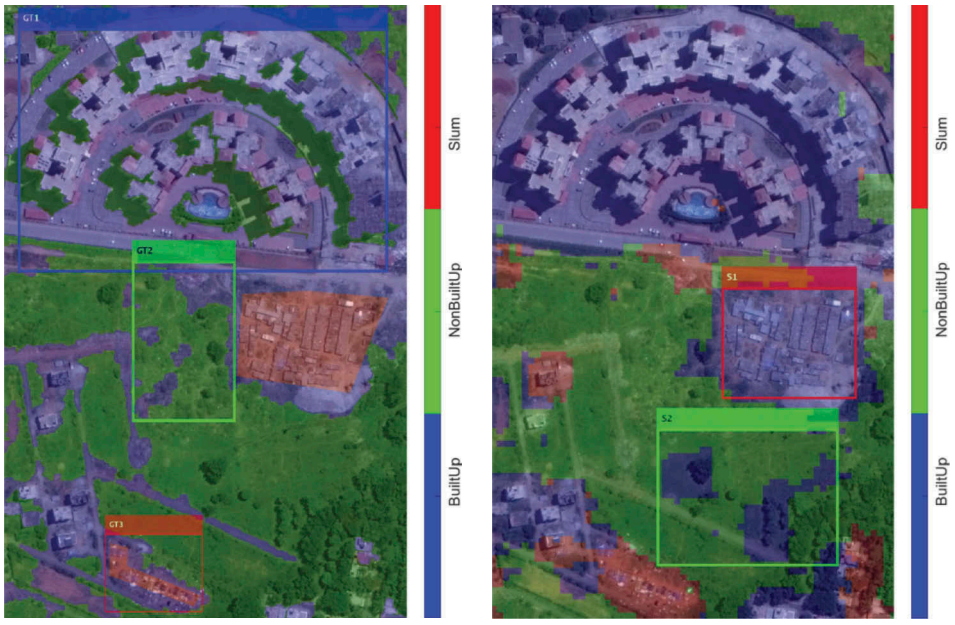

Figure 15. A subregion of Bangalore ROI 5. Left: ground truth and Right: the result from pixel-level segmentation overlaid on the image. Some areas are delineated roughly as annotated rectangles indicating the largest errors in relation to the observable truth (not always the same as the generated ground truth). GT stands for Ground Truth and S for Segmentation.

algorithm); however, it misses the area S3 (where VVI performs well). The GT2 area is particularly interesting, as our method detected this as a slum. Visually this region has a slum-like appearance and could be a morphological slum. Area GT3 indicates a limitation of the way we created the ground-truth. The vegetation detection algorithm indicated this area as vegetation and although some trees are certainly there, there seems to be confusion with deep shadows. Areas S1 and S2 are mistakes made by the segmentation method due to the presence of a highway. In general, slums are often located near large transportation axes using available land reserves.

Bangalore

For the Bangalore ROI segmentations (see Section Step 6: Image pixel segmentation and Figures 11 and 12), we also observe that usually most of the slum areas are detected successfully, along with many false positives. Figures 14 and 15 illustrate some areas of discrepancies between the ground truth or segmentation and the observable truth in the ROIs of Bangalore.

Figure 14 illustrates that most mistakes in this area are *Built-up* vs. *Non-built-up* classification. There are several instances where our method detected *Non-built-up* areas that were not labelled as such in the ground truth: GT1-5. Figures 7 and 8 show that tiles used for the training of the slum class often contain also moderate amounts of vegetation. Therefore, the BoVW histograms of slum tiles and *Non-built-up* tiles could be similar. Regions GT6 and GT7 again show examples of areas that visually are very slum-like but are not included in the ground truth as such. This indicates the possibility that our method could be used to detect missing slums in the ground provided they have a visual appearance to (some of) the slum training data. Finally, S6 shows an example of an incorrect segmentation.

Very rarely, we observe missed slum areas. One such omission of a large slum happened in ROI5 (see Figure 12). We studied this instance more in depth by zooming into that area as illustrated on Figure 15.

