Optimal Threshold Value Determination for Land Change Detection

Sangram Panigrahi[1], Kesari Verma[1], and Priyanka Tripathi[2]
[1]Department of Computer Applications, National Institute of Technology Raipur, India
[2]Department of Computer Engineering and Applications, National Institute of Technical Teachers Trainingand Research Bhopal, India

Abstract: *Recently data mining techniques have emerged as an important technique to detect land change by detecting the sudden change and/or gradual change in time series of vegetation index dataset. In this technique, the algorithms takes the vegetation index time series data set as input and provides a list of change scores as output and each change score corresponding to a particular location. If the change score of a location is greater than some threshold value, then that location is considered as change. In this paper, we proposed a two step process for threshold determination: first step determine the upper and lower boundary for threshold and second step find the optimal point between upper and lower boundary, for change detection algorithm. Further, by engaging this process, we determine the threshold value for both Recursive Merging Algorithm and Recursive Search Algorithm and presented a comparative study of these algorithms for detecting changes in time series data. These techniques are evaluated quantitatively using synthetic dataset, which is analogous to vegetation index time series data set. The quantitative evaluation of the algorithms shows that the Recursive Merging (RM) method performs reasonably well, but the Recursive Search Algorithm (RSA) significantly outperforms in the presence of cyclic data.*

Keywords: *Data mining, threshold determination, EVI and NDVI time series data, high dimensional data, land change detection, recursive search algorithm, recursive merging algorithm.*

Received October 28, 2015; accepted June 26, 2016

1. Introduction

Land cover and land use change detection it a thrust area of research in recent era. The land cover change detection problem explore when the land cover of a particular location has been converted from one type to another i.e., conversion of forested land to barren land due to fires, droughts, insect damage, logging, etc.

In literature lands cover change detection has often been performed by comparing two or more satellite snapshot images acquired on different dates [9]. For example, Let I_1 and I_2 are the two snapshots of a particular location acquired at time t_1 and t_2 respectively. The pixel I_{1ij} of image I_1 at time t_1 compares with the pixel I_{2ij} of image I_2 at time t_2, if any radical discrepancy found between two pixel values then the pixel location is assumed as change. These comparisons based method includes extensive variety of techniques ranging from simple differencing to more sophisticated regression approaches [1, 19, 20]. These images based change detection techniques have numbers of limitations, i.e., complex to find the rate of change (sudden change or gradual change). It is also difficult to find an actual change date due to bi-temporal approaches compare snapshots between two dates only. The most studies have focused on relatively small areas. The changes that occur outside the image acquisition windows are not mapped and they are inherently unsuited for application at global scale.

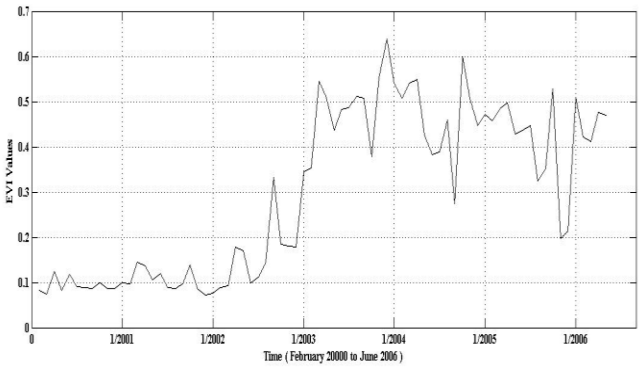
The time series data address to overcome the problem of image based technique [25] or imaged based data. Potter *et al*. [29] and Roy *et al*. [32] employed time series approach to detect land cover changes. For example, Figure 1 shows case of a land cover change found in time series data. The location of the point corresponds to Metropolitan golf links, Oakland, which was in fact opened in 2003 shown in Figure 1-b. The same site which was previously used as disposal sites for waste materials, shown in Figure 1-a. The time series for this location shows the low level of vegetation prior to 2003, then abrupt jump in vegetation in 2003, after which the vegetation is relatively uniform, and consistent in Figure 1-c. This golf course was constructed in 2003, and it is clearly observed in the time series of that location.

The data mining technique (or time series based change detection) has significant advantages over the comparison of snapshot images. In this technique, detection of changes is based on the pattern of spectral response of the landscape over time rather than the differences between two or more images collected on different dates. This technique addresses the various limitations of image based techniques i.e., when the changes occurred (time of change) and rate of change (sudden change or gradual change, etc.,), the extent, and pattern of re-growth etc.

Very large or long time series databases of ecosystem, environmental and climate data are difficult to analyze and interpret. Availability of large amounts of data necessitates data mining techniques to alleviate the automatic extraction and analysis of interesting patterns from the Earth Science data [18].

a) Image of location before construction of golf course.

b) Image of location after construction of golf course.

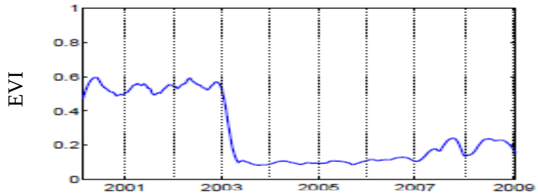


c) EVI time series of metropolitan golf links, Oakland.

Figure 1. An example of a land cover change in time series data.

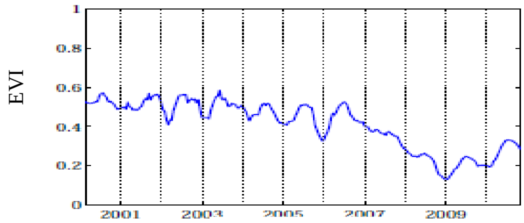
Figure 1-a shows the location correspond to before construction of metropolitan golf links Oakland, CA, Figure 1-b shows the location after construction of golf course in 2003, Figure 1-c shows the time series of the corresponding gulf course location from 2000 to 2006. The year 2003, which corresponds to the time step at which the time series exhibits a change [4]. (Image Source: Google Earth.)

