Search-based semi-supervised clustering algorithms for change detection in remotely sensed images

Conference Paper *in* International Journal of Knowledge Engineering and Soft Data Paradigms · December 2012

DOI: 10.1109/INDCON.2012.6420670

CITATIONS
2

READS
99

3 authors:

Dr. Moumita Roy
Indian Institute of Information Technology Guwahati
14 PUBLICATIONS **207** CITATIONS
SEE PROFILE

Susmita Ghosh
Jadavpur University
109 PUBLICATIONS **2,047** CITATIONS
SEE PROFILE

Ashish Ghosh
Indian Statistical Institute
250 PUBLICATIONS **6,185** CITATIONS
SEE PROFILE

Search-based Semi-supervised Clustering Algorithms for Change Detection in Remotely Sensed Images

Moumita Roy, Susmita Ghosh

Department of Computer Science and Engineering
Jadavpur University
Kolkata 700032, INDIA
Email: moumita2009.roy@gmail.com
susmitaghoshju@gmail.com

Ashish Ghosh

Center for Soft Computing Research
Indian Statistical Institute
Kolkata 700108, INDIA
Email: ash@isical.ac.in

***Abstract*—In real life change detection for remotely sensed images suffers due to the problem of inadequate labeled patterns. When a few labeled patterns can be collected by experts, semi-supervised (learning) clustering can be opted for change detection instead of the unsupervised approach to make full utilization of both labeled and unlabeled patterns. In the present work, a study has been carried out by applying some of the semi-supervised clustering techniques for changed detection. A comparative analysis between K-Means, COP-KMeans, Seeded-KMeans and Constrained-KMeans algorithms is being performed based on the results obtained using two multi-temporal remotely sensed images. It can be concluded from the experiments that the Constrained-KMeans is well suited for changed detection of remotely sensed images under semi-supervised framework.**

***keywords* - Change detection, multi-temporal images, semi-supervised clustering.**

I. Introduction

Change detection [1], [2] is a process of detecting temporal effects of multi-temporal images. This process is used for finding out changes in a land cover over time by analyzing remotely sensed images of a geographical area captured at different time instants. The changes can occur due to natural hazards (*e. g.* , disaster, earthquake), urban growth, deforestation etc. There are various applications of change detection like analysis of damage assessment, land use change analysis, day/night analysis of thermal characteristics, burned area identification etc.

The problem of change detection can be viewed as an image segmentation one where two groups of pixels are formed, one for the changed class and the other for the unchanged one. Methodology of change detection can be broadly classified into two categories: supervised [3] and unsupervised [4]-[6], [7]. Supervised technique has certain advantages like it can explicitly recognize the kinds of changes occurred and it is robust to different atmospheric and light conditions at two acquisition dates. Various methods exist in literature to carry out supervised change detection techniques *e. g.* , post classification method [1], direct multi-date classification method [1], kernel based methods [3] etc. The applicability of supervised methods in change detection is limited owing to the mandatory requirement of sufficient amount of ground truth information which is expensive, hard and monotonous to collect. On the contrary, in unsupervised approach, there is no need of additional information like ground truth. Due to the depletion of labeled patterns, unsupervised techniques seem to be compulsory for change detection. Generally, three consecutive steps are followed for unsupervised change detection. These are image preprocessing, image comparison and image analysis [1]. Images of the same geographical area, captured at different time instants, constitute the input of the change detection process. In preprocessing step, these images are made compatible by operations like radiometric and geometric corrections, co-registration, noise reduction etc. After preprocessing, image comparison is carried out, pixel by pixel, to generate a difference image (DI) which is subsequently used for change detection. The difference image can be generated in many ways *e. g.* , univariate image differencing, change vector analysis (CVA), image ratioing [1]. In the present work, CVA technique [1] is used for creating of DI. Unsupervised change detection process can be classified into two categories: context insensitive [1] and context sensitive [4]-[8]. Histogram thresholding [1], [9] is the simplest unsupervised context insensitive change detection method which has the limitations of not considering the spatial correlation between neighborhood pixels in the decision process. To overcome this difficulty, context sensitive methods using Markov random field (MRF) [8] are developed. These techniques also suffer from certain difficulties like requirement of the selection of a proper model for statistical distribution of the changed and the unchanged class pixels. Change detection methodologies based on neural networks, both using supervised and unsupervised learning [4], [5], are found in literature which are free from such limitations.