In this area, we highlight three ground truth and two segmentation mistakes. Firstly, we can see in GT1 another instance of shadows confusing the vegetation detection algorithm using VVI. Secondly GT2 shows an area where the VVI has missed some vegetation. GT3 shows a larger area that visually looks like a morphological slum and may be an omission in the ground truth. Areas S1 and S2 are miss-classifications by our algorithm, S1 is a slum area with larger roofs compared to other slum areas and a more regular layout pattern. Because of these visual characteristics, the area has been not classified as a slum. The problems shown in S2 could have been prevented when using a NIR band. Many of the ground truth issues are observed in *Non-built-up* areas. This does suggest a significant limitation of images with only R, G and B bands, where detection of vegetation is difficult.

Conclusions and future work

In this paper, we have shown a method for detecting slums at city scale using RGB images, which are more easily available compared to often expensive spectral imagery, using techniques from the computer vision domain. Our method combines the BoVW based on SURF features with an SVM. In order to simplify the problem of delineating slums we have converted it into a classification problem by selecting representative tiles from the satellite imagery for our three classes: *Slums, Built-up* and *Non-Built-up*. The VW (generalized SURF features) obtained from these tiles were used to create histograms, which served as an input to the SVM. The SVM was then trained to classify the three classes. Afterwards the best trained SVM was used to classify every 10th pixel of the satellite image. Finally, the remaining unprocessed pixels have been interpolated using majority filtering to create a full image segmentation.

In our research, we have focused on three main research questions. Firstly, we were interested in how well our method works for detecting slums. From the results, we presented, we can conclude that for the classification task our method achieves high accuracy and F1 scores for both case studies. The segmentation results are also reasonable: in most cases our algorithm detects slums where the ground truth indicates slums. We also took an in-depth look at what kind of mistakes our method makes. Most of them are slum false positives. However, since the ground truth is not perfect, in many cases our method detects slums in areas that are not marked as slum in the ground truth, but after visual inspection do look like morphological slums. For a final verification of slum locations, ground checks are required. Therefore, the low level of missed slums (false negatives) is a desirable feature of our proposed methodology.

The other research questions concerned the applicability of our method for images of different resolutions and the potential for generalization over different locations. We can see that our method performs similarly for both Kalyan and Bangalore, even though the difference in morphology of slums in these two cities is rather large. Also, the resolution of the images used for both case studies is different - 0.6 m for Kalyan and 0.15 m for Bangalore. These different resolutions require different parameters for the algorithm; however, for Kalyan, tiles of 150 m were used while for Bangalore, tiles of 40 m performed best. Larger tiles sizes were tried but resulted in too few tiles to effectively train the classifier.

This study showed that techniques used in computer vision have the potential to map slums while being transferable across two rather different Indian cities. Given limited access to standard VHR multi-spectral imagery,

the developed framework would allow for a frequent extraction of slum maps using available GE imagery and visualization of slum dynamics across large, complex and very dynamic cites. However, for a conclusion on slum locations ground checks are necessary to deal with false positives. Such EO-based information, which is validated on the ground, would allow to support planning and policy development as well as monitoring the implementation of slum policy in large and fast-growing cities in the global South, where information on slum locations and dynamics is often scarce. Future research should assess transferability of the method to different regions and consider the correlation between spatial morphology and concepts like segregation, polarisation, exclusion and marginality (Roy et al., 2018).

We have used only one type of image features (SURF) in order to delineate visually very complex and sometimes hard to distinguish class even by a human observer. In the future, we plan to implement a library of image features for training a much more complex model, which will be released as open source. Also, we are planning to use multi-spectral satellite images, which will give a much better input data for the complex segmentation problem.

Acknowledgments

The authors acknowledge the support of Lourens Veen for generating the ROIs.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work has been supported by Netherlands eScience Center [027.015.G05] and DynaSlum: Data Driven Modelling and Decision Support for Slums and SimCity project [C.2324.0293].

ORCID

Elena Ranguelova http://orcid.org/0000-0002-9834-1756
Debraj Roy http://orcid.org/0000-0003-1963-0056
Monika Kuffer http://orcid.org/0000-0002-1915-2069

References

ESRI. (2011). *ArcGIS Desktop: Release 10.3*. Redlands, CA: Environmental Systems Research Institute.

Baud, I., Pfeffer, K., van Dijk, T., Mishra, N., Richter, C., Bon, B., & Saharan, T. (2013). The development of Kalyan Dombivli; fringe city in a metropolitan region. (City Report No. 2), City Growth and the Sustainability Challenge Comparing Fast Growing Cities in Growing Economies. Bonn: Chance 2 Sustain Project. (pp. 58).