There are broadly three types of vegetation changes in time series based data: sudden change, gradual change and land cover type change as shown in the below Figure 2. The sudden change is abrupt and unexpected reduction in large vegetation index extends over multiple months (greater than seasonal months) or multiple years at the time. The sudden change is also called abrupt change. The example of sudden changes is forest degradation due to event like forest fire, flood, mechanized clearings, etc. The gradual change approach is to look for gradual increase or decrease in vegetation trends spanning over multiple years in the time. The illustration of gradual changes is forest degradation due to long-term droughts, beetle infestations or gradual logging, etc. The land covers type change method, divide the time series into homogeneous segment regions and the boundary of the segments indicate changes in vegetation. It identifies any change in the vegetation type such as change from one land cover to other or changes in cropping patterns, clearing of forests for agriculture, urban expansion, and so on.



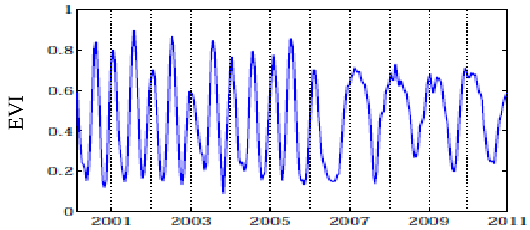
(Time 2000 to 2009)

a) Sudden change.



(Time 2000 to 2011)

b) Gradual change.



(Time 2000 to 2011)

c) Land cover type change.

Figure 2. Example of three types of vegetation changes in time series based data.

Earth science data consists of global snapshots of measurement values for a number of variables like temperature, pressure, precipitation and Vegetation Index (VI) collected for all land and sea surfaces. The VI is an important ecosystem variable and has been widely used for the phonologic monitoring [12]. The two global-based VI are Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). These vegetation indices essentially serve as a measure of the amount and “greenness” of vegetation at a particular location. The NDVI and EVI global dataset available from 1981 and 2000 respectively, with 250m, 500m, 1KM and 0.05 degree (or 5600m) spatial resolution, each 16 days and each calendar month. These dataset include Quality Assurance (QA) information, which address under which conditions each pixel was acquired and processed [12]. The MODIS Land Discipline Group provides consistent spatial and temporal information regarding NDVI and EVI, which are available for public download [36].

1.1. Key Contributions

The key contributions of this paper are as follows:

- We systematically study the two algorithms for land cover change detection. We quantitatively evaluate their performance for synthetic dataset.
- We proposed a novel algorithm for determination of threshold for land change detection algorithm and experimental validation is performed.

2. Related Work

This section briefly discusses about various data mining techniques for detecting changes in vegetation index time series [30]. These techniques are able to take advantage of the inherent structure present in vegetation index time series data, i.e., seasonality. Furthermore, these data mining techniques for time series change detection can be categories into four key approaches, statistical parameter change approaches, predictive model based approaches, unsupervised based approach and segmentation based approaches. Boriah *et al*. [3] have slightly modified Lunetta *et al*. [21] algorithm by taking the absolute maximum value of computed z-score value of consecutive annual sum differences present in time series and named as Modified Lunetta algorithm. The high values of z-score indicate decrease in vegetation and vice-versa. Boriah *et al*. [6] have further developed an adaptation of CUSUM algorithm [16, 26, 34] and called CUSUM_MEAN. This algorithm calculated the cumulative sum (CS_k for $k = 1. . . n$ observations) for each observation by comparing the deviation of every observation to the expected value (μ), which is computed by taking the average of the first annual cycle of the time series. The change score assign to a location is set the highest value of the cumulative sum, i.e. $|\max (CS_1, CS_2, ... ,CS_n)|$ or $|\min (CS_1, CS_2, ... , CS_n)|$ for positive change or negative change respectively. Chamber *et al*. [8] proposed PDELTA approach, first computes the Delta-series (*Δ-series*) by taking the corresponding time step difference between two consecutive annual segments and then identifies the window of maximum reliable drop. The maximum reliable drop window is the representative window of the time series. The higher the score value, represent the more severe the change. This technique use to identify the gradual vegetation change. Yearly Delta algorithm (YD) has been proposed by Boriah [7], which is based on EWMA [22] technique and used by Mithal *et al*. [23, 24]. This algorithm builds a model by predicting the expected EVI values for the each time step of future years. The YD score for a location is assigned by taking the maximum of calculated mean deviation of the observed value from the predicted value of the successive time steps over a year. The time step with the maximum score is considered as the change time. In unsupervised learning researcher Boriah *et al*. [5] and Potter *et al*. [28] used the k-means algorithm [14] to cluster the 250m EVI time series to solve land cover change detection problem. The two techniques have been discussed based upon clustering, one Distance to cluster Centroid and other Confidence Intervals around Cluster Centroids. Chen *et al*. [10] develops unsupervised spatio-temporal data mining methods by an integrated analysis in both the EVI and AF (Active Fire) datasets [15] to identify fire events. They defined 3 different scoring mechanisms: K-month Delta (KD), Local Instant Drop (LID) and Near Drop (ND). Authors consider a pixel is as initial pixels (highest stratum) forest fire events, based on present in AF dataset and satisfying scoring criteria. To increase coverage, they consider the 24 spatial neighbors in a 5×5 spatial grid around the initial pixels and apply the same scoring mechanism. The resultant data is called middle stratum. The lowest stratum is generated by relaxing scoring criteria for similar events in a spatial window around the other two strata. Verbesselt *et al*. [37] have proposed Breaks for Additive Season and Trend (BFAST) technique, which is iteratively estimate the trend and seasonal components and allow for an individual estimation of breakpoints in the seasonal and trend component. The optimal position of these breaks can be determined by minimizing the residual sum of squares, and the optimal number of breaks can be determined by minimizing Bayesian Information Criterion (BIC). BFAST decomposes a time series data into trend, seasonal, and remainder components, so that the intra-segment models are constant, while inter-segment models are dissimilar. Garg *et al*. [11] demonstrates a novel model-free segmentation approach, which key idea is to find the two segments in a given time series, such that the annual years present in a segment are similar and between the segments are dissimilar. The similarity and dissimilarity between the annual years for each segment can be expressed by Cohesion and Separation. The difference between the cohesion and separation values indicates the amount of change in the time series with respect to the natural variation. The maximum resultant difference value is considered as change score for that particular location.

