IDENTIFICATION OF SLUM SETTLEMENTS USING LOGISTIC REGRESSION

Vaishnavi Upadhyay (1), Ojasvi Saini (2), Kamal Pandey (2), Ashutosh Bhardwaj (2)

[1] Banasthali Vidyapith, Banasthali, Rajasthan, 304022, India
[2] Indian Institute of Remote Sensing, Dehradun, Uttarakhand, 248001, India
Email: vaishnaviupadhyay1299@gmail.com; ojasvi@iirs.gov.in; kamal@iirs.gov.in; ashutosh@iirs.gov.in

KEYWORDS: Remote Sensing, Slum Settlements, Informal Settlements, Spectral, Textural, Structural Features.

ABSTRACT: Identifying slum settlements in urban areas is a very important step in the process of the formulation of environment-friendly government policies. There are several methods, which can be used for the identification and delineation of slum areas. Field surveys are time-consuming as well as costlier for this purpose. Whereas remote sensing with the integration of image processing techniques provides an easy, efficient, and quick demarcation of the slum settlements. The presented study highlights a machine learning (ML) logistic-regression based method for the identification of the slums (informal settlements) using remote sensing datasets in Agra city, Uttar Pradesh, India. Agra being a tourist hotspot with the presence of the Taj Mahal needs proper planning for the urban development and improvement or relocation of slum dwellers by way of better future opportunities along with good living conditions. The method utilizes the spectral, textural, and spatial features, which are extracted from openly accessible high-resolution imagery from google earth. The algorithm delivers the confusion matrix performance score of 0.74.

1. INTRODUCTION

Urbanization has led to the uncontrolled growth of informal settlements in developing metropolitan areas across the world. The identification of the slums is a challenging task because of their place to place varying definitions across the world (Mahabir et al., 2016). A globally and most commonly used definition by UN-Habitat defines a slum by the deficiency of one or more of the following: Stable and durable housing, sanitized and sufficient area for living, access to clean and pure water, access to healthy and safe food, and security of tenure (Nolan, 2015). The identification of slums is a very important aspect in urban environments of major and metropolitan cities. The marginal condition of the slum dwellers, which day by day are losing their mental and physical health, needs to be highlighted. Moreover, the youngsters and the small kids who can prove to be the brightest minds and future of the country are deprived of access to a healthy environment and needful resources resulting in degradation of their overall health. The information on location, boundaries, and population in the informal settlements is of the great need for socio-economic studies and therefore will brief us with beneficial and important information for a healthy and safe environment along with the positive growth and development of the slum dwellers (Basiago, 1451). The status and lifestyle of mass living in the informal settlements, the access and availability to sanitized, sufficient, and affordable housing with the everyday essentials and important daily needs and resources has become one of the aims in achieving a healthy environment for sustainable growth and development. At present, about one-quarter of the world's total urban mass is living in informal settlements which are defined by the United Nations (UN) as areas lacking in access to safe water, sanitation, and stable housing (Kuffer et al., 2016). Access to fresh and clean water for the population living in these areas is a major goal of government according to the United Nations (*HABITAT III ISSUE PAPERS 7 – MUNICIPAL FINANCE*, 2015). Informal Settlements are most commonly featured by crowded areas with not so well-built buildings with irregular spatial data and an almost absence of vegetation while formal settlements comprise a presence of vegetation, well settled and maintained buildings, and a regular spatial piece of data (Kuffer et al., 2014) (Gueguen & Hamid, 2015). The employment factors give importance to a particular area that automatically attracts the population of rural areas. Due to this reason a mass migration of people taking place from rural to urban areas. The lack of money and resources compel these migrated people to live in slums and unorganized settlements. According to the 11th goal (Sustainable Cities and Communication) under Sustainability Development Goals (SDG), about one billion people are living in informal settlements of big and metropolitan cities.

