Chapter 37
Spatial Data Mining of Agricultural Land Area Using Multi-spectral Remote-Sensed Images

Parminder Kaur Birdi, Karbhari Kale, and Varsha Ajith

1 Introduction

Satellite sensors data is available for larger areas making it suitable for agricultural land monitoring. This data is acquired at frequent intervals which is useful to understand changes occurring in agricultural land areas. Due to the availability of satellite images covering large-scale areas, it is possible to design efficient agriculture land use monitoring systems by applying different spatial data mining methods [1]. The traditional data mining methods have been majorly focusing on transactional and relational databases. Remote-sensed data (spatial or geo-referenced) collected by various sensors record spatial relations present in the data. These systems record energy reflected at different wavelength ranges present in the electromagnetic spectrum. Satellite images have a huge amount of spatial information of different land covers recorded from the earth's surface. The traditional agricultural monitoring systems are dependent on extensive human efforts, and they usually cover smaller areas. This problem can be overcome using remotely sensed systems as the larger area is covered and human efforts and cost is also lowered. The temporal and spatial resolutions of the input data need to be considered appropriately.

P. K. Birdi (✉) · V. Ajith
Jawaharlal Nehru Engineering College, MGM University, N-6, CIDCO, Aurangabad, Maharashtra 431003, India
e-mail: dhingra.param@gmail.com

K. Kale
Department of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra 431003, India
e-mail: kvkale91@gmail.com

I. Jeena Jacob et al. (eds.), *Data Intelligence and Cognitive Informatics*,
Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-15-8530-2_37

Remote-sensed data are used from multi-spectral (MS), hyper-spectral (HS), and panchromatic sensors. Remote-sensed data from AVIRIS, MODIS, IRS-LISS-III, IV, SPOT-4,5, Landsat-7,8, ASTER, GF-2, IKONOS, and Sentinel-2 sensors are mostly used for agricultural land monitoring and crop growth analysis systems. It has been observed that spatial resolutions typically range from 250 to 5.8 m [2, 3].

Spatial data mining (SDM) has developed from the field of data mining and statistics. A large number of tasks including clustering, classification, association rule mining and data visualization is part of spatial data mining. SDM has a number of steps, wherein the selection of appropriate remote sensed images can be started with, followed by preprocessing and normalization of images, application of the spatial mining method for the classification task, visualization, and performance evaluation of results produced. Agricultural monitoring systems have been designed using spatial data mining methods [4]. There are two approaches to remote-sensed image classification; supervised and unsupervised. Maximum likelihood classifier (MLC), support vector machines (SVM), neural networks (NN), and decision tree (DT) are most commonly used supervised methods. DT and NN methods have been observed to generate higher classification accuracy in comparison with the traditional method, MLC. The performance of classifier depends on the training dataset created using handcrafted features. Finding an optimal decision tree design is a compute-intensive task, as the decision tree should have minimum overall classification error, appropriate depth of the tree, and an optimal subset of features to form decision rules.

In this study, a decision tree is constructed using rule-based modeling technique based on the foundation that different land use and land covers (LULC) have different spectral responses. The rules are formed using spectral signature values for the target land use and land covers. To improve classification accuracy and reduce time complexity, NDVI is also used in rule mapping. Rule-based mapping is done at pixel level. The spectral reflectance values and NDVI values of the training dataset are used as input.

2 Literature Review

Elodie et al. [5] used remote-sensed images from SPOT-5 sensor for mapping of cropland. Object-based supervised classification method is designed where expert rules are framed in the first step, followed by extraction of patterns from the dataset, and then, patterns are used to build classification rules. Naive Bayes, support vector machine, random forest, and decision tree methods are used for image classification. Remote-sensed images from Sentinel-2 sensor is also used for experiments. Overall accuracies for all methods are computed, and SVM is found to attain the highest accuracy of 84% as compared to other methods. It has also been concluded that data mining methods are suitable to process huge amounts of data [5]. Schultz et al. [6] used temporal images from Landsat-8 to classify study scene into different crop classes. An automated supervised technique is designed using random forest (RF) classifier, where features for segmentation step are automatically selected reducing