3. Recursive Merging Algorithm

Boriah *et al*. [4, 5] proposed a segmentation based algorithm, which partitioned time series into homogeneous segments and boundaries between the segments represent change points. It also exploits seasonality in order to detect land cover change of a particular location. If a given location has not had a land cover change, then the seasonal cycles look very similar going from one year to the next; if this is not, then assign a change score to a land location based on the extent to which the seasons are different. Recursive Merging follows a bottom-up strategy [17] of merging annual segments that are consecutive in time and similar in value. A cost corresponding to each merge is calculated by taking Manhattan distance between the segments. The Manhattan distance is considers the seasonality of the time series into account because it takes difference between the corresponding months. The key idea of this algorithm is that, it will merge similar annual cycles and calculate the cost, if final merging cost is maximum, it correspond to the change occurred in the time series. If final merging cost is low, it is likely that no change occurred in the time series.

The strength of recursive merging algorithm is its robustness to noisy, missing data and scalable to large global scale data sets. This approach is able to detect changes in both directions. The algorithm takes into account the seasonality of the data but not the variability. The limitation of recursive merging approach is change point may not be identified accurately. This approach does not handle multiple change points in time series. Although the Earth Science data sets exhibit significant spatio-temporal autocorrelation, but algorithm does not make use of spatial information that is present in the data.

4. Recursive Search Algorithm

Panigrahi *et al*. [31] proposed a novel, simple and efficient segmentation based land change detection technique. The Recursive Searching Algorithm (RSA) not only detect the changes in time series data, it also detects the time of change depends on the temporal resolution (each 16 days and each calendar month) present in the time series and label with the change type (whether sudden increase or decrease type) of time series.

The key idea of the RSA technique is to find two consecutive segments in the time series such that the intra-segment within each segment is very similar while inter-segment being significantly different. For this, the RSA technique first calculates the difference between each two consecutive segment presents in the time series using City-block distance, which is exploiting the seasonality of the time series data. The distance between two consecutive segments indicate, how one segment dissimilar from the other. The RSA technique tries to find the highest consecutive segment difference value and its index value by employing the two way searching method. This technique compares the first consecutive segmentation difference value with last consecutive segmentation difference value, the largest difference value among two is stored in another variable S (initially S=0) and the corresponding index of difference array (d) is stored in j. subsequently, the second first consecutive segmentation difference value compares with second last consecutive segmentation difference value, the largest difference value among two is compared with variable S, then greatest among two and the corresponding index of difference array (d) are assigned to S and j respectively (which is replaced the previous stored value of S and j). This comparison process is called recursively until satisfy the stopping criteria. From the resultant consecutive segment, the first segment is unchanged (j^{th}) segment and the second segment is changed segment ($j+1^{th}$), which is shown in Figure 3. In the time series data, 1^{st} to j^{th} segments follows the similar pattern and which is different from the from $(j+1)^{th}$ segment. To find the Change Score, the month wise EVI value average (avg) of 1^{st} to j^{th} segments are calculated first, then takes the annual sum, S_{avg}. Then S_{avg} is subtracted from the $(j+1)^{th}$ segment's computed annual sum, S_{org} and store in S_{diff}. The Change Score of time series can be computed by taking the absolute value of S_{diff}. The change type for each location is determined by the value of S_{diff}, i.e., if S_{diff} value is greater than zero, then the location is labeled as Positive change; if S_{diff} value is lesser than zero, then the location is labeled as Negative change; otherwise the location is labeled as No change. To determine the change point or time of change, the RSA technique, first compute the absolute value of the month wise difference between before change (j^{th}), change ($j+1^{th}$) and change ($j+1^{th}$), after change ($j+2^{th}$) segments and the resultant values store in Y. The authors assume that change point lies in "before change" segment, so find the confidence of first 12 value of the Y. The confidence of an element Y_i can be calculated by checking how many times an element Y_i greater than to its next 12 elements such as, Y_{i+1} to Y_{i+12}. The change point is a point with the confidence percentage greater than the certain user defined threshold (*Th*) and highest differential value (means highest Y_i value).

The RSA technique focuses on sudden changes and identifying the single most substantial changes in a time series dataset. This technique is scalable to large global scale data sets. It also detect the time of change close to the actual change with accuracy up to the temporal resolution of the input data set. The RSA algorithm has limitation in a number of scenarios; this algorithm fails to detect change, if a change event occurs during the first and last year of the time series data. This method always finds the change point at before change segment, but it may be possible that the change point may occurs at the initial time of the change segment.

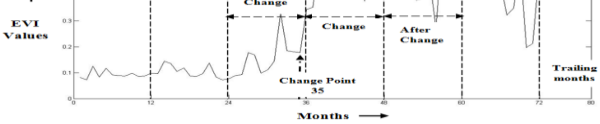


Figure 3. The RSA algorithm determines the change segment as well as change point, which corresponds to the time step at which the time series exhibits a sudden change [31].

5. Evaluation

We employ the quantitative evaluation technique to examine and understand the relative performance of Recursive Search Algorithm with Recursive Merging algorithm using standard datasets. In the following, we describe about validation data utilized in this study and also provides a general idea of the evolution methodology.

5.1. Dataset

A standard data set of synthetically generated time series is used [2, 27], which is downloaded from UCI KDD [33]. This data set contains 600 time series of 6 different types, namely: Normal, Cyclic, Increasing trend, Decreasing trend, upward shift, downward shift and each type consist of 100 time series. Each time series consists of 60 values.

We created a data set, DS by taking the few different types of time series data i.e., Normal, Cyclic, Upward shift and downward shift time series from the above standard data set. The data set DS comprises 200 Normal types, 100 Cyclic type, 100 Upward shift type and 100 Downward shift type time series data. Each time series with 60 values and assumes that each value is corresponding EVI value for a particular month. Here both Normal and Cyclic type's data are labelled as unchanged class, where as both Upward and Downward shift type data are labelled as changed class.