VHR optical data acquired by GeoEye-1, IKONOS-2, QuickBird-2, and WorldView-3, available through Google Earth coupled with the image processing algorithms can be utilized for the identification of different areas from all over the world (Tapete & Cigna, 2019) and also gives us the benefit to differ informal settlements from well-settled places or formal settlements based on the different characteristics of urban areas and the urban setting or infrastructure. VHR images, obtained from google earth are capable of providing structural, spectral, and textural information of the ground targets and can be efficiently used in the identification of various land use and land covers (LULC) (Hu et al., 2013) (Vatsavai et al., 2013). Spatial data points to the spatially arranged spectral data in the image and the contextual information describe the information that is extracted from a neighborhood. The spatial, contextual, and textural information is very important to improve the ultimate classification and minimize the errors from satellite images (Mboga et al., 2017) (Shekhar, 2019). Detecting the informal settlements can be taken as a positive land cover classification problem because it needs higher classes with a good level of semantic abstraction. In the presented study the ML Logistic regression classification algorithm is used to assign urban LULC classes as slum or non-slum classes. Linear regression outputs continuous numerical values. Whereas, logistic regression utilizes logistic sigmoid function to converts its output to a probability value (Lavanya & Pandey, 2007).

2. LITERATURE REVIEW

According to the study conducted by Olanrewaju and Akinbamijo (2002) (Owoeye & Omole, 2012), the surroundings have great and obvious effects on health. As poor dwellers are often seen at zones, it is an indication of their poor health. slum residents observed by a group of people who live in unsettled houses with bad infrastructures situated in poor surroundings where they are uncovered to ill health mental as well as physical health from polluted and unsafe water. Moreover, unhygienic conditions are also responsible for such poor health. Such houses are without washrooms and other daily use essentials. Their drains are often covered with refuse deposits and other unhygienic materials, which enables the free flow of runoffs. Informal settlements in developing cities have a large extent and in some cases, they are also forming the major urban land use. Slum settlements are growing fast and can sometimes be scattered within the no slum areas (Sliuzas, 2004). Slum Settlements have specific features depending on the geographical area. Morphological attributes differentiate between formal and informal settlements are unplanned areas and planned areas. Unplanned areas comprise of high densities, the organic layout structure, lack of public(green) spaces in the vicinity of the residential areas, and small sub-standard building sizes. Whereas planned areas consist of low moderate density areas, regular layout patterns, presence of public (green) spaces mainly herbs and the shrubs within or vicinity of residential areas, and generally larger building sizes. Unplanned or slum settlements form part of urban residential land. Moreover, it is a fact that there are different definitions of what constitutes and unplanned settlements, and this is also dependent on locality. And now it has become very important to stop the increase of slum settlements as the future of the country also matters. There is a great need to tour the use of geo-information technologies to enable the mapping of the physical state of slum settlements. Many studies have been conducted to map informal settlements from high-resolution satellite imagery. However, the authors have focused only on the morphological attributes in their try to define such unplanned settlements.

3. STUDY AREA

The selected city for the transferability of this approach is Agra. It is situated in southwest Uttar Pradesh. It stretches across 26º44'N to 27º25'N and 77º26'E to 78º32'E. The city faces different types of atmospheric and weather conditions as it represents the semi-arid climate that borders on a humid subtropical climate. The city also possesses mild winters, summers that are hot and dry, and one of the most important monsoon seasons. It also represents different customs as well as cultures and also different building elements. Agra is a city on the banks of the river Yamuna in the Indian state of Uttar Pradesh. It is governed by Municipal Corporation which comes under Agra metropolitan region. The Agra city is located in the Uttar Pradesh State of India. As per provisional reports of Census India, the population of Agra 2011 was 15,74,000. Tajganj is about 450 years old informal settlement and amongst oldest of the Agra city. It spreads along the southern-east edge of the city situated at the south side of the river Yamuna. Tajganj includes 15 slums distributed in 3 wards covering 2725 houses with a population of 18137. The average household size of 6.7 is higher than the city average of 6.08. In 2008, the CDP estimated the Tajganj slum population to be 20% of the total ward population. Recent slum surveys for Slum Free City Plan suggest that the population may be much higher at 35% (Kulshresth & Verma, 2017). The study site (Tajganj slum settlement) is shown in figure 1.