human efforts. The proposed method produced a satisfactory overall accuracy for classification [6]. Many studies show that decision trees achieve higher classification accuracies as compared to classifiers like MLC, neural network as the time required to train a DT is less. Decision tree classifier has been applied for cropland monitoring and mapping. Some other studies illustrate the improvement in classification accuracies when spectral signatures are integrated with derived features including texture values and different vegetation indices [2, 3, 7]. In another study by Lebourgeois et al. [8] remote-sensed data from SPOT-5, Landsat-8, and PLEIADES sensors are used for agriculture land mapping. In this study, random forest classifier is applied to produce land cover maps at different levels into two classes and further into 25 crop subclasses. Random forest classifier is optimized by reducing variable count and this, in turn, resulted in less computing time [8]. Gervais et al. [9] carried out experiments using Landsat-5 and MODIS datasets to study the influence of urbanization on the growth life cycle of plants. NDVI values were computed and used to extract three seasons of the vegetation, starting, end, and length of the season. The performance of the experiments was found to be satisfactory [9].

From different research studies, it has been observed that the selection of appropriate features plays a major role in attaining higher classification accuracies using decision tree classifier.

3 Study Area

For conducting experiments, study scene is selected by considering the presence of different land covers with an abundance of sugarcane crop at different growth stages and availability of remote-sensed images. Navin Kaygaon is the study scene, close to Aurangabad district of Maharashtra state and can be located at 19° 20′ 1.4″ to 20° 15′ 19.65″ North and 74° 30′ 43.61″ to 75° 5′ 67″ East [10]. Figure 1 shows the study area acquired from LISS-IV sensor.

The land covers present are waterbody, barefarm, road, settlement, and sugarcane crop at different growth stages.

4 Data Used

Remote sensing data is collected using satellite sensors, and the appropriate sensor is selected based on its spatial, spectral, and temporal resolution. The spectral reflectance values recorded by the sensor can be related to its phenological growth cycle of sugarcane crop. Different stages of the life cycle of a crop are its sowing time, budding, ripening, and harvesting. Spatial resolution, i.e., the size of a pixel, and the number of pixels present in a typical a sugarcane farm also affects the classification accuracy. Tables 1 and 2 show detailed specifications remote-sensed data used for different experiments.

Fig. 1 Satellite image of study area represented by LISS-IV data

Table 1 Landsat-8 OLI sensor data specifications

Mode	Band	Spectral region	Spectral resolution (μm)	Spatial resolution (m)
Panchromatic	Band 8	Panchromatic	0.50–0.68	15
Multispectral	Band 1	Visible	0.43–0.45	30
	Band 2	Visible	0.450–0.51	
	Band 3	Visible	0.53–0.59	
	Band 4	Red	0.64–0.67	
	Band 5	Near-infrared (NIR)	0.85–0.88	
	Band 6	SWIR 1	1.57–1.65	
	Band 7	SWIR 2	2.11–2.29	

Table 2 LISS-IV data specifications

Mode	Band	Spectral region	Spectral resolution (μm)	Spatial resolution (m)
Panchromatic	Single	Panchromatic	0.50–0.75	5.8
Multispectral	Band 2	Visible	0.52–0.59	5.8
	Band 3	Visible	0.62–0.68	
	Band 4	Near-infrared (NIR)	0.77–0.86	

In situ data is also collected for precise remote-sensed image interpretation. For information related to crops cultivated, the growth cycle of crops was collected with the help of local farmers who owned farms of the study area. GPS enabled handheld devices were used to record geographical coordinates of a total of 150 sugarcane farms which ranged in size from half an acre to sixteen acres.

5 Proposed Methodology and Experimentation

This section describes the proposed research methodology used to classify multi-spectral remote-sensed images. Spatial data mining method decision tree has been used as a classifier here [11]. For experiments, multistage decision tree has been designed by applying rules created using band spectral reflectance values and NDVI. Figure 2 shows the proposed flow diagram of the proposed methodology. Since dataset specifications are different, separate decision trees are developed for LISS-IV and Landsat-8 images. According to Tucker et al. [12] and Schmidt et al. [13], physical characteristics of crops change with different growth stages and can be better interpreted using vegetation indices (VI) [12, 13]. The most recommended vegetation index, NDVI, has been used in experiments. NDVI represented by Eq. 1 is calculated as the ratio between red and NIR bands and spectral reflectance values [14, 15].

$$\text{NDVI} = \frac{R_{\text{NIR}} - R_{\text{Red}}}{R_{\text{NIR}} + R_{\text{Red}}} \tag{1}$$

5.1 LISS-IV and Landsat-8 Dataset Pre-processing

Since satellite images are not directly usable due to radiometric, geometric, and atmospheric errors present in them. LISS-IV dataset is acquired from IIRS and images that