Bay, H., Ess, A., Tuytelaars, T., & van Gool, L. (2008). SURF: Speeded up robust features. *Computer Vision and Image Understanding*, *110*(3), 346–359. doi:10.1016/j.cviu.2007.09.014

Chang, A.Y., Parrales, M.E., Jimenez, J., Sobieszczyk, M.E., Hammer, S.M., Copen-Haver, D.J., & Kulkarni, R.P. (2009, Jul). Combining Google Earth and GIS mapping technologies in a dengue surveillance system for developing countries. *International Journal of Health Geographics*, *8*(1), 49. doi:10.1186/1476-072X-8-49

Chen, S., & Tian, Y. (2015, April). Pyramid of spatial relations for scene-level land use classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(4), 1947–1957. doi:10.1109/TGRS.2014.2351395

Du, Q., Zhang, L., & Zhang, L. (2016). A scene change detection framework for multi-temporal very high resolution remote sensing images. *Signal Processing*, *124*, 184–197. doi:10.1016/j.sigpro.2015.09.020

Duque, J.C., Patino, J.E., & Betancourt, A. (2017a). Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, *9*, 895. doi:10.3390/rs9090895

Duque, J.C., Patino, J.E., & Betancourt, A. (2017b). Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, *9*(9), 895. doi:10.3390/rs9090895

Engstrom, R., Newhouse, D., Haldavanekar, V., Copenhaver, A., & Hersh, J. (2017, March). Evaluating the relationship between spatial and spectral features derived from high spatial resolution satellite data and urban poverty in Colombo, Sri Lanka. In *2017 IEEE Joint Urban Remote Sensing Event (JURSE)* (pp. 1–4).

Friesen, J., Taubenböck, H., Wurm, M., & Pelz, P. (2018). The similar size of slums. *Habitat International*, *73*, 79–88. doi:10.1016/j.habitatint.2018.02.002

Gitelson, A.A., Kaufman, Y.J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, *80*(1), 76–87. doi:10.1016/S0034-4257(01)00289-9

Government of India. (2011). *Slums in India a statistical compendium*. New Delhi: Ministry of Housing and Urban Poverty Alleviation.

Graesser, J., Cheriyadat, A., Vatsavai, R., Chandola, V., Long, J., & Bright, E. (2012). Image based characterization of formal and informal neighbourhoods in an urban landscape. *Computers Environment and Urban Systems*, *36*(2), 154–163.

Gunter, A.W. (2009). Getting it for free: Using google earth and IL WIS to map squatter settlements in Johannesburg. *International Geoscience and Remote Sensing Symposium (IGARSS), Cape Town, South Africa, 12-17 July, 2009. (3)* (pp. III–388).

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. doi:10.1109/72.991427

Hunt, E.R., Doraiswamy, P.C., McMurtrey, J.E., Daughtry, C.S., Perry, E.M., & Akhmedov, B. (2013). A visible band index for remote sensing leaf chlorophyll content at the canopy scale. *International Journal of Applied Earth Observation and Geoinformation*, *21*, 103–112. doi:10.1016/j.jag.2012.07.020

Jun, Y., Jiang, Y.-G., Hauptmann, A.G., & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. International Mutlimedia Conference, MM'07 - *9th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval, MIR'07* (pp. 197–206). New York, NY: ACM.

Kalyan-Dombivli Municipal Corporation (KDMC). (2007). *The Kalyan-Dombivli city development plan. (KDMC, in*

association with Subash Patil and Associates.), Kalyan-Dombivli.

Kit, O., Ldeke, M., & Reckien, D. (2012). Texture-based identification of urban slums in Hyderabad, India using remote sensing data. *Applied Geography*, *32*(2), 660–667. doi:10.1016/j.apgeog.2011.07.016

Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. *Computers Environment and Urban Systems*, *36*(2), 154–163. doi:10.1016/j.compenvurbsys.2011.11.001

Kohli, D., Stein, A., & Sliuzas, R. (2016). Uncertainty analysis for image interpretations of urban slums. *Computers, Environment and Urban Systems*, *60*(60), 37–49. doi:10.1016/j.compenvurbsys.2016.07.010