In change detection, a situation may occur where the categorical information of a few labeled patterns can be collected easily by experts. If the number of these labeled

patterns is low, then this information may not be sufficient enough for developing supervised methods. In such a scenario, knowledge of labeled patterns will be completely unutilized if unsupervised approach is carried out. Under this circumstance, semi-supervised approach [10], [11] can be opted for instead of unsupervised or supervised one. Semi-supervision has been used successfully for improving the performance of clustering and classification [12] when insufficient amount of labeled data are present. Semi-supervised classification uses abundant unlabeled patterns for training along with a few labeled patterns, whereas semi-supervised clustering utilizes a few labeled patterns for finding out more accurate clusters. Semi-supervised clustering techniques can be carried out in two ways: search-based approach and similarity-based approach [13]. Search-based approach uses the labeled patterns to search for an accurate partitioning. In similarity-based approach, the labeled patterns are utilized to adopt the underlying similarity metrics.

In the present work, we consider the search-based semi-supervised clustering approach. Here, three variants of semi-supervised K-Means algorithm namely, COP-KMeans [14], Constrained-KMeans [12] and Seeded-KMeans [12], have been studied for change detection problem. Comparative analysis between these techniques and the standard K-Means algorithm [15] is also carried out for two multi-temporal remotely sensed images. It is found that Constrained-KMeans algorithm is more suitable for change detection under the scarcity of labeled patterns.

II. Generation of Input Pattern

The difference image $D = \{l_{mn}, 1 \leq m \leq p, 1 \leq n \leq q\}$ is produced by the CVA technique [1] from the two co-registered and radiometrically corrected γ -spectral band images Y_1 and Y_2 of size $p \times q$ of the same geographical area at different times T_1 and T_2. Here, gray value of the difference image D at spatial position (m, n), denoted as l_{mn}, is calculated as,

$$l_{mn} = (int)\sqrt{\sum_{\alpha=1}^{\gamma}(l_{mn}^{\alpha}(Y_1) - l_{mn}^{\alpha}(Y_2))^2},$$

where, $l_{mn}^{\alpha}(Y_1)$ and $l_{mn}^{\alpha}(Y_2)$ are the gray values of the pixels at the spatial position (m, n) in the α^{th} band of the images Y_1 and Y_2, respectively.

From the difference image D, the input pattern for a particular pixel position is generated by considering the gray value of the said pixel as well as those of its neighboring ones to exploit (spatial) contextual information from neighbors. In the present methodology, 2^{nd} order neighborhood system is used. Here, each input pattern consists of nine features, one gray value of its own and eight gray values of its eight neighbors. Here, the y-dimensional input pattern of the $(m, n)^{th}$ pixel position of DI is denoted by $\overrightarrow{X_{mn}} = [x_{mn,1}, x_{mn,2}, ..., x_{mn,y}]$.

III. Background: Methodologies used

Semi-supervised variants of K-Means algorithm utilize insufficient labeled information either in the form of seed data or constraint during the iterative partitioning process of standard K-Means algorithm. A brief description of the standard K-Means, COP-KMeans, Constrained-KMeans and Seeded-KMeans algorithms is given in the following subsections.

A. K-Means algorithm

In K-Means algorithm [15], initially, C patterns (number of clusters) are randomly selected from a set of unlabeled patterns and they corresponds to initial cluster centers. Let, $v_1, v_2, ..., v_C$ represent these C cluster centers. In each iteration, the Euclidean distance of the unlabeled patterns from each of the cluster center is computed. Then, the unlabeled pattern is assigned to the cluster for which the distance measure is minimum. After that, each cluster center, v_i is updated by the arithmetic mean of the patterns (feature wise) assigned to the i^{th} cluster. This process (partitioning and assignment) continues until the following objective function is minimized:

$$O_{kmeans} = \sum_{l=1}^{C} \sum_{X_{mn} \in \chi_l} ||X_{mn} - v_l||; \quad (1)$$

where, χ_l is the set of patterns assigned to the cluster l. This process is stopped when no changes occur from the partitioning point of view.