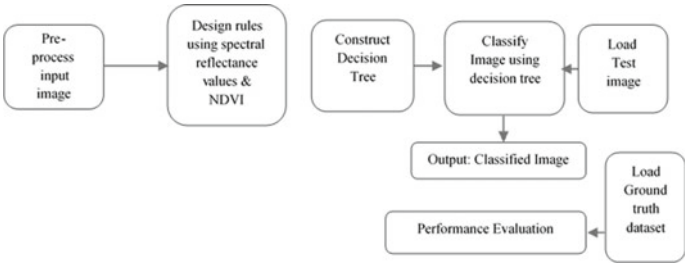


Fig. 2 Flow diagram of the proposed methodology using decision tree classifier

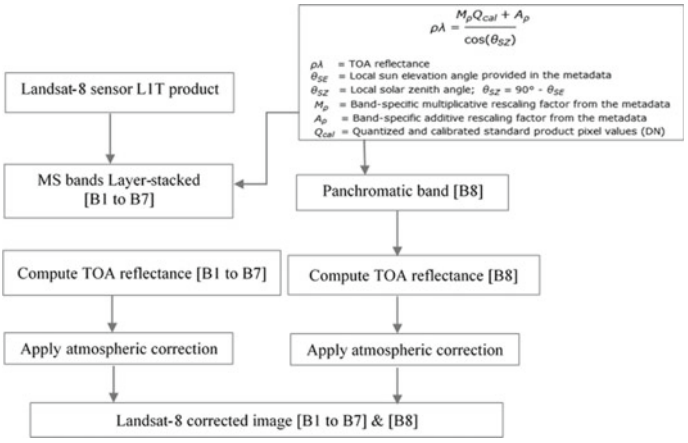


Fig. 3 Landsat-8 dataset preprocessing

are geometrically and radiometrically corrected. So no preprocessing is performed on them. The study area is cropped and layer stacked for further use. Landsat-8 images are preprocessed after acquiring them in L 1T format. The steps performed to preprocess them are shown in Fig. 3. The preprocessing is carried out using ENVI software by applying quick atmospheric correction method on all the images.

5.2 Experimentation

An exponential number of decision trees can be constructed from a given set of attributes or features, where some of the trees can be more accurate than others. Finding an optimal decision tree is computationally infeasible as the search space is exponential in size. Various decision trees were built using trial and error. A widely used decision tree algorithm C4.5 is applied to generate decision trees since reaching an optimum tree is time consuming. A greedy algorithmic model is used to overcome this, where at every level a locally optimal decision is made to construct the decision tree. The set of features are used to partition the data. A training set is constructed by investigating the existing data, where a class label is assigned to every pixel. A recursive algorithm is developed by making a combination of feature values from the training data. This algorithm works fine if every feature combination has a class label assigned to it. The alternate condition can arise for a case, where for some feature combinations no training data gets associated with the identified class labels.

Once a decision tree is constructed, classifying a test data is fast as worst-case complexity is $O(d)$, where d is the maximum depth of the tree. Maximum depth of the decision tree reached a depth of 16, generating 65 leaf nodes and a total of 128 nodes for Landsat-8 image with four classes and seven attributes. The total time

taken is 9 min and 19 s with an accuracy of only 50% on test data. The minimum depth as indicated by Landsat-8 decision tree is three and four for LISS-IV image.

A minimum of 50 training pixel values (for every target class) are considered from all spectral bands. The mean values are computed along with standard deviation and further used to frame the rules for decision making. Four land covers as target classes could be selected for Landsat-8 image due to fewer number of training pixels for road and barefarm, whereas six land covers were selected for LISS-IV image classification.

Table 3 shows minimum (min), maximum (max), mean, and standard deviation (SD) values obtained from regions of interest marked as a training dataset for Landsat-8 image. Similarly, Table 5 shows spectral reflectance distribution in all bands for the target classes of LISS-IV image. The two most separable group of classes are processed first, and the subtlest class pair will be processed last. This helps in reducing cumulative error. To ensure better accuracy, error generated at every decision node needs to be reduced by making use of optimal or sub-optimal features for portioning of the data, as given in Table 4.

The highest class separability is shown by water from Crop-Hv and Crop-Gr with a value of 2.0. So, for this design at level 1, water is separated from other land covers using Band 7 (spectral reflectance values are less than 0.048). It has no overlap with other class range. For the next decision rule, class 'Others' show the next best TD value w.r.t to Crop-hv and Crop-Gr. The optimal decision tree for Landsat-8 image is shown in Fig. 4a.