The EVI value near to 0 correspond to barren areas of rock, sand, or snow means no vegetation, while 1 indicates the highest possible density of green leaves means highly saturated. For the change detection problem, the earth science domain experts suggested to removal of EVI value less than or equal to 0.1 and above 0.9 [6]. We scale the all values of DS data set into the ranges from 0.1 to 0.9, by applying min-max normalization techniques [35], which perform a linear transformation on the original data. The min-max normalization is computed by using following Equation (1):

$$v' = \frac{v - \min}{\max - \min}\left(new\_\max - new\_\min\right) + new\_\min \qquad (1)$$

Where, v is the value to be mapped and v' is the mapped value in the ranges from 0.1 to 0.9. The max and min are the maximum and minimum value of DS data set respectively. Here *new*_max and *new*_min are the maximum and minimum value of normalized data set respectively. Here *new*_max=0.9 and *new*_min= 0.1.

The RSA algorithm assumes that change point does not occur on the first and last year of the time series data, thus it is excluded these time series data from DS data set. Then the data set DS contains 152 Normal types, 76 Cyclic type, 100 Upward shift type and 100 Downward shift type, total 428 time series data.

5.2. Evaluation Methodology

For quantitatively comparing the performance of Recursive Search Algorithm with Recursive Merging algorithm, we employ Precision, Recall, F-score and Accuracy as evaluation metrics, because these metrics are generally used to measure the performance of algorithms in information retrieval, machine learning and data mining [13, 35].

Given a time series data set D with N time series, each time series represent for one respective location. Each change detection algorithm calculates change score for each location and assign. Finally, it returns a list of N change scores, where each change score is a measure of the degree of change for the corresponding location. Here, we have taken the standard data sets which consist of the true labels of each of the location; let M be the actual number of sudden increase and decrease in time series data present in the data set, which is labelled as disturbance and called validation data. Then according to the descending order of their change score the locations are ranked. The algorithm flags the top n ranked locations as change events ($1 \leq n \leq M$) and the lower ranked locations as unchanged. By computing the intersection with the validation data, we find the number of True Positives (TP_n), changes detected by the scheme also present in the validation data; False Positives (FP_n), changes found by the scheme but not in the validation data; True Negatives (TN_n), changes neither found by the scheme also not present in the validation data and False Negatives (FN_n), changes noted in the validation data but not found by the scheme for each algorithm, as shown a confusion matrix in Table 1.

Table 1. Confusion matrix.

		Predicted	
		Fire	No
Validation Data	Fire	TP_n	FN_n
	No	FP_n	TN_n

The Precision, Recall, F-score and Accuracy can be defined as:

$$precision(p_n) = \frac{TP_n}{TP_n + FP_n} \qquad (2)$$

$$recall(r_n) = \frac{TP_n}{M}, \quad Where \quad M = TP_n + FN_n \qquad (3)$$

$$F - Score(F) = 2 * \left(\frac{precision * recall}{precision + recall}\right) \qquad (4)$$

$$Accuracy = \frac{TP_n + TN_n}{TP_n + TN_n + FP_n + FN_n} \qquad (5)$$

Precision can be perceived as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Often, there is an inverse relationship between precision and recall. F-score is a measure of a test's accuracy and Accuracy presents the overall correctness of the model.

6. Optimal Threshold Determination and Result with Discussions

The Recursive Merging (RM) and Recursive Search Algorithm (RSA) were executed on DS time series dataset. Both the algorithms provide a list of change scores as output and each change score corresponding to a particular location. The locations under study can be ranked according to their change score given by the algorithm. The locations with higher scores are likely to have changed. Figures 5 and 7 shows the histograms of the change score obtained by RSA and RM algorithm respectively. Among the large coverage of land cover data, only small region will actually exhibit a change. A location with change score greater than some threshold value is exhibit a change. Determine the optimal threshold value for change detection problem is a tedious task, especially when dealing with time series data. Here we discuss the procedure for selection of optimum threshold value for both above discussed land change detection algorithm and then find the performance of both algorithm and compare. The procedure for optimum threshold determination is a two step process: first upper and lower boundary detection for threshold and second find the optimal point between upper and lower boundary. The process is explained in Figure 4.

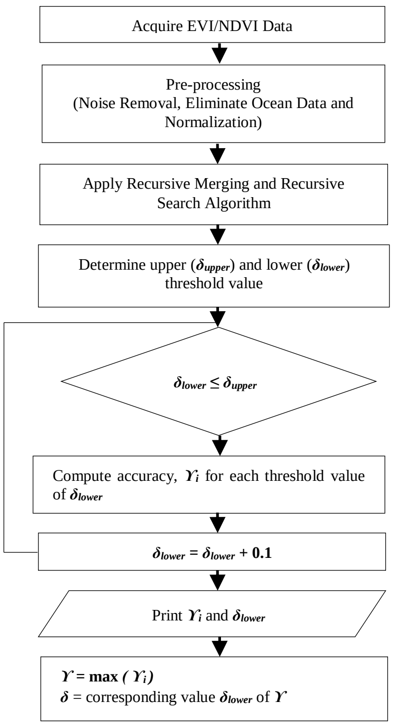


Figure 4. Proposed methodology for threshold determination.