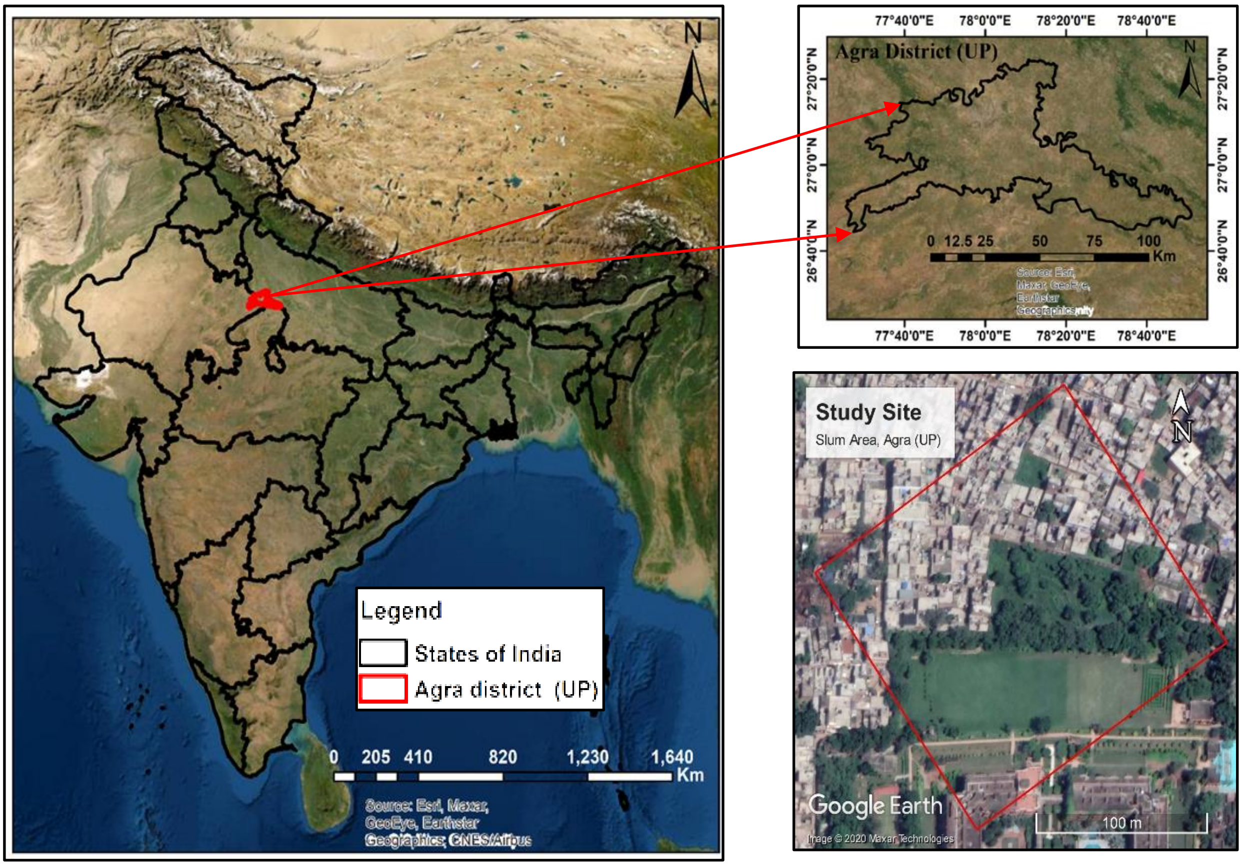

__Figure 1__. Location of study site.

4. METHODOLOGY

4.1 Feature Extraction

The overall methodology adopted in the study is shown in figure 2. The feature extraction was performed in Envi 5.3. It is an efficient GIS image analysis software that can be used for extracting features from the image. The example-based feature extraction method was performed using this software. Different image derived attributes can be used to differentiate informal settlements from the formal settlements. Textural and spatial image derived attributes can be used to differ in slum areas from no slum areas. Google Earth image was used to extract data. Three sets of variables were calculated for image feature extraction namely spectral variables, textural variables, and spatial (structural) variables. The image features were extracted from the images by using the example-based feature extraction in which training data is selected to assign objects to unknown identity to known features. Spectral features provide data regarding color and appearance of different elements in the image while structural features provide a piece of data regarding the arrangement of different elements in the image. The infrastructure of the slum areas can be defined with more crowded places ad mess up pattern than for the formal settlements and well-maintained places. Texture and spatial features can prove to be helping to differentiate Informal settlements from the formal ones.