The best features selected are considered for constructing final decision trees for LISS-IV image as shown in Fig. 4b.

6 Results and Interpretation

The performance of decision tree classifier has been evaluated using a confusion matrix. The confusion matrix shows the relationship between the predicted values and actual (ground truth) values. The confusion matrix is further used to calculate overall classification accuracy which is defined as the ratio between correctly predicted values and a total number of predicted values. Kappa coefficient is also computed to measure the agreement between predicted values and actual values, where a zero coefficient value means no agreement and one means complete agreement. With the increase in kappa value, classification accuracy becomes better. The same images have also been classified using maximum likelihood classifier (MLC) for comparison of results, and classified images are shown in Figs. 5 and 6.

Table 3 Spectral reflectance distribution in all bands for Landsat-8 data

	Crop-Hv			Crop-Gr			Water			Others		
	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD
B1	0.131	0.145	0.139 ± 0.003	0.132	0.138	0.132 ± 0.001	0.126	0.136	0.129 ± 0.001	0.128	0.228	0.146 ± 0.011
B2	0.112	0.132	0.123 ± 0.004	0.113	0.125	0.116 ± 0.001	0.105	0.117	0.109 ± 0.002	0.108	0.226	0.132 ± 0.015
B3	0.097	0.123	0.110 ± 0.005	0.102	0.116	0.107 ± 0.003	0.081	0.102	0.088 ± 0.004	0.085	0.235	0.122 ± 0.029
B4	0.086	0.137	0.113 ± 0.008	0.081	0.107	0.090 ± 0.005	0.063	0.085	0.071 ± 0.004	0.082	0.257	**0.130 ± 0.029**
B5	0.151	0.329	0.185 ± 0.038	0.279	0.404	0.341 ± 0.022	0.059	0.095	0.068 ± 0.005	0.124	0.372	0.211 ± 0.058
B6	0.141	0.203	0.166 ± 0.012	0.141	0.194	0.151 ± 0.010	0.035	0.062	0.041 ± 0.004	0.105	0.376	0.190 ± 0.045
B7	0.083	0.164	0.126 ± 0.013	0.067	0.105	0.081 ± 0.007	0.023	**0.044**	**0.027 ± 0.003**	0.089	0.329	0.141 ± 0.035

Bold values represent attributes used to form rules for decision tree for Landsat-8

Table 4 Best features for class pairs of Landsat-8 data used in optimal DT

Class pairs	No. of features selected	Best feature selected
Crop-Hv + Crop-Gr	NDVI, B5	NDVI
Crop-Hv + Water	B6, B7	B7
Crop-Gr + Water	B6, B7	B7
Crop-Hv + Others	B5,B4	B4
Crop-Gr + Others	B7, B4	B4
Water + Others	B5, B6, B7	B7

The results show that DT classifier has higher classification accuracy as compared to MLC as optimal features are selected to form classification rules while constructing a decision tree. The first classification rule consists of spectral features which have the highest separability, measured using transformed-divergence matrix. Further classification rules also follow the same selection criterion. Another factor which resulted in improvement of accuracy is the inclusion of NDVI values for classification of sugarcane crop at harvest and grown stage. Table 6 shows the classification results obtained on both the datasets. Since the spatial resolution of LISS-IV dataset is 5.8 m, user's and producer's accuracy (UA, PA) for six classes is found to be better than Landsat-8 dataset as per results obtained by Kaur and Kale [16].

7 Conclusions and Future Scope

Traditionally, agricultural land monitoring is being done using manual methods which are time inefficient and tedious. The work done using manual surveys is also error prone and results in inaccurate data recording. Satellite images are found to be appropriate for this task as they cover larger areas, and availability of periodic information makes them beneficial for doing the task of agricultural land monitoring automated. While performing spatial data mining methods such as decision trees, the accuracy of supervised classification methods depends on correct labeling of training samples and also a suitable amount of training samples. This study shows that multi-spectral images from Landsat-8 and LISS-IV sensors are suitable to carry out this task of agricultural land monitoring using a decision tree. As new techniques such as deep learning have also been used for similar work, the task can be performed using deep convolutional neural networks for more accurate results.

