Clustering methods to asses land cover samples of MODIS vegetation indexes time series

Lorena A. Santos, Rolf E. O. Simoes, Karine R. Ferreira, Gilberto R. de Queiroz, Gilberto Camara, Rafael D. C. Santos

National Institute for Space Research - INPE
Av. dos Astronautas, 1758, 12227-010
São José dos Campos - SP - Brazil
{lorena.santos, rolf.simoes, karine.ferreira, gilberto.queiroz, gilberto.camara, rafael.santos}@inpe.br

Abstract. MODIS vegetation indexes time series have been widely used to build land cover change maps on large scales. In this scope, to obtain good quality maps using supervised classification methods, it is crucial to select representative training samples of land cover change classes. In this paper, we evaluate two clustering methods, Hierarchical and Self-Organizing Map (SOM), to assess land cover samples of MODIS vegetation indexes time series. As we show, these techniques are suitable tools for assisting users to select representative land cover change samples from MODIS vegetation indexes time series. We present the accuracy of both methods for a case study in Ipiranga do Norte municipality in Mato Grosso state, Brazil.

Keywords: time series clustering, MODIS vegetation indexes, land cover change classification, self-organizing map (SOM)

1 Introduction

Remote sensing images play a crucial role in land cover change classification on global and continental scales. Recently, time series of vegetation indexes, such as NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index), from MODIS (Moderate Resolution Imaging Spectroradiometer) products have been widely used to build land cover change maps on large scales [1, 2, 4, 3].

Aguiar et.al [1] identify pasture land and its different levels of degradation in Mato Grosso do Sul state, Brazil, using MODIS NDVI time series and a J48 classifier with wavelet technique. Arvor et.al [2] use MODIS EVI time series to quantify the evolution of agricultural area from 2000 to 2006 in Mato Grosso state, Brazil. Bagan et.al [3] propose an approach to classify land cover from MODIS EVI time series using the Self-Organizing Map (SOM) neural network technique and present a case study in eastern China from March to December in 2002. Maus et.al [4] propose an algorithm called Time-Weighted Dynamic Time Warping (TW-DTW), based on the classical Dynamic Time Warping (DTW) method, for land cover and land use classification and present a case study using MODIS EVI time series in the Porto dos Gaúchos municipality in Mato Grosso state, Brazil.

In general, supervised classification methods require a training step, which consists in gathering training samples to represent the classes to be identified. The quality of such samples is crucial in the classification process. Representative samples of classes lead to good classification results. Therefore, specially in land cover change classification, there is a need for techniques that help users to select representative land cover change samples from vegetation indexes time series of remote sensing images.

Remote sensing time series usually contain interferences of spatio-temporal phenomena that impact on land use and land cover monitoring, i.e. atmospheric influences or cloud cover. NDVI index has several limitations that affect the accuracy of classification, including atmospheric conditions. To account for these limitations, EVI index was proposed to remove residual atmosphere contamination [7]. However, Heute et. al [6] studied vegetation index products and argued that EVI and NDVI of the biotic formation in different regions were sensitive to seasonal change, land cover change and biophysical parameters change. They demonstrated that NDVI was highly relevant to EVI and the value of NDVI was always bigger than EVI when the soil background and the atmospheric aerosol vary less [6].

In this context, time series clustering techniques can assist in an exploratory analysis to evaluate which vegetation indexes are better to select representative land cover change samples. Representative samples allow to extract temporal patterns that concern the seasonal periodicity of vegetation. Thus, such techniques can improve the training step of the land cover change classification.

It is important to develop an appropriate and validation scheme to assess the performance and limitations of clustering algorithms. In this paper, we present a ground truth based comparative study to evaluate the accuracy and performance of two clustering methods, Hierarchical and Self-Organizing Map (SOM), for assessing the separability of land cover samples of MODIS vegetation indexes time series [5, 10, 11]. In this work, we present the accuracy of these two clustering methods for a case study in Ipiranga do Norte municipality in Mato Grosso state, Brazil.

2 Background

In this section, we present concepts and algorithms used in our study: the Self-Organizing Map (SOM) Neural Network, the Dynamic Time Warping Distance (DTW) and the Hierarchical Clustering Algorithm.