B. COP-KMeans algorithm

Wagstaff et al. proposed this algorithm in 2001 [14]. Here, the labeled information is used during the partitioning process of K-Means algorithm in the form of "must link" and "cannot link" constraints. "Must-link" constraint ensures that a pair of patterns must be in the same group. On the contrary, "cannot-link" constraint specifies that the said two patterns can not belong to the same group. To apply COP-KMeans algorithm for the change detection process, we require some constraints. In the present technique, for experimental purpose, labeled patterns are picked up from the ground truth for both the groups with equal percentage. After that, from the pool of these labeled patterns, each combination of the pattern pair is considered. Now, if they are in the same class, then "must-link" constraint is generated. Otherwise, "cannot-link" constraint is generated.

In this algorithm, both the constraints must be satisfied for assigning a pattern to a particular cluster. In the standard K-Means algorithm, it is mentioned that an unlabeled pattern is assigned to the nearest cluster. On the contrary in this algorithm, before assigning an unlabeled pattern to a specific cluster, a sorted list of clusters (in ascending order based on the distance of a pattern from each of the cluster centers) is generated for each of the unlabeled patterns . Initially, the first one from the sorted list is selected and an unlabeled pattern is assigned to the corresponding cluster if no constraints are violated. This means, the patterns, those are already assigned in that cluster, are not in "cannot-link" constraint with the unlabeled pattern. Also the patterns, those are already assigned in different clusters, are not in the "must-link" constraint with the unlabeled pattern. If any of the constraints is violated, the

next cluster in the sorted list is checked for assignment. This process is continued until a valid cluster is found or the list is exhausted. If no valid cluster is obtained then it can be said that the partitioning is not possible with the initial cluster centers without violating the constraints. After the assignment, the rest of the steps are similar to those of the standard K-Means algorithm.

C. Seeded-KMeans algorithm [12]

As mentioned earlier, in the standard K-Means algorithm, the initial cluster center is randomly chosen from a set of patterns. While, in Seeded K-Means algorithm, labeled patterns (seed data) are utilized for assigning the initial cluster center. The i^{th} cluster center is initialized by the arithmetic mean of the labeled pattern belongs to the i^{th} cluster. In this algorithm, labeled patterns are only used in initialization step. After that, this class label information may be changed during the iterative clustering process.

D. Constrained-KMeans algorithm [12]

In Constrained-KMeans algorithm, the initial cluster centers are assigned in the same way as it is done in Seeded K-Means algorithm. But, in this algorithm, labeling of the seed data is not re-estimated during the iterative clustering procedure.

IV. Description of data sets

To evaluate the effectiveness of the proposed methodology, experiments are carried out on two multi-temporal remotely sensed images corresponding to the geographical areas of Mexico and Sardinia Island of Italy.

A. Data set related to Mexico area

This data set consists of two multi-spectral images of the Landsat-7 satellite captured by the Landsat Enhanced Thematic Mapper Plus (ETM+) sensor over an area of Mexico taken on 18^{th} April, 2000 and 20^{th} May, 2002. From the entire available Landsat scene, a section of 512×512 pixels has been selected as test site. A fire destroyed a large portion of the vegetation in the considered region between two acquisition dates. Initially, we performed some trials in order to determine the most effective spectral bands for detecting the burnt area in the considered data set. On the basis of the results of these trials, band 4 is observed to be more effective to locate the burnt area. Figures 1(a) and 1(b) show the band 4 images corresponding to April, 2000 and May, 2002. The difference image (Figure 1(c)) created by spectral band 4 using CVA technique is only used for further analysis. For evaluation of the proposed approach, a reference map (Figure 1(d)) was used. The reference map contains 25599 changed and 236545 unchanged pixels.