Spectral Features

These provide the details about the statistics values per pixel. These features brief us with the data regarding the spectral functionalities of objects, which always differs for different places across the world and also for different types of land, vegetation, soil constitution, and also differs for a different pattern of buildings, etc. The majority statistic is extracted within this feature. The attributes which were calculated are Mean, Max, Min, Std for each RGB band, and therefore major statistic from the image is extracted from this set of features. These attributes of spectral features are not difficult to grab and also are capable of providing better performance about this feature and also provide differences across the places. Attributes like Mean and Standard deviation provide much more clear statistics than other spectral attributes like Max and Min.

Texture Features

Textural features are calculated on each band of the target image. The attributes computed under texture features are Range, Mean, Variance, and Entropy. Calculation of these attributes comprises a two-step process wherein the first step a square Kernel of defined size is applied to the band of the input image. The attributes are computed per pixel in the Kernel window and the computed result is processed to the center Kernel pixel. In the second step of the two-step process, the results of the attributes are averaged per pixel in the segment to create the value of the attribute for the segmentation label of that particular band.

Spatial Features

Spatial features are also called structural features which help determine the arrangement of the elements in the input image. The different spatial attributes calculated are Area, Length, Compactness, Convexity, Solidity, Roundness, Form_Factor, Elongation, Rectangular_Fit, Main_Direction, Major_Length, Minor_Length, Number_Of_Holes, Hole_Area/ Solid_Area. Structural attributes are also calculated for each RGB band. Image variables are shown in Table 1.

Table 1. *Image derived variables.*

Variables	Attributes	Description
Spectral Variables	Mean	Mean value of the pixels
	Max	Maximum value of the pixels
	Min	Minimum value of the pixels
	STD	Standard deviation value of the pixels
Texture Variables	Range	Average data range of the pixels comprising the region
	Mean	Average value of the pixels
	Variance	Average variance of the pixels
	Entropy	Average entropy value of the pixels
Spatial Variables	Area	Total area of polygon minus area of holes
	Length	The combined length of all the boundaries of polygon
	Compactness	A shape measure that indicates the compactness of polygon
	Convexity	A shape measure that indicates the convexity of polygon
	Solidity	A shape measure that compares the area of polygon to the area of convex hull surrounding the polygon
	Roundness	A shape measure that compares the area of the polygon to the maximum diameter of the polygon
	Form_Factor	A shape measure that compares the area of the polygon to the square of total perimeter
	Elongation	A shape measure that indicates ratio of major axis to the minor axis of the polygon
	Rectangular_Fit	A shape measure that indicates how well the shape is described by the rectangle
	Main_Direction	The angle subtended by the major axis and the x-axis
	Major_Length	The length of the major axis of an oriented bounding box enclosing the polygon
	Minor_Length	The length of the minor axis of an oriented bounding box enclosing the polygon**.**
	Number_of_holes	The number of holes in the polygon
	Solid_Area	The ratio of total area of the polygon to the area of the outer contour of the polygon

After the image feature extraction using ENVI 5.3 as shown in Figure 3, the next and the most important step is to create the dataset. In this step, a ground truth sample was selected for the city. The polygons of the chosen samples are manually labeled as slum or no-slum. The prior available data and studies were used to identify informal settlements in each area of the city. The areas having informal settlements can be considered with some characteristics, but it can show different appearances depending on the contextual features. However, each country and each city can differ in the definition of slums. The location of slum areas in Agra is identified on the “tciurbanhealth” website. It provides interactive and updated information about the slums in Agra. The slum areas were selected based on a specific appearance. The no slum areas in the city included high- and low-rise areas with well-maintained buildings and different areas with malls, transport facilities, parks, green areas filled with herbs and shrubs, and forests areas. The used technique is known as a binary classification scheme and is also the most common practice in remote sensing for identifying informal settlements. The obtained image derived variables majorly textural variables and spatial variables can differentiate better between slum and no slum areas.