2.1 SOM Neural Network

A SOM (Self-Organizing Map) is an unsupervised neural network that consists in competitive learning for providing a topology-preserving mapping from a high-dimensional input onto a low-dimensional output. The structure of a SOM is composed by input and output layers. The training data or input data are in the input layer whereas the output layer is formed by a set of neurons that are trained to extract patterns from the input data [5].

An important property of SOM is the neighborhood relationship among neurons in the output layer, i.e., vectors in the input layer with similar characteristics can be mapped into either a neuron that represents those characteristics or neighboring neurons in the output layer [3].

Each neuron j in the set of J neurons has a n-dimensional weight vector $w_j = [w_{j1}, \ldots, w_{jn}]$ associated to it. At each training step t, an input vector $x(t) = [x(t)_1, \ldots, x(t)_n]$ is randomly chosen, and then the Euclidean distance D_j is calculated between this input vector and each neuron j for all the neurons in the output layer (equation 1).

$$D_j = \sum_{i=1}^{N} \sqrt{(x(t)_i - w_{ji})^2}. \tag{1}$$

The next step is to determine the Best-Matching-Unit (BMU), i.e. the neuron d_b with weight vector closer to $x(t)$ (equation 2):

$$d_b = min\left\{D_1, \ldots, D_J\right\}. \tag{2}$$

The weight vector of the neuron chosen from the BMU is updated, i.e. adjusted to be closer to the input vector (equation 3. The weights of the neurons $N_b(t)$, neighbors of the BMU, are also updated with a smaller weight.

$$w_{ji}(t+1) = w_{ji}(t) + \alpha(t)[x(t)_i - w_{ji}(t)], \tag{3}$$

In equation 3, $\alpha(t)$ is the learning rate, set as $0 < \alpha(t) < 1$. An iteration ends when all vectors of the input layer are trained, then $\alpha(t)$ must be reduced [8]. The number of iterations must be high in order to allow the neurons to fit accurately to the data sets [9].

2.2 Dynamic Time Warping Distance

Dynamic Time Warping (DTW) is a classical algorithm that produces the most robust distance used to align two time series, allowing the alignment of similar sequences that match even if they are out of phase in the time axis [13].

Consider two time series $Q = [q_1, \ldots, q_i, \ldots, q_n]$ and $C = [c_1, \ldots, c_i, \ldots, c_m]$. The first step of DTW is to compute a cost matrix, $n \times m$, given by the squared distance between the elements of the two time series:

$$\Psi_{i,j} = (q_i - c_j)^2 \tag{4}$$

From Ψ, we can find the best matching between two time series, getting an optimal path that minimizes the cost warping. This warping path can be found using dynamic programming, transforming a complex global problem into a number of local optimization subproblems [11].

$$d_{i,j} = \Psi_{i,j} + \min \begin{cases} d_{i-1,j} \\ d_{i-1,j-1} \\ d_{i,j-1} \end{cases} \tag{5}$$

2.3 Hierarchical Clustering

Hierarchical Clustering is another well-known method used to cluster data points. There are two types of hierarchical clustering: agglomerative and divisive. In this paper we use the agglomerative type, where each sample starts in its own cluster, and the clusters are then grouped with large clusters based on linkage criteria, until all samples are contained in a single cluster.

The linkage criterion determines the distance between sets of data as a function of the pairwise distances between the data [12]. There are several linkage criteria, some of them are presented next.

Ward's criterion merges two clusters that result in the smallest increase in the value of the sum-of-squares variance. At each clustering step, all possible mergers of two clusters are tried. The sum-of-squares variance is computed for each cluster, and the one with the smallest value is selected [11]. Other popular linkage criteria are *average*, *single* and *complete*. All these are used to determine which pair of clusters are going to be merged in the next step of the hierarchical algorithm: the *average* criterion calculates, for each pair of clusters, the average distance between all data points in each cluster; the **single** criterion calculates the distance between two clusters A and B as $Dist(A, B) = \min_{a,b} d(a, b)$, and the **complete** criterion calculates the distance as $Dist(A, B) = \max_{a,b} d(a, b)$ [12]. For each criterion, the smallest distance between the two clusters is selected and the clusters are merged, and the process repeated until there is only one cluster with all data points.