Table 5 Spectral reflectance distribution in all bands for LISS-IV data

	Crop-Hv			Crop-Gr			Water			Settlement			Road			Barefarm		
	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD
B1	0	168	46.58 ± 25	0	205	45.86 ± 19.9	0	155	1.7 ± 9.15	24	255	221 ± 38.05	15	177	60.7 ± 20.8	0	255	160 ± 38.4
B2	0	151	**33.60 ± 19**	**23**	255	**165 ± 40.6**	0	255	0.8 ± 7.94	53	255	**155 ± 33.7**	1	40	**20.24 ± 10.4**	31	255	151.36 ± 48
B3	0	176	64.40 ± 22.5	0	255	26.0 ± 17.2	0	133	**1.17 ± 6.82**	29	255	208 ± 43.5	12	126	58.42 ± 16.6	28	255	164.5 ± 37.2

Bold values represent attributes used to form rules for decision tree for LISS-IV

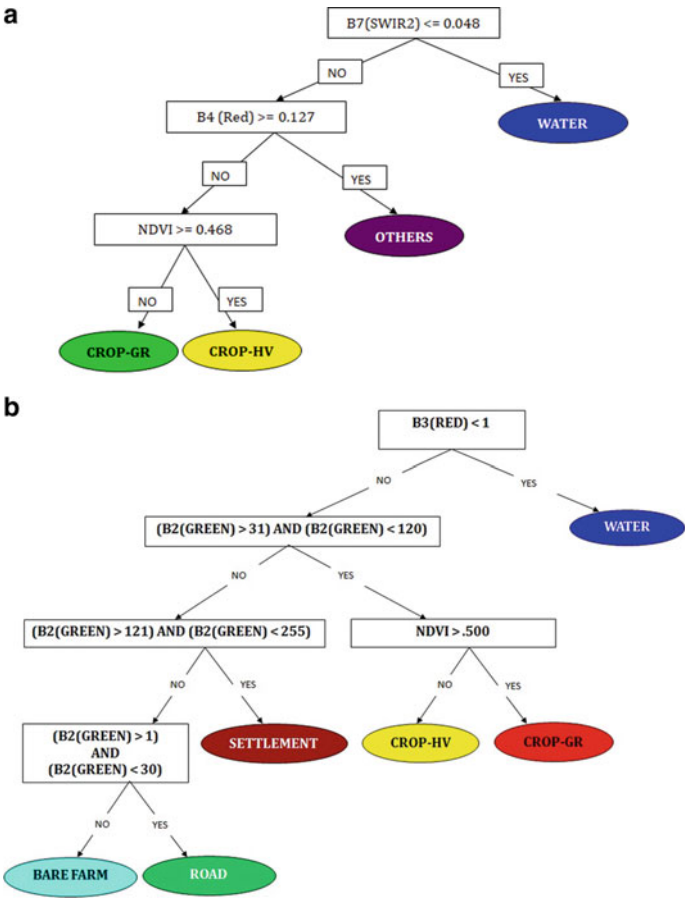
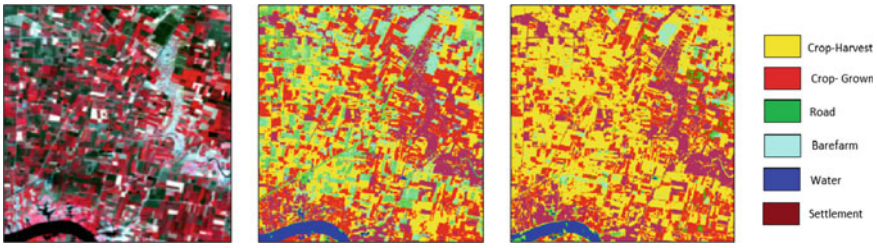
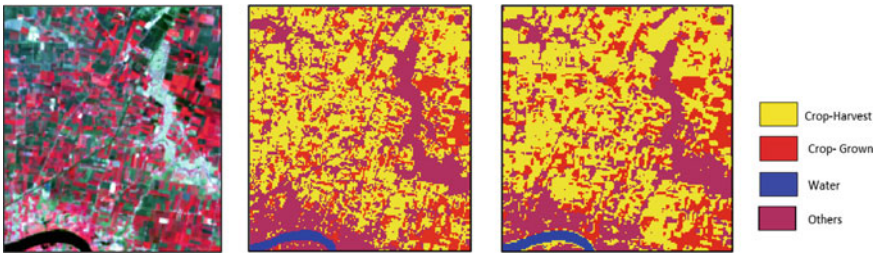