B. Data set related to Sardinia Island, Italy

Two multi-spectral images are acquired by the Landsat Thematic Mapper (TM) sensor of the Landsat-5 satellite in September, 1995 and July, 1996. The test site of 412×300 pixels of a scene includes the lake Mulargia on the Island of Sardinia (Italy). The water level of the lake increased (see lower center part of the image) between two acquisition dates. Figures 2(a) and 2(b), respectively, show the 1995 and 1996 images of band 4. We applied CVA technique on spectral bands 1, 2, 4, and 5 of the two multi-temporal images to generate the difference image (Figure 2(c)), as elementary experiments show that the above channels contain useful information on the changes of water body. In the reference map (Figure 2(d)), 7480 changed and 116120 unchanged pixels were identified.

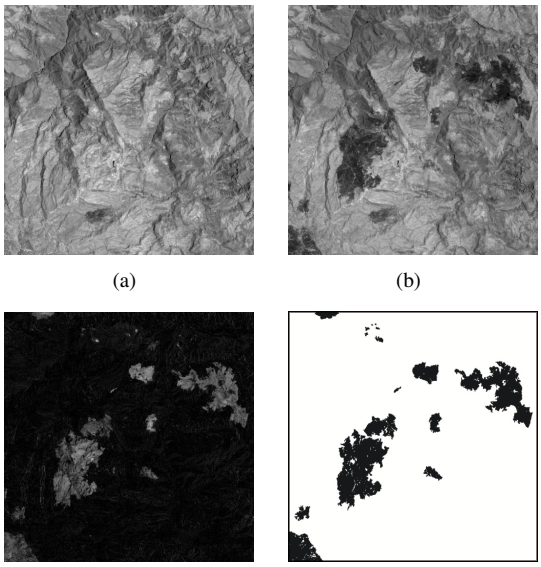
V. Results and Analysis

As already mentioned, to investigate the effectiveness of the standard K-Means, COP-KMeans, Constrained-KMeans and Seeded-KMeans algorithms, experiments are conducted on two multi-temporal remotely sensed images. For experimentation, different percentages of training patterns (from 0.5% to 5%) are considered and 10 simulations are conducted in each case. For typical illustrations, the results corresponding to three different percentages (0.5%, 2%, and 5%) are given. Performance measuring indices used are the number of missed alarms (changed class pixels identified as unchanged ones-MA), the number of false alarms (unchanged class pixels classified as changed ones-FA), the number of overall error (OE) and Kappa measure (KM) [16]. The best result in terms of minimum overall error (over 10 simulations) is depicted in the tables. Average CPU time (over 10 simulations) is also considered for comparison. Results of Mexico data set and Sardinia data set are put in Table I and Table II, respectively.

From the Tables I and II, it is observed (for both the data sets) that COP-KMeans algorithm (considering all the percentages of training patterns used) is significantly better than the standard K-Means algorithm in terms of all the measuring indices except the CPU time requirement. It has also been noticed that the required CPU time is increased by increasing the number of labeled patterns or the constraints. This may be due to the time required for checking of the constraint violation in each case before assignment.

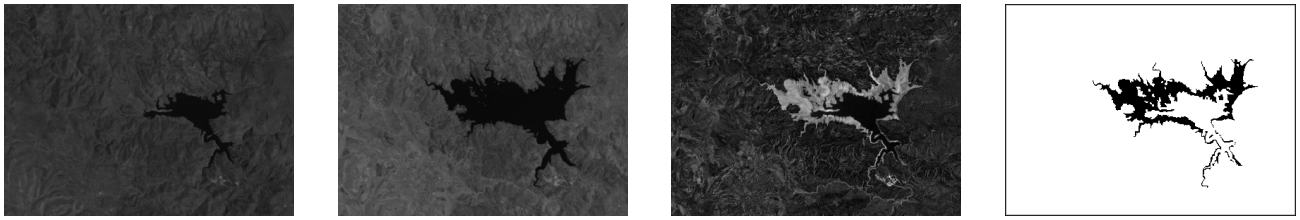
From the Tables I and II, it has also been noticed that Seeded-KMeans algorithm is better than the standard K-Means algorithm in terms time requirement. During experimentation with standard K-Means algorithm (different simulations with various initial cluster centers), it has been observed that the results are not very much sensitive to the choice of the initial cluster centers. So, it is obvious that Seeded K-Means algorithm can not attain the betterment over K-Means algorithm, because in Seeded-KMeans algorithm, labeled patterns are only used for assigning initial cluster centers. Thereafter, the rest of the steps are identical. Due to more accurate initial cluster center assignment than random guessing, convergence time of this algorithm is quicker than that of the standard K-Means algorithm.