Krishna, A. (2013). Stuck in place: Investigating social mobility in 14 Bangalore slums. *The Journal of Development Studies*, *49*(7), 1010–1028. doi:10.1080/00220388.2013.785526

Krishna, A., Sriram, M., & Prakash, P. (2014). Slum types and adaptation strategies: Identifying policy-relevant differences in Bangalore. *Environment and Urbanization*, *26*(2), 568–585. doi:10.1177/0956247814537958

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space - 15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6), 455. doi:10.3390/rs8060455

Kuffer, M., Pfeffer, K., Sliuzas, R., & Baud, R. (2016). Extraction of slum areas from VHR imagery using GLCM variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *32*, 1830–1840. doi:10.1109/JSTARS.2016.2538563

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, R., & van Maarseveen, M. (2017). Capturing the diversity of deprived areas with image-based features: The case of Mumbai. *Remote Sensing*, *9*, 384. doi:10.3390/rs9040384

Kuffer, M., van Maarseveen, M., Sliuzas, R., & Pfeffer, K. (2017). *Spatial patterns of deprivation in cities of the global south in very high-resolution imagery*. Enschede: University of Twente Faculty of Geo-Information and Earth Observation (ITC.

Laboratory, P. H. (n.d.). Visible vegetation index (vvi). Retrieved fromhttp://phl.upr.edu/projects/visible-vegetation-index-vvi.

Li, -F.-F., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05 (2)* (pp. 524–531).

Lowe, D.G. (1999). Object recognition from local scale-invariant features. In Proceedings of the international conference on computer vision. Washington, DC: IEEE Computer Society (2), pp. 1150–1157.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. doi:10.1080/01431160600746456

Mahabir, R., Croitoru, A., Crooks, A., Agouris, P., & Stefanidis, A. (2018). A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: Trends, challenges and emerging opportunities. *Urban Science*, *2*, 8. doi:10.3390/urbansci2010008

Motohka, T., Nasahara, K.N., Oguma, H., & Tsuchida, S. (2010). Applicability of green-red vegetation index for remote sensing of vegetation phenology. *Remote Sensing*, *2*(10), 2369–2387. doi:10.3390/rs2102369

Olson, D.L., & Delen, D. (2008). *Advanced data mining techniques*. Springer-Verlag Berlin Heidelberg 2008

Panchal, P.M., Panchal, S.R., & Shah, S. (2013). *A comparison of SIFT and SURF*. International Journal of Innovative Research in Computer and Communication Engineering.

Pratomo, J., Kuffer, M., Martinez, J., & Kohli, D. (2016, September). Uncertainties in analyzing the transferability of the generic slum ontology. Retrieved from http://proceedings.utwente.nl/428/

Pratomo, J., Kuffer, M., Martinez, J., & Kohli, D. (2017). Coupling uncertainties with accuracy assessment in object-based slum detections, case study: Jakarta, Indonesia. Remote Sens., vol. 9, no. 11, p. 1164, 2017.

Ranguelova, E., Kuffer, M., Pfeffer, K., Roy, D., & Lees, M. (2017, June). Image based classification of slums, built-up and non-built-up areas in Kalyan, India. In *37th Earsel symposium Smart Future with Remote Sensing*.

Roy, D., Lees, M.H., Palavalli, B., Pfeffer, K., & Sloot, M.P. (2014). The emergence of slums: A contemporary view on simulation models. *Environmental Modelling & Software*, *59*, 76–90. doi:10.1016/j.envsoft.2014.05.004

Roy, D., Lees, M.H., Pfeffer, K., & Sloot, P.M. (2017). Modelling the impact of household life cycle on slums in Bangalore. *Computers Environment and Urban Systems*, *64*, 275–287. doi:10.1016/j.compenvurbsys.2017.03.008

Roy, D., Lees, M.H., Pfeffer, K., & Sloot, P.M. (2018). Spatial segregation, inequality, and opportunity bias in the slums of Bengaluru. *Cities*, *74*, 269–276. doi:10.1016/j.cities.2017.12.014

Roy, D., Palavalli, B., Menon, N., King, R., Pfeffer, K., Lees, M., & Sloot, P.M. (2018). Survey-based socio-economic data from slums in Bangalore, India. *Scientific Data*, *5*, 170–200. doi:10.1038/sdata.2017.200

Schenk, H. (2001). *Living in India's slums: A case study of Bangalore*. New Delhi: IDPAD/Manohar.