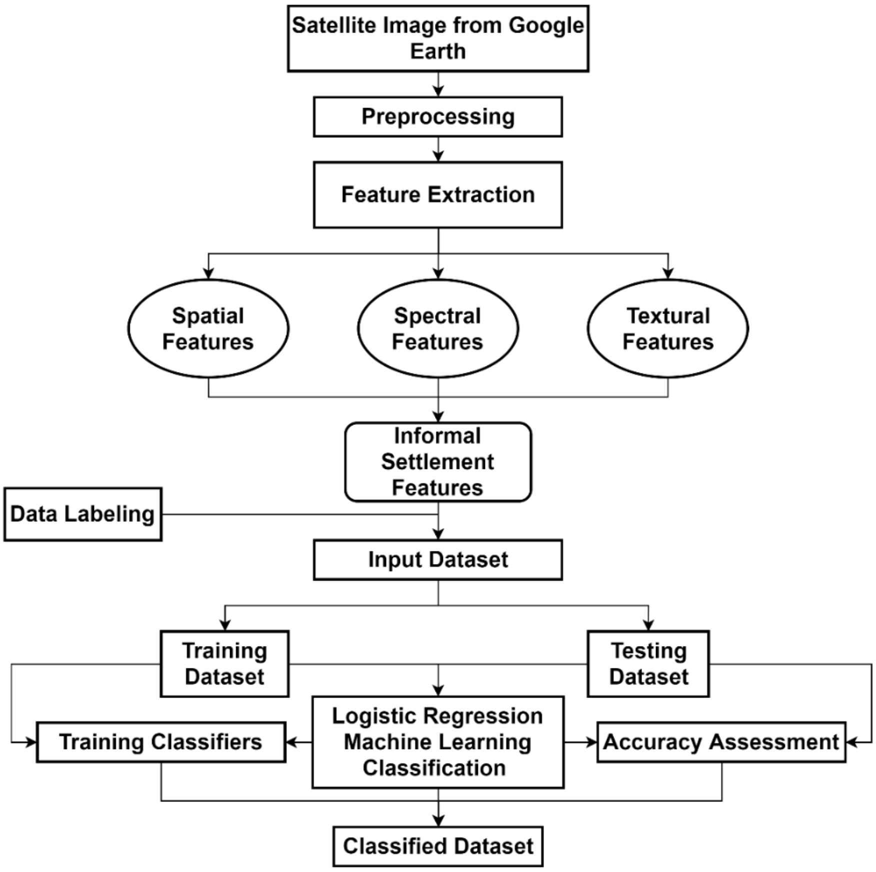

Figure 2. *Flow Diagram of the proposed approach for slum detection.*

The next step is to choose the classification model to train the dataset. Numerous algorithms had been proposed overtime for the slum area detection. The primary aim is to develop a classification method that has the capability of knowing and also possesses the capability of generalizing a relationship between independent variables and dependent variables. The independent variable is a set of variables (X) called the matrix of features and the dependent variable called the categorical variable (Y). For a particular classification problem, X is a matrix that contains the computed values of spectral, textural, and structural (spatial) features of the area in the image and Y is a dependent or the categorical variable that will predict the value either 1 or 0 if the specific area is a slum or having informal settlements or not respectively. There are two types of classifiers Linear and Non-linear classifiers and there are two factors that define the ability of classification models. Firstly, the definition of the classification boundary of the particular classifier (it can be linear or non-linear). Secondly to what extent the data possess complexity. The Internal complexity and also the extent of complexity cannot be determined easily or understood, specifically for high dimensional data. The most instinctive method to grab the complexity of the data is by analyzing the respective features and their classes. This way is generally confined to low dimensional data. However, it becomes very important to test the feasibility to find strong tools for slum detection in different parts of the world with good performance in different surroundings.