The initial threshold value ($\delta_{initial}$) for change detection algorithm can be determined by selecting the score, where the sudden reduce in change score frequency along with highest difference in change score from the histogram of the change score produced by that algorithm for a Data set. Then compute accuracy ($\Upsilon_{initial}$) of the algorithm at $\delta_{initial}$. After that increase the threshold value with step size one ($\delta_{new} = \delta_{initial} + 1$) and compute the corresponding accuracy value (Υ_{new}). Then assign the $\Upsilon_{privious}$, $\Upsilon_{current}$, $\delta_{privious}$, and $\delta_{current}$ respectively with $\Upsilon_{initial}$, Υ_{new}, $\delta_{initial}$ and δ_{new}. After that compare the previous accuracy value ($\Upsilon_{privious}$) with current accuracy value ($\Upsilon_{current}$), if $\Upsilon_{current}$ is greater than or equal to $\Upsilon_{privious}$, then $\Upsilon_{privious}$ and $\delta_{privious}$ are assigned with $\Upsilon_{current}$ and $\delta_{current}$ respectively. Now increase the $\delta_{current}$ by one and compute its corresponding accuracy (Υ_{new}), and store in $\delta_{current}$ and $\Upsilon_{current}$ respectively, such that, $\delta_{current} = \delta_{new} + 1$ and $\Upsilon_{current} = \Upsilon_{new}$. This procedure is continued until satisfy the condition, $\Upsilon_{current} < \Upsilon_{privious}$. When this condition satisfied, the current threshold value ($\delta_{current}$) is considered as upper threshold boundary, mathematically, $\delta_{upper} = \delta_{current}$ and the lower threshold boundary can be calculated by subtracting current threshold value minus total increased step (S), here S is two, mathematically, ($\delta_{lower} = \delta_{current} - S$). Actually the optimal threshold value lies in between the lower and upper threshold boundary. After finding lower and upper threshold boundary value, the threshold value increase with step size 0.1 from lower to upper threshold boundary value and compute accuracy percentage at each threshold values. From the list, the threshold value with high percentages of accuracy is considered as optimal threshold (δ). The pseudo-code for optimal threshold determination for land change detection algorithm is explained below:

Algorithm 1: Optimal threshold determination

//symbol δ represents for threshold and Υ represents for accuracy
//Initialize $\delta_{initial}$ from the histogram of the change score produced by change detection algorithm
// compute the accuracy value for $\delta_{initial}$
compute $\Upsilon_{initial}$
// Initialize the total increased step
$S \leftarrow 0$
// increase the threshold value with step size one
$\delta_{new} \leftarrow \delta_{initial} + 1$
// compute the accuracy value for δ_{new}
compute Υ_{new}
// assign $\Upsilon_{privious}$ and $\delta_{privious}$ as $\Upsilon_{initial}$ and $\delta_{initial}$ respectively.
$\Upsilon_{privious} \leftarrow \Upsilon_{initial}$
$\delta_{privious} \leftarrow \delta_{initial}$
// assign $\Upsilon_{current}$ and $\delta_{current}$ as Υ_{new} and δ_{new} respectively.
$\Upsilon_{current} \leftarrow \Upsilon_{new}$
$\delta_{current} \leftarrow \delta_{new}$
while ($\Upsilon_{current} >= \Upsilon_{privious}$) do
    //assign $\Upsilon_{privious}$ and $\delta_{privious}$
    $\Upsilon_{provious} \leftarrow \Upsilon_{current}$
    $\delta_{privious} \leftarrow \delta_{current}$
    //threshold value increase by one
    $\delta_{current} \leftarrow \delta_{current} + 1$
    //compute new accuracy for $\delta_{current}$
    compute Υ_{new}

// assign Υ_{new} to $\Upsilon_{current}$
    $\Upsilon_{current} \leftarrow \Upsilon_{new}$
    // increase total increased step by 1

$S \leftarrow 0 + 1$
end while
//upper boundary threshold value is δ_{upper}
$\delta_{upper} \leftarrow \delta_{current}$
//lower boundary threshold value is δ_{lower}
$\delta_{lower} \leftarrow \delta_{current} - S$
for $i = \delta_{lower}$ *to* δ_{upper} *step 0.1 do*
    compute accuracy (Υ_i) *for each threshold value* (δ_i)
end for
//optimal accuracy
$\Upsilon \leftarrow max\ (\Upsilon_i)$
//optimal threshold
$\delta \leftarrow$ *Find the corresponding value* δ_i *of* Υ

The initial threshold value for RSA has been determined from the histogram of the change scored produced by RSA algorithm for DS Data set. Figure 5, shows the histograms of the change score obtained by RSA algorithm and from this histogram we determine the initial threshold value is 7. By employing above discussed threshold determination procedure, we first obtain the lower and upper threshold boundary for RSA algorithm, which are 7 and 9 respectively. After that we obtain the list of threshold values by increasing step size 0.1 from lower to upper threshold boundary and evaluate the performance of RSA algorithm for each threshold value, which is shown in the Table 2. From the list, the threshold value with high percentages of accuracy is considered as optimal threshold, i.e., here 8.2. The graph plotted between threshold value and its corresponding accuracy produced by RSA algorithm for DS Data set is shown in Figure 6. The optimum threshold value for RSA algorithm is depicted in the graph.

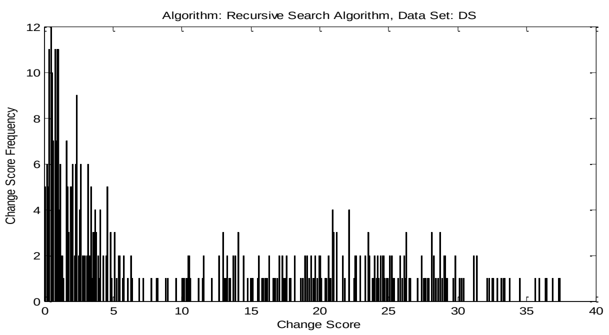


Figure 5. Histogram of change scores produced by RSA for DS Data set.

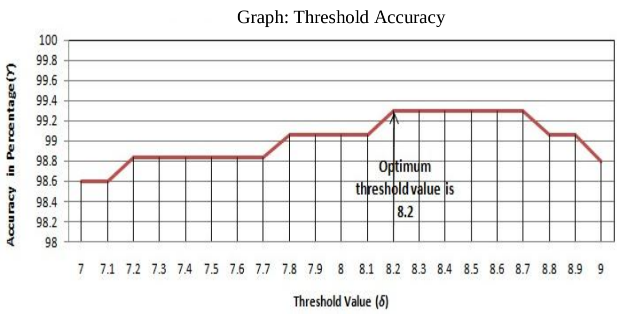


Figure 6. The graph plotted between threshold value and its corresponding accuracy produced by RSA algorithm for DS Data set.