Tanaka, S., Goto, S., Maki, M., Akiyama, T., Muramoto, Y., & Yoshida, K. (2007). Estimation of leaf chlorophyll concentration in winter wheat [*Triticum aestivum*] before maturing stage by a newly developed vegetation index-RBNDVI. *Journal of the Japanese Agricultural Systems Society, 2007*, Volume 23, Issue 4, Pages 297-303,.

Taylor, J.R., & Lovell, S.T. (2012). Mapping public and private spaces of urban agriculture in Chicago through the analysis of high-resolution aerial images in google earth. *Landscape and Urban Planning*, *108*(1), 57–70. doi:10.1016/j.landurbplan.2012.08.001

UN Habitat. (2016). *Slums almanac 2015-16. Tracking improvement in the lives of slum dwellers. (Nairobi, Kenya:UNON, Publishing Services Section)*.

Van, D. (2014). *Subaltern Urbanism in India Beyond the Mega-city Slum: The Local Politics of Occupancy and Locality Development* (PhD Thesis). University of Amsterdam, the Netherlands.)

Wurm, M., & Taubenböck, H. (2018). Detecting social groups from space assessment of remote sensing-based mapped morphological slums using income data. *Remote Sensing Letters*, *9*(1), 41–50. doi:10.1080/2150704X.2017.1384586

Wurm, M., Taubenböck, H., Weigand, M., & Schmitt, A. (2017). Slum mapping in polarimetric SAR data using spatial features. *Remote Sensing of Environment*, *194* (Complete), 190–204. doi:10.1016/j.rse.2017.03.030

Wurm, M., Weigand, M., Schmitt, A., Gei, C., & Taubenböck, H. (2017, March). Exploitation of textural and morphological image features in sentinel-2a data for slum mapping. In *2017 joint urban remote sensing event (JURSE)* (pp. 18), IEEE.

Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors(1):1-17*

Appendix 1. Supplementary material for Kalyan

Tile construction

Table A1. Number of all possible image tiles from the Kalyan ROI per class for different tile sizes.

		Tile size, N				
ROI	Class	50 m 84 px	100 m 167 px	150 m 250 px	200 m 334 px	250 m 417 px
1	Slum	707	115	33	10	2
	Built-up	8149	1810	723	387	222
	Non-built-up	1926	281	80	32	14

Tile classification

Table A2. Accuracy for Kalyan tile classification during training, [%].

		Tile size, *N*		
Vocabulary size	Class	50 m 87 px	100 m 167 px	150 m 250 px
10	Slum	74.9	76.8	73.0
	Built-up	75.2	86.9	82.0
	Non-built-up	74.2	78.9	78.2
20	Slum	74.9	84.8	85.9
	Built-up	76.7	90.9	89.7
	Non-built-up	73.1	89.5	98.7
50	Slum	79.7	92.8	98.7
	Built-up	78.0	92.4	97.4
	Non-built-up	76.6	89.5	98.7

Table A3. F1 score for Kalyan tile classification during training.

		Tile size, *N*		
Vocabulary size	Class	50 m 87 px	100 m 167 px	150 m 250 px
10	Slum	0.65	0.50	0.43
	Built-up	0.61	0.82	0.78
	Non-built-up	0.59	0.74	0.69
20	Slum	0.66	0.78	0.79
	Built-up	0.63	0.86	0.86
	Non-built-up	0.56	0.71	0.80
50	Slum	0.71	0.89	0.98
	Built-up	0.64	0.88	0.96
	Non-built-up	0.64	0.84	0.98

Appendix 2. Supplementary material for Bangalore

Tile construction

Figure A1. Bangalore ROI 1. Left: original image and Right: overlaid ground truth for slums, non built-up (mostly vegetation) and the remaining areas (built-up).

Table A4. Number of all possible image tiles per Bangalore ROI and per class for different tile sizes.