The performance of the logistic regression classification model is determined which is present in the Scikit-learn library. Logistic regression is the most commonly used linear classifier and is used by different policymakers as well. It is a mathematical way whose general aim is to make use of the logistic function to predict the dependent variable (Y) given the matrix of feature or the independent variable (X). For this classification model, we use the sigmoid function, which is more specifically defined as an activation function. This function fixes the output range between

0 and 1 which in return becomes helpful in determining the prediction or the probability of the output. training sets and test sets. The classifiers were trained and then the confusion matrix was made to find the accuracy assessment.

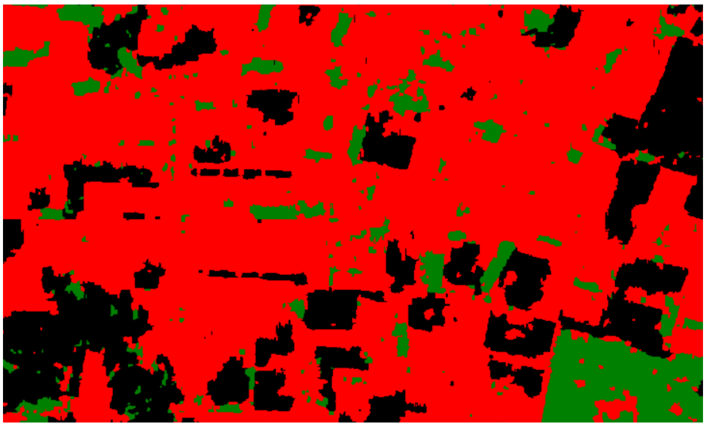

Figure 3. *Slum feature extraction from image using ENVI 5.3.*

5. RESULTS AND DISCUSSION

The results obtained from the logistic regression classification model suggest that the distributions of all image-derived variables are significantly different for slum and non-slum areas. The major variables which contribute to differentiate between slum and non-slum-areas are texture (Range, Mean, Variance, Entropy) and spatial (Mean, Max, Min, STD) features. The confusion matrix accuracy derived from the test and validation data utilizing the ML Logistic Regression model is found to be 0.74. The confusion matrix amongst the slum and non-slum classes and classification report for logistic regression ML method are shown in figure 4 and 5 respectively. The macro average and weighted average for the logistic regression classification are found to be 0.72 and 0.72 respectively with precision values of 0.73 and 0.74 respectively. The statistics generated from the input spatial, spectral and textural variables for slum features are listed in table 2, 3 and 4 respectively.

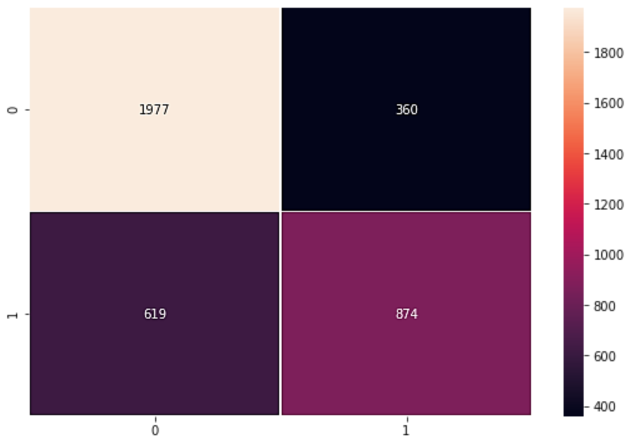

Figure 4. *Confusion Matrix accuracy analysis.*

	precision	recall	f1-score	support
no	0.76	0.85	0.80	2337
yes	0.71	0.59	0.64	1493
accuracy			0.74	3830
macro avg	0.73	0.72	0.72	3830
weighted avg	0.74	0.74	0.74	3830