By analyzing the tables, it can be concluded that Constrained-KMeans algorithm is integrating the superiority of both COP-KMeans algorithm (in terms of different performance measuring indices used) and Seeded-KMeans algorithm (*w. r. t.* CPU time requirement). In short, Constrained-KMeans



(a) (b) (c) (d)

Fig. 1. Images of Mexico area. (a) Band 4 image acquired in April 2000, (b) band 4 image acquired in May 2002, (c) corresponding difference image, and (d) a reference map of the changed area.



(a) (b) (c) (d)

Fig. 2. Images of Sardinia Island, Italy. (a) Band 4 image acquired in September 1995, (b) band 4 image acquired in July 1996, (c) difference image generated by CVA technique using bands 1, 2, 4, & 5, and (d) a reference map of the changed area.

TABLE I
RESULTS ON MEXICO DATA SET

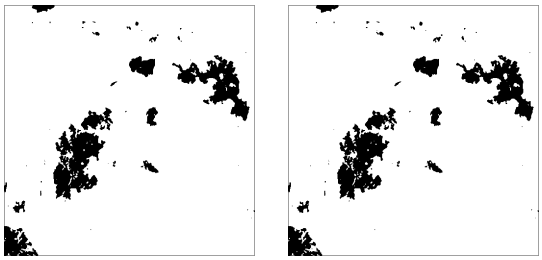
Techniques used	Training patterns	MA	FA	OE	KM	Time (in second)
K-Means	-	3107	665	3772	0.914725	9.7
COP-KMeans	0.5%	3077	661	3738	0.915534	33.3
	2%	3016	657	3673	0.917088	194.3
	5%	2868	645	3513	0.920895	221.9
Seeded-KMeans	0.5%	3107	665	3772	0.914725	6.8
	2%	3107	665	3772	0.914725	6.4
	5%	3107	665	3772	0.914725	6.5
Constrained-KMeans	0.5%	3077	661	3738	0.915534	6.1
	2%	3016	657	3673	0.917088	7.3
	5%	2868	645	3513	0.920895	5.9

TABLE II
RESULTS ON SARDINIA DATA SET

Techniques used	Training patterns	MA	FA	OE	KM	Time (in second)
K-Means	-	637	1879	2516	0.833887	7.1
COP-KMeans	0.5%	635	1865	2500	0.834828	21.8
	2%	638	1792	2430	0.838742	62.8
	5%	622	1698	2320	0.845338	105.8
Seeded-KMeans	0.5%	637	1879	2516	0.833887	4.6
	2%	637	1879	2516	0.833887	4.1
	5%	637	1879	2516	0.833887	4.5
Constrained-KMeans	0.5%	635	1858	2493	0.835224	5
	2%	638	1791	2429	0.838799	4.8
	5%	622	1693	2315	0.845626	4

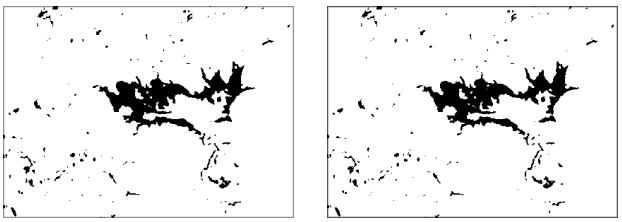
algorithm is able to produce better results by consuming comparable amount of CPU time. In this algorithm, constraints or labels are fixed for labeled patterns during iterative partitioning process. So, there is no need to check for violating the constraints resulting in less CPU time requirement. Though, the values of performance measuring indices (considering all the percentage of training patterns) for Mexico data Set (in Table I) are similar to those of the COP-KMeans algorithm but those are obtained in lesser time than COP-KMeans algorithm. For Sardinia data set (in Table II), it has been observed that the values of performance measuring indices are also better using Constrained-KMeans algorithm than using COP-KMeans algorithm.