		Tile size, *N*					
ROI	Class	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px	50 m 334 px	60 m 400 px
1	Slum	761	147	50	19	11	2
	Built-up	14,300	3026	1251	667	405	285
	Non-built-up	4311	705	237	101	57	32
2	Slum	729	132	45	19	10	6
	Built-up	9349	1892	751	369	215	136
	Non-built-up	3877	607	180	83	40	25
3	Slum	387	58	20	6	3	2
	Built-up	17,806	3282	1200	566	319	208
	Non-built-up	17,270	3200	1171	551	302	191
4	Slum	614	88	21	6	2	0
	Built-up	41,938	8716	3667	1952	1187	774
	Non-built-up	22,635	4313	1685	863	545	370
5	Slum	539	100	28	10	4	1
	Built-up	23,134	5284	2282	1251	776	534
	Non-built-up	7499	1427	584	250	144	85

Tile classification

Table A5. Accuracy for Bangalore tile classification during training, [%].

		Tile size, *N*			
Vocabulary size	Class	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px
10	Slum	70.86	80.08	85.24	83.33
	Built-up	67.53	72.89	79.89	78.47
	Non-built-up	67.44	79.6	84.98	79.86
20	Slum	71.59	83.89	89.06	90.97
	Built-up	67.56	75.55	81.18	81.94
	Non-built-up	69.15	79.92	87.53	90.27
50	Slum	71.91	86.9	91.35	94.44
	Built-up	68.51	78.89	86.26	84.72
	Non-built-up	68.55	83.57	89.82	90.27

Table A6. F1 score for Bangalore tile classification during training.

		Tile size, *N*			
Vocabulary size	Class	10 m 67 px	20 m 134 px	30 m 200 px	40 m 268 px
10	Slum	0.60	0.73	0.79	0.78
	Built-up	0.23	0.53	0.71	0.69
	Non-built-up	0.61	0.69	0.74	0.64
20	Slum	0.61	0.77	0.79	0.78
	Built-up	0.28	0.59	0.73	0.734
	Non-built-up	0.62	0.71	0.80	0.75
50	Slum	0.62	0.8	0.84	0.91
	Built-up	0.26	0.65	0.79	0.77
	Non-built-up	0.62	0.77	0.84	0.86

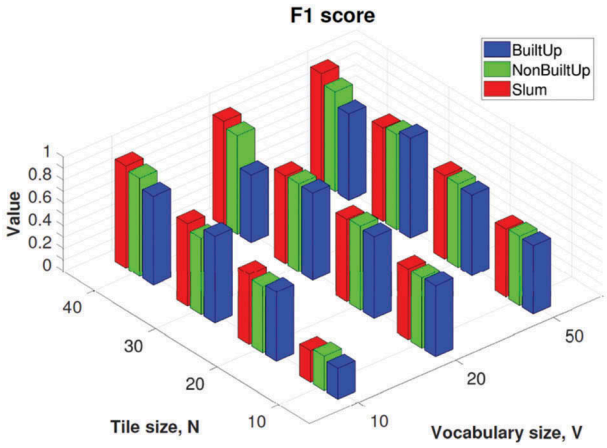

Figure A2. F1 score performance of the classifier on the Bangalore test tile datasets.

Segmentation Results

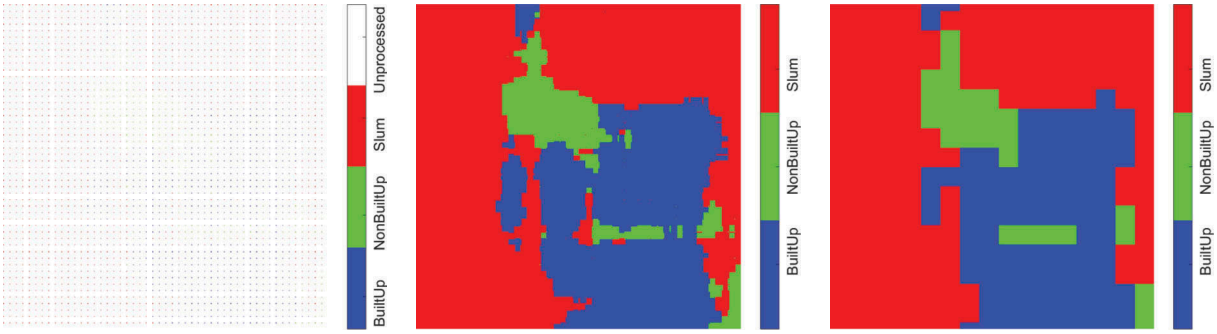

Figure A3. Pixel-level segmentation of Bangalore ROI2 (zoom). Left: Initial class label via classification of a tile with size 40m centred around these pixels; Middle: Assignment of labels to all unprocessed pixels with window size 22; Right: Final segmentation – all pixels within each window of size M = 30 get the majority label.