Figure 5. *Classification report for logistic regression ML model.*

Table 2. *Statistics for spatial variables of slum features.*

	Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10	Col.11	Col.12	Col.13	Col.14
Mean	969.40	174.30	0.21	1.19	0.83	0.50	0.49	1.78	0.63	83.56	13.18	7.63	1.58	0.99
Median	16	20	0.22	1.12	0.84	0.46	0.50	1.57	0.56	90	6.52	4.18	0	1
Minimum	2	6	0.02	1	0.25	0.06	0.00	1	0.16	0	2	1	0	0.68
Maximum	4815775	705190	0.28	52.11	1	1.27	0.78	9.23	1	180	4867.46	2427.15	8651	1

Representation of columns in table 2 are following:

*Col.1- **AREA**, Col.2- **LENGTH**, Col.3- **COMPACTNESS**, Col.4- **CONVEXITY**, Col.5- **SOLIDITY**, Col.6- **ROUNDNESS**, Col.7- **FORM_FACTOR**, Col.8- **ELONGATION**, Col.9- **RECTANGULAR_FIT**, Col.10- **MAIN_DIRECTION**, Col.11- **MAJOR_LENGTH**, Col.12- **MINOR_LENGTH**, Col.13- **NUMBER_OF_HOLES**, and Col.14- **SOLID_AREA**.*

Table 3. *Statistics for spectral variables of slum features.*

	Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10	Col.11	Col.12
Mean	147.42	8.25	132.18	162.32	134.01	8.27	118.75	148.89	135.54	8.06	120.56	150.13
Median	145.83	6.35	129	160	132.73	6.31	117	146	134.67	6.13	120	148
Minimum	77.14	0	54	79	59.95	0	32	69	57.29	0	21	73
Maximum	240.59	46.28	222	255	216	48.10	207	246	216.82	56.01	201	249

Representation of columns in table 3 are following:

*Col.1- **MEAN** in Band 1, Col.2- **STANDARD DEVIATION** in Band 1, Col.3- **MINIMUM** in Band 1, Col.4- **MAXIMUM** in Band 1, Col.5- **MEAN** in Band 2, Col.6- **STANDARD DEVIATION** in Band 2, Col.7- **MINIMUM** in Band 2, Col.8- **MAXIMUM** in Band 2, Col.9- **MEAN** in Band 3, Col.10- **STANDARD DEVIATION** in Band 3, Col.11- **MINIMUM** in Band 3, Col.12- **MAXIMUM** in Band 3.*

Table 4. *Statistics for textural variable of slum features.*

	Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9	Col.10	Col.11	Col.12
Mean	16.05	147.41	42.08	-0.49	15.95	134.00	42.17	-0.49	15.74	135.53	40.87	-0.49
Median	14.89	145.76	31.04	-0.48	14.84	132.64	30.57	-0.48	14.67	134.38	29.46	-0.48
Minimum	3.14	78.10	1.52	-1.25	3.25	61.57	1.07	-1.08	3.44	59.35	1.64	-1.16
Maximum	60.5	238.60	922.53	-0.06	62.5	216.11	852.21	-0.06	68	215.65	921.08	0

Representation of columns in table 4 are following:

*Col.1- **RANGE** in Band 1, Col.2- **MEAN in Band 1**, Col.3- **VARIANCE** in Band 1, Col.4- **ENTROPY** in Band 1, Col.5- **RANGE** in Band 2, Col.6- **MEAN** in Band 2, Col.7- **VARIANCE** in Band 2, Col.8- **ENTROPY** in Band 2, Col.9- **RANGE** in Band 3, Col.10- **MEAN** in Band 3, Col.11- **VARIANCE** is the Band 3, Col.12- **ENTROPY** in Band 3.*

6. CONCLUSIONS

Understanding urban systems always remains a critical challenge for different planners and lawmakers. This study explored implementing a standardized method for slum areas using spectral, textural and spatial features which are extracted from openly accessible high-resolution imagery and accessed the capability of ML logistic regression to classify urban areas as slum or no slum. The aim of this paper is to contribute to the methods of urban planning by using ML logistic regression in identifying slum areas. The paper focused on answering two different research questions. First, how can the ML logistic regression be applied for the analysis of slum areas and second, how the different characteristics are responsible to recognize slum areas from no slum areas. The image-derived variables performed well in the city because the slum area in this city have a different spatial and texture pattern than slum areas and exhibit differences in different areas. The textural and structural variables are mainly responsible to differentiate between slum and no slum areas. A suggestion for future studies is to use algorithms for object and scene recognition on images to generate a new set of features that can enhance the performance of the algorithms.