A dendrogram can be used to visualize the hierarchy obtained from the hierarchical clustering method. The dendrogram helps visualization of the merging of the clusters, and can be used to evaluate the height in where the largest change in dissimilarity occurs, so it can be cut at such height for the clusters extraction. It is also possible to specify the number of clusters and then cut the dendrogram in such a way that the chosen number is obtained [11, 10].

3 Materials and Methods

3.1 Data

The data used in this study was extracted from MODIS sensor of the Terra satellite developed by NASA. This sensor monitors the state of Earth's environment. The MOD13Q1 product from MODIS provides per pixel values of vegetation indexes. These indexes are used for global monitoring of vegetation conditions and land cover classifying on large spatial scales [14]. In this product, there are two vegetation layers, the Normalized Difference Vegetation Index (NDVI) and

Enhanced Vegetation Index (EVI) [15]. Time series MODIS data, used in this paper, produce vegetation indexes at each 16-day with 250-meter spatial resolution. During plant growth periods, different vegetation styles can be distinguished by vegetation indexes time series [17].

The study area comprises a region of $9.6km \times 8km$ and it is located in Ipiranga do Norte (*Mato Grosso*, Brazil) municipality as shown in Figure 1.

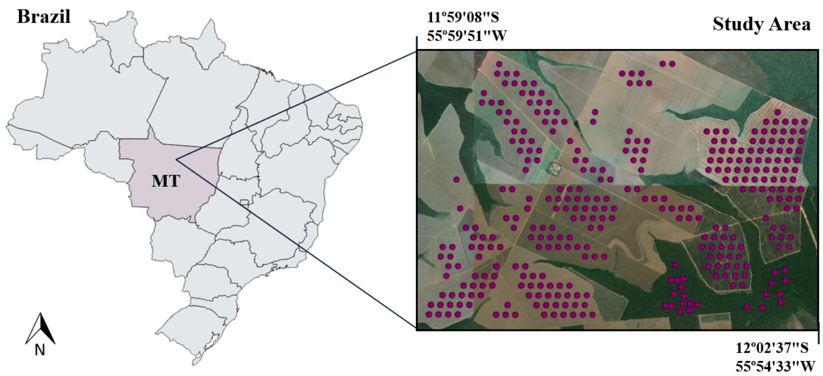

Fig. 1. The area of study corresponds to approximately $76.8Km^2$ in the Ipiranga do Norte municipal

The study region was chosen considering the existence of a dataset containing 603 ground truth sample points, from 2007 to 2013. This dataset is organized in five classes: 138 samples for "forest", 68 for "cotton-fallow", 79 for "soybean-cotton", 134 for "soybean-maize" and 184 for "soybean-millet". Each data sample has the spatial location (latitude and longitude), the start and end dates, with an one year interval, and the label representing the class.

From the ground truth, the MODIS vegetation indexes time series of each sample was extracted, for each spatial location and date. In total, we have collected 603 one-year-spanned time series from different years. Figure 2, shows the EVI and NDVI time series between 2011 and 2012. This time series correspond to the point lat=-12.0385, lon=-55.9844 for the "cotton-fallow" class.

3.2 Analysis

In order to evaluate the accuracy of the clustering result we use the Shannon information entropy [16] over all clusters and its capability of representing one class. If there is confusion between two or more known classes belonging to the same cluster then the entropy metric will increase up to a maximum. The entropy e can be obtained for each cluster taking into account all its classes frequencies p_i in each cluster, using equation 6.

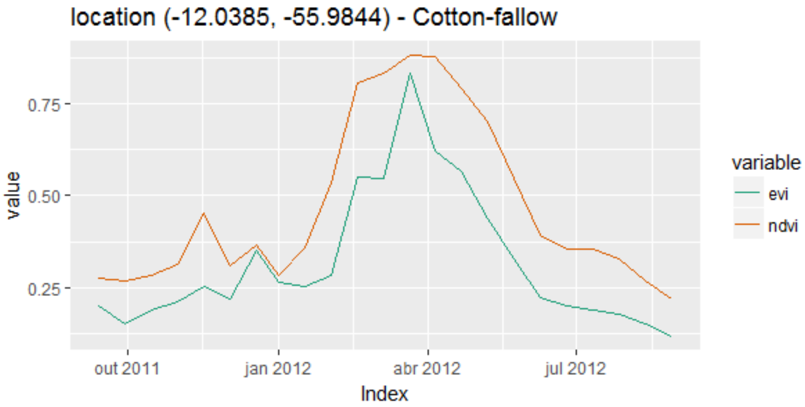

Fig. 2. EVI and NDVI Time series from a random chosen sample point.