Similarly, we plot the histogram for change score obtained by RM algorithm for DS dataset, shown in Figure 7, and determine the initial threshold value, i.e., 21, from this histogram. Then obtain the lower and upper threshold boundary for RM algorithm, which are 23 and 25 respectively, by employing above discussed threshold determination procedure. The list of threshold values are produced by increasing step size 0.1 from lower to upper threshold boundary and then evaluate the performance of RM algorithm for each threshold values, which is shown in the Table 3. From the list, the threshold value with high percentages of accuracy is considered as optimal threshold, i.e., here 24.2. The optimum threshold value has been depicted in the graph plotted between threshold value and its corresponding accuracy produced by RM algorithm for DS dataset is shown in Figure 8.

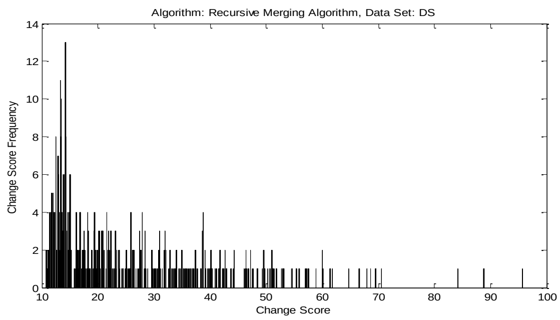


Figure 7. Histogram of change scores produced by RM algorithm for DS Data set.

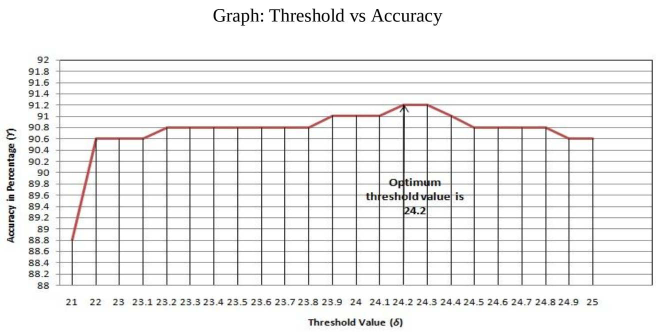


Figure 8. The graph plotted between threshold value and its corresponding accuracy produced by RM algorithm for DS Data set.

The optimum threshold values for RM and RSA algorithm for DS time series dataset are 24.2 and 8.2 respectively. Both the algorithms provide a list of change scores as output and each change score corresponding to a particular location. A location with change score greater than optimum threshold value is exhibit a change. The validation result and performance of both algorithms at their respective optimum threshold value are presented in Table 4 in the following we discuss their performance as well as the relative strengths and weaknesses in detail.

Table 4 presents the evaluation result of DS data set, which consists the normal, cyclic, upward shift (or sudden increase) and downward shift (or sudden decrease) time series data. The RM algorithm incurs 19 false positives and 25 false negative cause high precision and recall value, i.e. 0.9021 and 0.8750 respectively, which gives the accuracy 91.2%. The RSA technique finds one false positives and two false negative provoke very high precision and recall, which provides accuracy approximately 99.3%. The Figure 9 shows the comparison of precision, recall, F-score and accuracy of each discussed algorithm for DS data set.

Figure 9. Showing the performance Comparison of algorithms on DS dataset.

The Evaluation result shows that both RM and RSA techniques perform better on DS dataset. Also, both discussed algorithm able to find sudden changes such as sudden increases or sudden decreases in time series data and robust to the presence of noisy and missing values in the time series. RM algorithm performs better on DS data sets, but the evaluation result shows that RSA algorithm significantly outperforms. Both RM and RSA algorithm accounts for seasonality and variability in the time series but in particular, RSA algorithm significantly outperforms in the presence of cyclic data. Both algorithms are scalable to massive data sets, which is one of the key requirements for global scale land cover change studies.

Table 2. Performance of Recursive Search Algorithm at different Threshold values for DS dataset.

Technique	Threshold value (δ)	TP	FN	TN	FP	Precision	Recall	F-Score	Accuracy (ϒ)
Recursive Search Algorithm	7	198	2	224	4	0.9802	0.9900	0.9851	98.5981
	8	198	2	226	2	0.9900	0.9900	0.9900	99.0654
	9	196	4	227	1	0.9949	0.9800	0.9847	98.8
	7.1	198	2	224	4	0.9802	0.9900	0.9851	98.5981
	7.2	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.3	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.4	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.5	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.6	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.7	198	2	225	3	0.9851	0.9900	0.9875	98.8318
	7.8	198	2	226	2	0.9900	0.9900	0.9900	99.0654
	7.9	198	2	226	2	0.9900	0.9900	0.9900	99.0654
	8.1	198	2	226	2	0.9900	0.9900	0.9900	99.0654
	8.2	**198**	**2**	**227**	**1**	**0.9950**	**0.9900**	**0.9925**	**99.2991**
	8.3	198	2	227	1	0.9950	0.9900	0.9925	99.2991
	8.4	198	2	227	1	0.9950	0.9900	0.9925	99.2991
	8.5	198	2	227	1	0.9950	0.9900	0.9925	99.2991
	8.6	198	2	227	1	0.9950	0.9900	0.9925	99.2991
	8.7	198	2	227	1	0.9950	0.9900	0.9925	99.2991
	8.8	197	3	227	1	0.9949	0.9850	0.9899	99.0654
	8.9	197	3	227	1	0.9949	0.9850	0.9899	99.0654

Table 3. Performance of recursive merging algorithm at different threshold values for DS dataset.