Fig. 4 **a** Optimal decision tree for Landsat-8 image. **b** Optimal decision tree for LISS-IV image



a) Test image b) MLC classified c) DT classified d) Legend

Fig. 5 Classification results for LISS-IV image of date 27th April'17



a) Test image b) MLC classified c) DT classified d) Legend

Fig. 6 Classification results for Landsat-8 image of date 24th May'17

Table 6 Classification results for study images of LISS-IV and Landsat-8 images

Study image details		Decision tree		Maximum likelihood classifier	
	Class	UA%	PA%	UA%	PA%
LISS-IV Image of date 27th Apr'17	Crop-Hv	**91.92**	71.74	68.85	55.08
	Crop-Gr	**96.49**	87.4	85.33	98.81
	Waterbody	88.32	97.71	98.3	96.68
	Road	25	20	14.74	88.32
	Settlement	99.95	82.8	96.31	8.99
	Barefarm	10	8	12.4	80.59
	OA	79.45%		75.71%	
	k	0.716		0.692	
Landsat-8 image of date 24th May'17	Crop-Hv	**92.08**	74	79.41	79.41
	Crop-Gr	**97.22**	93.22	93.06	97
	Waterbody	100	100	100	100
	Others	76.55	96.52	88.24	86.54
	OA	88.97%		89.88%	
	k	0.86		0.86	

Bold values represent highest UA values observed

References

1. Openshaw S (1999) Geographical data mining: key design issues. In: Proceedings of geocomputation '99
2. Verma AK, Garg PK, Hari Prasad K, Dadhwal V (2016) Classification of LISS-IV imagery using decision tree methods. In: The international archives of the photogrammetry, remote sensing and spatial information sciences, vol XLI-B8
3. Peña JM, Gutiérrez PA, Hervás-Martínez C, Six J, Plant RE, López-Granados F (2014) Object-based image classification of summer crops with machine learning methods. Remote Sens 6:5019–5041
4. Tso B, Mather PM (2009) Classification methods for remotely sensed data. CRC Press, Taylor & Fransic Group, Boca Raton, pp 3–15

5. Elodie V, Valentine L, Agnes B, Dino I, Maguelonne T, Stephane D, Fidiniaina R (2015) Identifying cropped areas in small growers agricultural regions using data mining for food security. Accessed on https://agritrop.cirad.fr/574632/1/document_574632.pdf
6. Schultz B, Immitzer M, Formaggio A, Sanches I, Barreto Luiz A, Atzberger C (2015) Self-guided segmentation and classification of multi-temporal Landsat-8 images for crop type mapping in southeastern Brazil. Remote Sens 7:14482–14508
7. Huang J, Wang H, Dai Q, Han D (2014) Analysis of NDVI data for crop identification and yield estimation. IEEE J Sel Top Appl Earth Observations Remote Sens 7(11):4374–4384
8. Lebourgeois V, Dupuy S, Vintrou É, Ameline M, Butler S, Bégué A (2017) A combined random forest and OBIA classification scheme for mapping smallholder agriculture at different nomenclature levels using multisource data (Simulated sentinel-2 time series, VHRS and DEM). Remote Sens 9:259
9. Gervais N, Buyantuev A, Gao F (2017) Modeling the effects of the urban built-up environment on plant phenology using fused satellite data. Remote Sens 9(1):99
10. https://www.onefivenine.com/india/villages/Aurangabad-District/Gangapur/Kaygaon. Accessed on 2nd Mar 2017
11. Hunt EB, Marin J, Stone PJ (1966) Experiments in induction. Academic, New York
12. Tucker CJ, Holben BN, Elgin JH, McMurtrey JE (1980) Relationship of spectral data to grain yield variation. Photogramm Eng Remote Sens 46:657–666
13. Schmidt M, Pringle M, Devadas R, Denham R, Tindall D (2016) A framework for large-area mapping of past and present cropping activity using seasonal Landsat images and time series metrics. Remote Sens 8:312
14. Singha M, Wu B, Zhang M (2016) An object-based paddy rice classification using multi-spectral data and crop phenology in Assam, Northeast India. Remote Sens 8:479
15. Pereira RM, Casaroli D, Vellame LM, Junior JA, Evangelista AW (2016) Sugarcane leaf area estimate obtained from corrected normalized difference vegetation index (NDVI). Pesq Agropec Trop 46(2):140–148. www.agro.ufg.br/pat
16. Kaur P, Kale KV (2017) Identification of growth stage of sugarcane crop using decision tree for Landsat-8 data. In: Proceedings of 38th Asian conference on remote sensing