$$e = -\sum_i p_i \log p_i \tag{6}$$

This gives us a measure of separability for each cluster and informs how much a set of samples are distinguishable. In order to get a range from 0 to 1 we just divide the Eq. (6) by the logarithm of the number of classes obtaining a relative entropy metric. To compute an overall metric, we just have taken the weighted average of all clusters with respect to the amount of samples linked to it.

Differently from internal cluster validity indexes (e.g. Silhouette) that aims to measure compactness and separation between clusters, entropy aims to measure the clusters from a ground truth and can be viewed as an external cluster validity index [18]. This choice may overcome the disadvantage of internal cluster validity indexes in a high dimensionality context, such as time series data, due to the curse of dimensionality [19].

The clusters were produced by two methods. In the first method, the hierarchical clustering (**HC**) over a dissimilarity matrix among all sampled time series using DTW distance. Hence, a dendrogram can be computed according to a linkage method and once produced, the dendrogram can provide any number of clusters, from one to the number of samples, just informing a dissimilarity parameter: those samples with dissimilarity bellow that parameter will be tied together in a same cluster and all those samples above that value will pertaining others clusters.

In the second method, which uses the SOM as a preprocessing step of the HC (**SOM+HC**), we conducted a SOM neural network before the hierarchical clustering. In SOM, a neuron, through a weight vector represents a pattern of samples. These patterns are organized into a meaningful two-dimensional order in which similar models are closer to each other in the grid than the more dissimilar ones and the neighboring models are mutually similar,in this way, a neuron contains all the samples which are mapped to it. This step allows the samples to be represented by these patterns. As a result, we have obtained a set of representative neurons that were used as the input of the hierarchical clustering step. From them, we compute a dissimilarity matrix using DTW distances, and what follows is similar with the first method. Hence, the main difference here is that the hierarchical clustering is being applied to a reduced but representative number of time series.

Our experiments were conducted using some variations of the methods described in previous sections, with different parameters. For the **HC** method, we chose the parameters *number of clusters* (`n_clusters`), *vegetation index combination* (`bands`) and *linkage method* (`linkage`). For the **SOM+HC** method, the parameters chosen were *number of clusters*, *vegetation index combination*, *linkage method*, *SOM grid size* (`grid_size`), *SOM learning rates* (`learnr_init` and `learnr_fin`), and *SOM iteration steps* (`iterations`). The parameter's ranges were defined as described in Table 1.

Table 1. Experiments parameters' range

Parameter	Range	Methods
`n_clusters`	{3, 5, 7, 9}	HC and HC+SOM
`bands`	{EVI, NDVI, EVI&NDVI}	HC and HC+SOM
`linkage`	{Ward, average, complete, single}	HC and HC+SOM
`grid_size`	{49, 81, 121, 169}	HC+SOM
`learnr_init`	{0.1, 0.2, 0.3, 0.4}	HC+SOM
`learnr_fin`	{0.02, 0.04, 0.06, 0.08}	HC+SOM
`iterations`	{1000, 1200, 1400, 1600}	HC+SOM

All possible parameters values were combined producing 48 different experiments for the **HC** method and $12,288$ experiments for the **SOM+HC**. As output of both methods, we have got the same set of samples labeled with the corresponding computed cluster identification. This information allowed us to compute the separability level given by Equation 6. The methods are summarized in the Figures 3 and 4.

The experiments produced a database relating each parameter's values to the entropy dependent variable. The subsequent analysis were made on this dataset using descriptive statistics and correlation analysis in order to show the parameters-entropy behavior.