Technique	Threshold value (δ)	TP	FN	TN	FP	Precision	Recall	F-Score	Accuracy(ϒ)
Recursive Merging Algorithm	21	187	13	257	43	0.8130	0.9350	0.8698	88.80
	22	185	15	268	32	0.8525	0.9250	0.8873	90.60
	23	179	21	273	27	0.8689	0.8950	0.8818	90.60
	24	175	25	280	20	0.8974	0.8750	0.8861	91
	25	172	28	281	19	0.9005	0.8600	0.8798	90.60
	23.1	178	22	275	25	0.8768	0.8900	0.8834	90.60
	23.2	178	22	276	24	0.8812	0.8900	0.8856	90.80
	23.3	178	22	276	24	0.8812	0.8900	0.8856	90.80
	23.4	177	23	277	23	0.8850	0.8850	0.8850	90.80
	23.5	177	23	277	23	0.8850	0.8850	0.8850	90.80
	23.6	177	23	277	23	0.8850	0.8850	0.8850	90.80
	23.7	176	24	278	22	0.8889	0.8800	0.8844	90.80
	23.8	175	25	279	21	0.8929	0.8750	0.8838	90.80
	23.9	175	25	280	20	0.8974	0.8750	0.8861	91
	24.1	175	25	280	20	0.8974	0.8750	0.8861	91
	24.2	**175**	**25**	**281**	**19**	**0.9021**	**0.8750**	**0.8883**	**91.2**
	24.3	175	25	281	19	0.9021	0.8750	0.8883	91.2
	24.4	174	26	281	19	0.9016	0.8700	0.8855	91
	24.5	173	27	281	19	0.9010	0.8650	0.8827	90.80
	24.6	173	27	281	19	0.9010	0.8650	0.8827	90.80
	24.7	173	27	281	19	0.9010	0.8650	0.8827	90.80
	24.8	173	27	281	19	0.9010	0.8650	0.8827	90.80
	24.9	172	28	281	19	0.9005	0.8600	0.8798	90.60

Table 4. Evaluation result of DS Data Set.

Technique	Threshold value (Th)	TP	FN	TN	FP	Precision	Recall	F-Score	Accuracy
Recursive Merging Algorithm	24.2	175	25	281	19	0.9021	0.8750	0.8883	0.912
Recursive Search Algorithm	8.2	198	2	227	1	0.9950	0.9900	0.9925	0.992991

7. Conclusions

In this paper, we have suggested a simple process for determination of local optimal threshold value for land change detection algorithm. Subsequently, by engaging this process, determined the threshold value for both Recursive Merging Algorithm and Recursive Search Algorithm and presented a comparative study of these algorithms for detecting changes in time series data. The quantitative evaluation of the algorithms shows that the RM method performs reasonably best but the RSA algorithm significantly outperforms. Both RM and RSA algorithm accounts for seasonality and variability in the time series but in particular, RSA algorithm significantly outperforms in the presence of cyclic data. RSA detect the time of change close to the actual change with accuracy up to the temporal resolution of the input data set where as RM detects change points at boundaries between the segments. RSA algorithm has limitation in a number of scenarios; this algorithm fails to detect change, if a change event occurs during the first and last year of the time series data. This method always find the change point at before change segment but it may be possible that the change point may occurs at the initial time of the change segment. Future work includes Both the algorithms will be applied on 250m EVI of time series data set of a particular geographical region to detect land cover changes and make quantitative and qualitative evaluation. In future we will perform the experimental study in real EVI/NDVI dataset.

References

[1] Aach T. and Kaup A., "Bayesian Algorithms for Adaptive Change Detection in Image Sequences using Markov Random Fields," *Signal Processing: Image Communication*, vol. 7, no. 2, pp. 147-160, 1995.

[2] Alcock R. and Manolopoulos Y., "Time-Series Similarity Queries Employing a Feature-Based Approach," *in Proceedings of 7th Hellenic Conference on Informatics*, Ioannina, pp. 27-29, 1999.

[3] Boriah S., Kumar V., Potter C., Steinbach M., and Klooster S., "Land Cover Change Detection using Data Mining Techniques," Technical Report, 2008.

[4] Boriah S., Kumar V., Steinbach M., Potter C., and Klooster S., "Land Cover Change Detection: A Case Study," *in Proceedings of 14th International Conference on Knowledge Discovery and Data Mining*, Las Vegas, pp. 857-865, 2008.

[5] Boriah S., Kumar V., Steinbach M., Tan P., Potter C., and Klooster S., *Detecting Ecosystem Disturbances and Land Cover Change Using Data Mining*, CRC Press, 2009.

[6] Boriah S., Mithal V., Garg A., Kumar V., Steinbach M., Potter C., and Klooster S., "A Comparative Study of Algorithms for Land Cover Change," *in Proceeding of the Conference on Intelligent Data Understanding*, CA, pp. 175-188, 2010.

[7] Boriah S., Time Series Change Detection: Algorithms for Land Cover Change, PhD Thesis, University of Minnesota, 2010.

[8] Chamber Y., Garg A., Mithal V., Brugere I., Lau M., Krishna V., Boriah S., Potter C., and Klooster S., "A Novel Time Series based Approach to Detect Gradual Vegetation Changes in Forests," *in Proceedings of Conference on Intelligent Data Understanding*, California, pp. 248-262, 2011.

[9] Coppin P., Jonckheere I., Nackaerts K., Muys B., and Lambin E., "Digital Change Detection Methods in Ecosystem Monitoring: a Review," *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1565-1596, 2004.

[10] Chen X., Karpatne A., Chamber Y., Mithal V., Lau M., Steinhaeuser K., Boriah S., Steinbach M., Kumar V., Potter C., Klooster S., Abraham T., Stanley J., and Castilla-Rubio J., "A New Data Mining Framework for Forest Fire Mapping," *in Proceedings of Conference on Intelligent Data Understanding*, Boulder, pp. 104-111, 2012.

[11] Garg A., Manikonda L., Kumar S., Krishna V., Boriah S., Steinbach M., Toshnival D., Kumar V., Potter C., and Klooster S., "A Model-Free Time Series Segmentation Approach for Land Cover Change Detection," *in Proceedings of the Conference on Intelligent Data Understanding*, California, pp. 144-158, 2011.