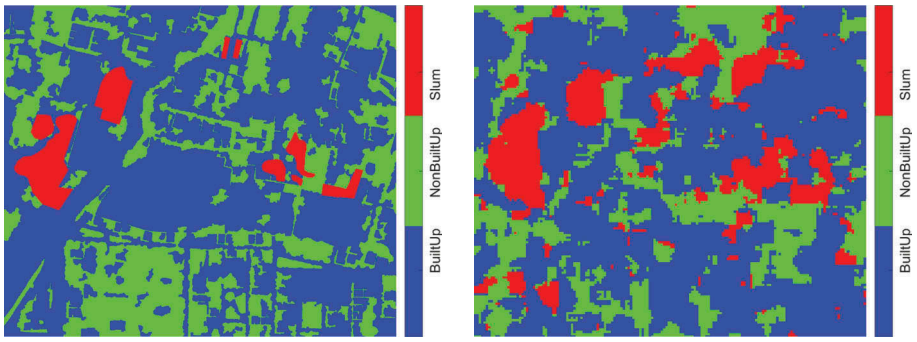

Figure A4. Bangalore ROI1: Left: ground truth and Right: result from pixel-level segmentation.

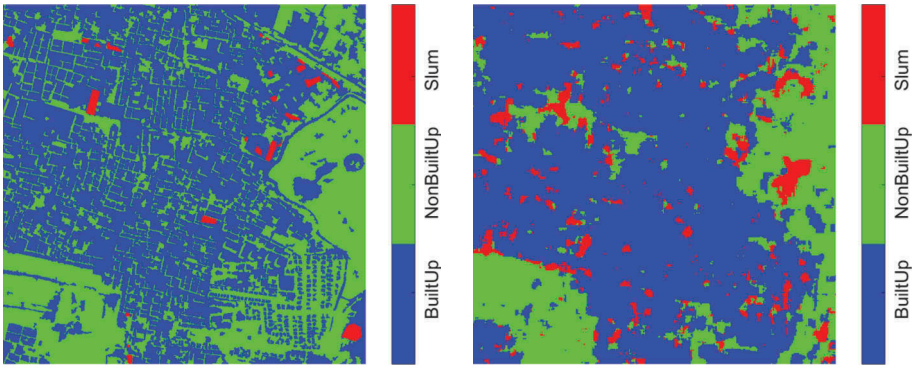

Figure A5. Bangalore ROI4: Left: ground truth and Right: result from pixel-level segmentation.

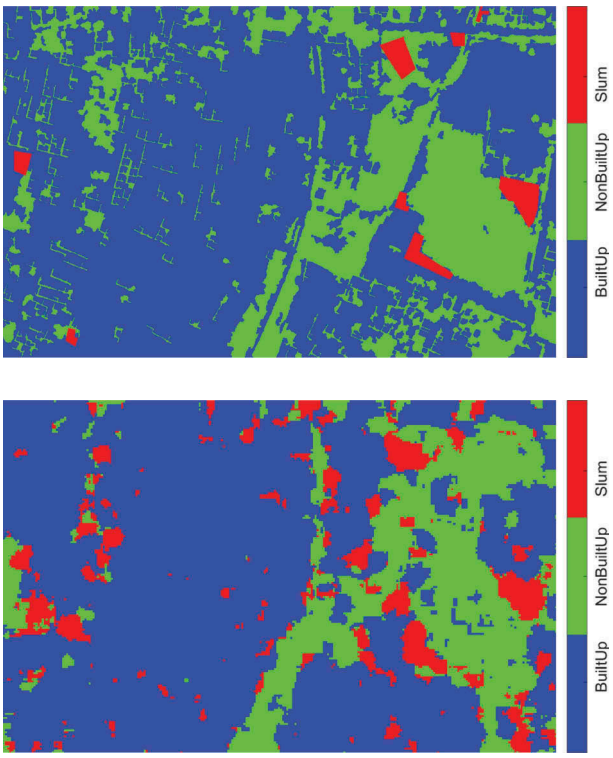

Figure A6. Bangalore ROI3: Top: ground truth and Bottom: result from pixel-level segmentation.