4 Results

In order to verify how the numeric variables correlate to the entropy we calculated the Pearson correlation. Both **HC** and **SOM+HC** methods showed a similar result. The correlation between the parameter `n_clusters` and `entropy` in the case of **HC** method was -0.55. For **SOM+HC** method, the `n_clusters` and `grid_size` correlations against `entropy` are -0.58 and 0.15, respectively (all

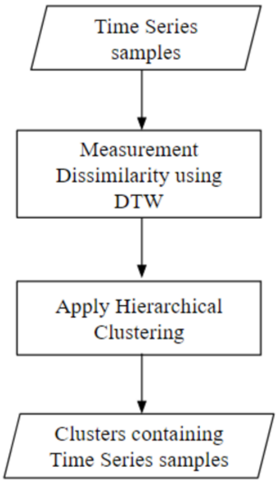

Fig. 3. Clustering Time Series Using Hierarchical clustering

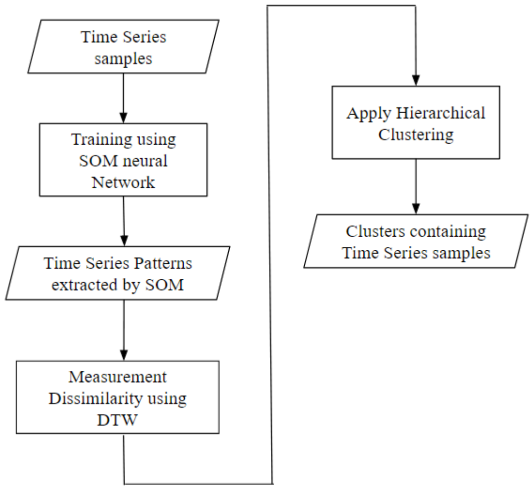

Fig. 4. Clustering Time Series Using SOM Neural Network and Hierarchical clustering

measures have a p-value less than 1.0×10^{-7}). The `n_clusters` correlation was expected as a fine grained clustering can capture more subtleties that the data may present by reducing the confusion and consequently the entropy for each cluster. However, the positive correlation between SOM grid size may suggest that this is not the case for SOM stage, at least in the range values used in the experiments.

The best achievement of **HC** experiments is shown in the Table 2, which parameters' values were `n_clusters`$= 5$, `bands`$= NDVI$, and `linkage`$= Ward$. The resulting entropy was 0.02782576 indicating a reasonably separability between classes using only the NDVI band. When considering EVI and NDVI bands, the lowest entropy (0.02872378) was achieved only by increasing the clusters to 9 with the same linkage criterion. The first non **Ward** linkage criterion with the lowest entropy comes at 7th position with 0.03011122.

All the **HC** experiments' results can be seen in Figure 5. As the number of clusters increases, the overall entropy stabilizes, suggesting an optimal `n_cluster` value. The graphs shows that **single** linkage criterion was outperformed in terms of separability and so are inadequate to our data samples. Maybe this is the case for all land use and land cover spectral data as different classes exhibit considerably variance and are, sometimes, very similar between them. The same is also observed in **SOM+HC** experiments where **single** linkage resulted in higher entropies. The Figure 6 depicts its entropy results for those experiments with `grid_size`$= 49$, `iterations`$= 1000$, `learnr_init`$= 0.1$, and `learnr_fin`$=$ 0.02. Figures 5 and 6 show that the number of clusters depends on the data bands and the linkage criterion. For example, using only the EVI band we have reached a minimum entropy with 7 clusters using Ward linkage while using the NDVI band the amount of clusters with best separability was 5 for the same linkage, suggesting that the NDVI band captures more differences among our sample classes.

Table 2. Separability matrix for the best HC clustering result. The resulting entropy was 0.02782576 from parameters' values `n_clusters`= 5, `bands`= $NDVI$, and `linkage`= $Ward$.

Classes	Clust.1	Clust.2	Clust.3	Clus.4	Clust.5
Forest	NA	NA	138	NA	NA
Cotton-Fallow	66	2	NA	NA	NA
Soybean-Cotton	3	76	NA	NA	NA
Soybean-Maize	NA	1	NA	133	NA
Soybean-Millet	NA	NA	NA	NA	184

The lowest entropy was achieved by **average** linkage criterion that outperformed **Ward** entropies only when the number of clusters were 9. The parameters used were: `n_clusters`= 9, `grid_size`= 49, `iterations`= 1000, `learnr_init`= 0.2, `learnr_fin`= 0.04, and `bands`=*EVI&NDVI*. The respective separability matrix is shown in Table 3.