REFERENCES

Basiago, A. D. (1451). Economic, social, and environmental sustainability in development theory and urban planning practice. In *The Environmentalist* (Vol. 19). Kluwer Academic Publishers.

Gueguen, L., & Hamid, R. (2015). Large-scale damage detection using satellite imagery. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 1321–1328. https://doi.org/10.1109/CVPR.2015.7298737

HABITAT III ISSUE PAPERS 7 – MUNICIPAL FINANCE. (2015). http://habitat3.org/wp-content/uploads/Habitat-III-Issue-Paper-7_Municipal-Finance-2.0.pdf

Hu, Q., Wu, W., Xia, T., Yu, Q., Yang, P., Li, Z., & Song, Q. (2013). Exploring the use of google earth imagery and object-based methods in land use/cover mapping. *Remote Sensing*, *5*(11), 6026–6042. https://doi.org/10.3390/rs5116026

Kuffer, M., Barros, J., & Sliuzas, R. V. (2014). The development of a morphological unplanned settlement index using very-high-resolution (VHR) imagery. *Computers, Environment and Urban Systems*, *48*, 138–152. https://doi.org/10.1016/j.compenvurbsys.2014.07.012

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from Space—15 Years of Slum Mapping Using Remote Sensing. *Remote Sensing*, *8*(6), 455. https://doi.org/10.3390/rs8060455

Kulshresth, V., & Verma, A. (2017). A Contrast to The Taj Mahal: Slums of Tajganj, Agra. *International Journal on Emerging Technologies*, *8*(1), 560–567. www.researchtrend.net

Lavanya, D., & Pandey, V. (2007). Code Development in Octave for Optimization Using Logistic Regression. *International Advanced Research Journal in Science, Engineering and Technology ISO*, *3297*. https://doi.org/10.17148/IARJSET.2018.5710

Mahabir, R., Crooks, A., Croitoru, A., & Agouris, P. (2016). The study of slums as social and physical constructs: challenges and emerging research opportunities. *Regional Studies, Regional Science*, *3*(1), 399–419. https://doi.org/10.1080/21681376.2016.1229130

Mboga, N., Persello, C., Bergado, J. R., & Stein, A. (2017). Detection of informal settlements from VHR images using convolutional neural networks. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111106

Nolan, L. B. (2015). Slum Definitions in Urban India: Implications for the Measurement of Health Inequalities. *Population and Development Review*, *41*(1), 59–84. https://doi.org/10.1111/j.1728-4457.2015.00026.x

Owoeye, J. O., & Omole. (2012). Built Environment Decay and Health Situation of Slum

Dwellers in Residential Cores of Akure, Nigeria. *American Journal of Human Ecology*, *1*(2), 33–39. https://worldscholars.org/index.php/ajhe/article/view/82/42

Shekhar, S. (2019). Effective management of slums- Case study of Kalaburagi city, Karnataka, India. *Journal of Urban Management*, *9*(1), 35–53. https://doi.org/10.1016/j.jum.2019.09.001

Sliuzas, R. V. (2004). *Managing informal settlements : a study using geo - information in Dar es Salaam, Tanzania*. https://research.utwente.nl/en/publications/managing-informal-settlements-a-study-using-geo-information-in-da

Tapete, D., & Cigna, F. (2019). Detection of archaeological looting from space: Methods, achievements and challenges. In *Remote Sensing* (Vol. 11, Issue 20, p. 2389). MDPI AG. https://doi.org/10.3390/rs11202389

Vatsavai, R. R., Bhaduri, B., & Graesser, J. (2013). Complex settlement pattern extraction with multi-instance learning. *Joint Urban Remote Sensing Event 2013*, *JURSE 2013*, 246–249. https://doi.org/10.1109/JURSE.2013.6550711