For visual illustration, the change detection maps corresponding to minimum overall error (obtained over 10 simulations) for Mexico and Sardinia data sets, using K-Means algorithm and Constrained-KMeans algorithm are depicted in Figures 3 and 4, respectively. Figure 3(a) shows the map obtained using K-Means algorithm while Figure 3(b) shows the map using Constrained-KMeans algorithm (with 0.5% training pattern from both the classes) for Mexico data set. Corresponding change detection maps for Sardinia data set are displayed in Figures 4(a) and 4(b), respectively. It has been observed that the maps obtained using Constrained-KMeans algorithm are more accurate resemblance of the reference map.



(a) (b)

Fig. 3. Change detection maps obtained for Mexico data set: (a) using K-Means algorithm, and (b) using Constrained-KMeans algorithm (with 0.5% training pattern)



(a) (b)

Fig. 4. Change detection maps obtained for Sardinia data set: (a) using K-Means algorithm, and (b) using Constrained-KMeans algorithm (with 0.5% training pattern)

VI. Conclusion

In this paper, three variants of search-based semi-supervised K-Means algorithms are studied for change detection technique incorporating a few labeled information. From the results, it has been observed that Constrained-KMeans algorithm is more appropriate for change detection under semi-supervised clustering framework in terms of both CPU time requirement and other performance measuring indices, used in the work.

VII. Acknowledgments

Moumita Roy is grateful to Council of Scientific & Industrial Research (CSIR), India for providing her a Senior Research Fellowship [No. 09/096(0684)2k11-EMR-I].

References

[1] A. Singh, "Digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] M. J. Canty, *Image Analysis, Classification and Change Detection in Remote Sensing*. Taylor & Francis: CRC Press, 2006.

[3] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Mari, J. L. Rojo-Álvarez, and M. Martinez-Ramón, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 46, no. 6, pp. 1822–1835, 2008.

[4] S. Ghosh, L. Bruzzone, S. Patra, F. Bovolo, and A. Ghosh, "A context-sensitive technique for unsupervised change detection based on Hopfield-type neural networks," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 45, no. 3, pp. 778–789, 2007.

[5] S. Ghosh, S. Patra, and A. Ghosh, "An unsupervised context-sensitive change detection technique based on modified self-organizing feature map neural network," *International Journal of Approximate Reasoning*, vol. 50, no. 1, pp. 37–50, 2009.

[6] A. Ghosh, N. S. Mishra, and S. Ghosh, "Fuzzy clustering algorithms for unsupervised change detection in remote sensing images," *Information Sciences*, vol. 181, no. 4, pp. 699–715, 2011.

[7] N. S. Mishra, S. Ghosh, and A. Ghosh, "Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images," *Applied Soft Computing*, vol. 12, no. 8, pp. 2683–2692, 2012.

[8] T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on Markov random field models," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 40, no. 8, pp. 1815–1823, 2002.

[9] S. Patra, S. Ghosh, and A. Ghosh, "Histogram thresholding for unsupervised change detection of remote sensing images," *International Journal of Remote Sensing*, vol. 32, no. 21, pp. 6071–6089, 2011.

[10] X. Zhu, *Semi-supervised Learning Literature Survey*, Computer Sciences TR1530, University of Wisconsin, Madison, 2008.

[11] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised Learning*. Cambridge: MIT Press, 2006.

[12] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proceedings 19th International Conference on Machine Learning (ICML-2002)*, C. Sammut and A. G. Hoffmann, Eds. Sydney, Australia: Morgan Kaufmann, San Fransisco, USA, 2002, pp. 27–34.

[13] S. Basu, M. Bilenko, and R. J. Mooney, "Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering," in *Proceedings of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining Systems*, Washington DC, USA, 2003, pp. 42–49.

[14] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means clustering with background knowledge," in *Proceedings 18th International Conference on Machine Learning (ICML-2001)*, C. E. Brodley and A. P. Danyluk, Eds. Williamstown, MA, USA: Morgan Kaufmann, San Fransisco, USA, 2001, pp. 577–584.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. Statistical Laboratory of the University of California, Berkeley: University of California Press, Berkeley, Calif, 1967, pp. 281–297.

[16] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 2nd ed. Boca Raton, USA: CRC Press, 2009.