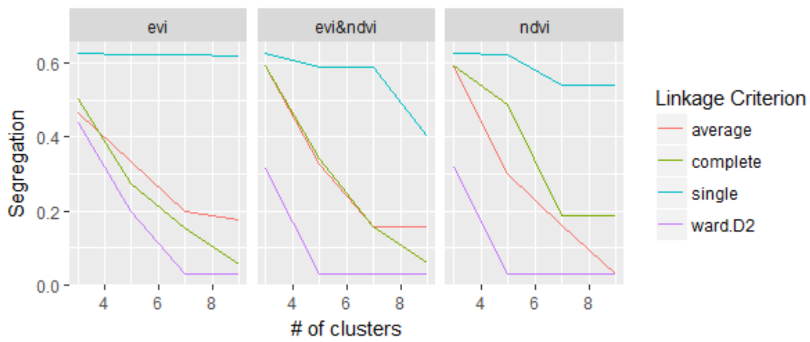

Fig. 5. HC experiments entropy results.

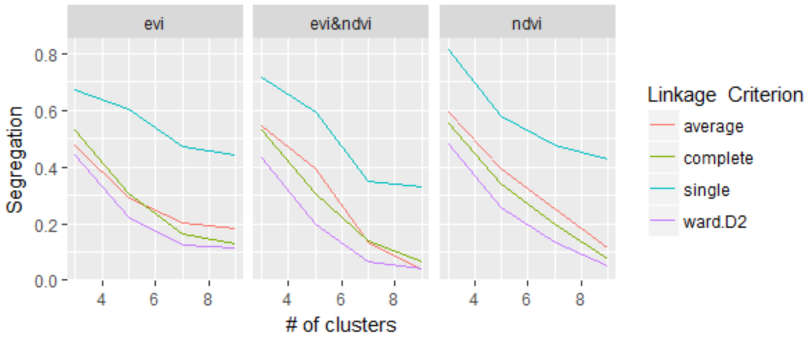

Fig. 6. **SOM+HC** experiments behavior for fixed parameters `grid_size`= 49, `iterations`= 1000, `learnr_init`= 0.2, `learnr_fin`= 0.04

Table 3. separability matrix of the best SOM+HC clustering separability result. The resulting entropy was 0.02872378 from parameters' values `n_clusters`= 9, `grid_size`= 49, `iterations`= 1000, `learnr_init`= 0.1, `learnr_fin`= 0.04, and `bands`=*EVI&NDVI*

Classes	Clust.1	Clust.2	Clust.3	Clust.4	Clust.5	Clust.6	Clust.7	Clust.8	Clust.9
Forest	NA	NA	25	113	NA	NA	NA	NA	NA
Cotton-Fallow	NA	NA	NA	NA	NA	NA	NA	NA	68
Soybean-Cotton	NA	NA	NA	NA	2	NA	NA	74	3
Soybean-Maize	NA	NA	NA	NA	NA	134	NA	NA	NA
Soybean-Millet	52	4	NA	NA	94	NA	34	NA	NA

We can see from the the best separability results (Tables 2 and 3) that some classes are mixed up inside a same cluster. This is the case of "cotton-fallow" and "soybean-cotton" for both methods. Specific investigations may provide some understanding why this may be the case for those classes or if we may consider discard the confusions cases as outliers or probably miss classified data sample.

Despite the fact that we can obtain the lowest relative entropy by setting `n_clusters` to the sample size, our results show that the relative entropy drops rapidly to an inflection point for a low value of `n_clusters` (1.5% of the sample size) after which it decreases very slowly. The existence of such inflexion point for a low value relative to the sample can give us a measure of data separability. However, an unexplored question is if such separability is stable.

5 Conclusions

Hierarchical clustering methods has disadvantage that the dissimilarity matrix is calculated for all given dataset. Then, for big dataset the process becomes expensive computationally, due to a complexity of $O(n^2)$, where n is the amount of data. Thus, hierarchical clustering may be prohibitive for large datasets.