[12] Huete A. and Justice C., *MODIS Vegetation Index (MOD13) Algorithm Theoretical Basis Document*, the University of Arizona, 1999.

[13] Han J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.

[14] Jain A. and Dubes R., *Algorithm for Clustering Data*, Prentice Hall, 1988.

[15] Justice C., Giglio O., and Roy L., "MODIS-Derived Global Fire Products," *in Proceedings of Land Remote Sensing and Global Environmental Change*, New York, pp. 661-679, 2011.

[16] Kucera J., Barbosa P., and Strobl P., "Cumulative Sum Charts-a Novel Technique for Processing Daily Time Series of Modis Data for Burnt Area Mapping in Portugal," *International Workshop on the Analysis of Multi-Temporal Remote Sensing Images*, Leuven, pp. 1-6, 2007.

[17] Keogh E., Chu S., Hart D., and Pazzani M., "Segmenting Time Series: A survey and Novel Approach," *in Proceedings of Data Mining in Time Series Databases*, pp. 1-21, 2003.

[18] Kumar V., Steinbach M., Tan P., Klooster S., Potter C., and Torregrosa A., "Mining Scientific Data: Discovery of Patterns in the Global Climate System," *in Proceedings of the Joint Statistical Meetings*, Athens, 2001.

[19] Kasetkasem T. and Varshney P., "An Image Change Detection Algorithm based on Markov Random Field Models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1815-1823, 2002.

[20] Karim S., Riadh F., Solaiman B., and Mohamed B., "Towards a Multi-Temporal Approach for Satellite Image Interpretation," *The International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 281-287, 2008.

[21] Lunetta R., Knight J., Ediriwickrema J., Lyon J., and Worthy L., "Land-Cover Change Detection using Multi-Temporal MODIS NDVI Data," *Remote Sensing of Environment*, vol. 105, no. 2, pp. 142-154, 2006.

[22] Lucas J. and Saccucci M., "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements," *Technometrics*, vol. 32, no. 1, pp. 1-12, 1990.

[23] Mithal V., Garg A., Boriah S., Steinbach M., Kumar V., Potter C., Klooster S., and Castilla-Rubio J., "Monitoring Global Forest Cover using Data Mining," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 4, 2011.

[24] Mithal V., Garg A., Brugere I., Boriah S., Kumar V., Steinbach M., Potter C., and Klooster S., "Incorporating Natural Variation into Time Series based Land Cover Change Detection," *in*

Proceedings of Conference on Intelligent Data Understanding, CA, pp. 45-59, 2011.
[25] Nemmour H. and Chibani Y., "Support Vector Machines for Automatic Multi-class Change Detection in Algerian Capital Using Landsat TM Imagery," *Journal of the Indian Society of Remote Sensing*, vol. 38, no. 4, pp. 585-591, 2010.
[26] Page E., "Continuous Inspection Schemes," *Biometrika*, vol. 41, no. 1-2, pp. 100-115, 1954.
[27] Pham D. and Chan A., "Control Chart Pattern Recognition using a New Type of Self Organizing Neural Network," *in Proceedings of the Institution of Mechanical Engineers*, vol. 212, no. 1, pp. 115-127, 1998.
[28] Potter C., Genovese V., Gross P., Boriah S., Steinbach M., and Kumar V., "Revealing Land Cover Change in California With Satellite Data," *EOS, Transactions, American Geophysical Union*, vol. 88, no. 26, pp. 269-276, 2007.
[29] Potter C., Tan P., Kumar V., Kucharik C., Klooster S., Genovese V., Cohen W., and Healey S., "Recent History of Large-Scale Ecosystem Disturbances in North America Derived from the AVHRR Satellite Record," *Ecosystems*, vol. 8, no. 7, pp. 808-824, 2005.
[30] Panigrahi S., Verma K., and Tripathi P., "Data Mining Algorithms for Land Change Detection: A Review," *Sādhanā Journal*, vol. 42, no. 12, pp. 2081-2097, 2017.
[31] Panigrahi S., Verma K., and Tripathi P., "An Efficient Approach to Detect Sudden Changes in Vegetation Index Time Series for Land Change Detection," *IETE Technical Review*, vol. 33, no. 5, pp. 539-556, 2016.
[32] Roy D., Lewis P., and Justice C., "Burned Area Mapping using Multi Temporal Moderate Spatial Resolution Data–a Bi-directional Reflectance Model based Expectation Approach," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 263-286, 2002.
[33] Synthetic Control Chart Time Series, Available: http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html, Last Visited, 2018.
[34] Taylor W., "Change-point Analysis: a Powerful New Tool for Detecting Changes," http://www.variation.com/cpa/tech/changepoint.html, Last Visited, 2000.
[35] Tan P., Steinbach M., and Kumar V., *Introduction to Data Mining*, Addison-Wesley, 2006.
[36] US Geological Survey and NASA. Land Processes Distributed Active Archive Center (LP DAAC). Available: https://lpdaac.usgs.gov/products/modis_products_table, Last Visited, 2018.
[37] Verbesselt J., Hyndman R., Newnham G., and Culvenor D., "Detecting Trend and Seasonal Changes in Satellite Image Time Series," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 106-115, 2010.

Sangram Panigrahi is currently pursuing Ph.D.program in Dept. of Computer Applications, NIT Raipur, India. His areas of research interests include data mining, NN and WSN. He has authored number of research papers in the above mentioned areas in international journals and conferences.

Kesari Verma Post graduated in 1998 and got doctorate in 2007. She is currently Assistant professor in National Institute of Technology, Raipur. She has 12 year R & D and Teaching experience. She has more than 30 publications in journals and conferences.

Priyanka Tripathy has received Ph.D. degree from MANIT, India, in 2009. She is currently working as Associate professor in the Dept. of Comp. Engg. and App., NITTTR, Bhopal, India. She has around 15 years of teaching and research experience. Her research interest includes web engg.,, ERP, NN & fuzzy logic, data mining and S/W engineering.