In remote sensing context, using hierarchical clustering may become a problem due the large amount of data. A way to soften the complexity problem is to use iterative techniques such as SOM to reduce the amount of input data by creating patterns to represent those data. In this way, the dissimilarity matrix can be computed with much less data. Here we showed that the results presented by SOM method produced a good separability compared with hierarchical clustering approach.

Our experiments showed that vegetation indexes and linkage criterion interfere directly in the separability result. An entropy metric was used to assess class separability on the input data. The samples quality in terms of separability can be given by the inflexion point exhibited by the entropy against number of clusters. However, further investigations on stability of cluster validity index are needed.

References

1. Aguiar,D., Adami,M., Silva, W., Rudorff, B., Mello, M., Silva,J.: Modis time series to assess pasture land. In: Geoscience and Remote Sensing Symposium (IGARSS), (2010)
2. Arvor, D., Meirelles, M., Dubreuil, V., Begue,A., Shimabukuro, Y,.: Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices. In:Applied Geography, vol. 22, pp.702–713, (2012)
3. Bagan, H., Wang,Q., Watanabe, M., Yang, Y., MA, J., : Land Cover Classification From Modis EVI Time-series data using SOM neural network. In: International Journal of Remote Sensing, vol.26, pp.4999–5012, (2005)
4. Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, F., Ribeiro, G.: A Time-Weighted Dynamic Time Warping method for land cover mapping. In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 20, (2015)
5. Augustijn,E., Zurita-Milla,R.: Self-Organizing Maps as an approach to exploring spatiotemporal diffusion patterns. In: International Journal of Health Geographics, (2013)
6. Huete,A., Didan, K., Miura.T.: Overview of the radiometric and biophysical performance of the MODIS vegetation indices. In:Remote Sensing of Environment, vol.83. No. 1. pp.195–213 (2002)
7. Xiao, X., Braswell. B., Zhang. Q., Boles.S., Frolking.S., Moore.B.: Sensitivity of vegetation indices to atmospheric aerosols: continental-scale observations in Northern Asia.In: Remote Sensing of Environment. vol.84. pp. 385-–392
8. Natita,W., Wiboonsak,W., Dusadee,S.: Appropriate Learning Rate and Neighborhood Function of Self-Organizing Map (SOM) for Specific Classification over Southern Thailand. In: International Journal of Modeling and Optimization, vol.6. (2016)
9. Introdução a Mineração de Dados: com Aplicações em R. Silva, A. L., Peres, M. S, Boscarioli. C. (2016)
10. Sarda-Espinosa, A.: Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package. In: `https://CRAN.R-project.org/package=dtwclust`, (2017)
11. Liao, T.: Clustering of time series data-a survey In: The Journal of The Pattern Recognition Society, (2004)

12. Nonlinear Mixture Models: A Bayesian Approach. Tatarinova, T., Schumitzky, A., (2014)
13. Keogh,E., Ratanamahatana, C.: Exact indexing of dynamic time warping. In: Knowledge and Information Systems, vol. 7, pp. 358–386 (2005)
14. Huete,A., Liu H. Q., Batchily, K., Leeuwen,V.:A comparison of vegetation indices over a global set of TM images for EOS-MODIS. In: Remote Sensing Environment. vol.59, pp.440–451 (1997)
15. Vegetation Indices 16-Day L3 Global 250m, http://pdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q
16. Bock, H. H. Information and entropy in cluster analysis.:Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: an informational approach. Springer Netherlands, (1994)
17. Boles, S.H., Xiao, X., Liu, J., Zhang, Q., Sharav, M., Chen, S. and Ojima, D.: Land cover characterization of temperate East Asia using multi-temporal VEGETATION sensor data. In: Remote Sensing of Environment. vol.90, pp. 477–489, (2004).
18. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., Perona, I.: An extensive comparative study of cluster validity indices. In: Pattern Recognition, 46(1), pp. 243–256 (2013)
19. Tomašev, N., Radovanović, M.: Clustering Evaluation in High-Dimensional Data. In: Unsupervised Learning Algorithms, Springer International Publishing (2016)