IMEKO International Conference on
Metrology for Archaeology and Cultural Heritage
Lecce, Italy, October 23-25, 2017

Machine Learning: a Toolkit for Speeding Up Archaeological Stratigraphic Identification

I. Cacciari[1], G. F. Pocobelli[2], S. Siano[1]

[1] *Istituto di fisica Applicata "Nello Carrara", Consiglio Nazionale delle Ricerche, v. M. del Piano 10, 50019 Sesto Fiorentino (FI), Italy, i.cacciari@ifac.cnr.it*
[2] *Cooperativa Archeologia, v. L. la Vista, 5, 50133 Firenze, Italy*

***Abstract* – Digging a site, recording the stratigraphic units and interpreting the results in order to comprehend the historical processes of the site formation are part of archaeological excavation work. As archaeologists dig, they consider the extension, color, texture, hardness, and composition of the soil that they are removing. These processes are time-consuming, and may be affected by human skill. The main idea of this work is to automatize stratigraphic unit detection and characterization. To this end, a Machine Learning algorithm has been applied to digital images of archaeologic excavation sites for classifying regions that are similar in color and the contours of which represent stratigraphic units. Each stratigraphic unit has been characterized in terms of texture according to the mean energy. This combined approach speeds up the documentation work: since the results are readily digitalized during an excavation, they could offer a prompt guide for archaeologists.**

I. INTRODUCTION

The long-term buildup of sequential layers of soil material due to geological and human activities is commonly referred to as stratigraphy. One of the most important parts of post-excavation is represented by the study of these layers in order to comprehend the historical processes of the site formation. In modern archaeology, the basic stratigraphic units are defined according to lithological criteria: for example color, texture and particle characteristics, rather than to the objects that they may contain. Since excavations come across different layers at various elevations under the surface, the task of the archaeologist is to distinguish the layers during an excavation. Since layers fade into each other and are rarely completely distinct, this task is challenging and difficult.

In general, after a layer has been completely excavated, the walls and floor are cleaned and prepared for documentation. The archaeologists take photos of both the sides and the bottom of the excavation, and sketch what they see in the trench. These drawings delineate the extent, shape of the features, artifacts, and layers in the horizontal plane. This stage is not only considerably time-consuming and may be affected by human skill, but it also complicates the digitalization of the results. In this context, any attempts at automatizing the identification of stratigraphic units during excavation work is considered challenging.

This work is mainly focused on exploring the possibility of an "automated archaeologist" [1] who is capable of recognizing the stratigraphic unit from digital images during the excavation. The automatization of this procedure is strongly motivated by the need for a prompt guide for "human archaeologists": it could simplify the drawing step and hence speed up the documentation of the excavations.

The Machine Learning (ML) approach has been considered suitable for classifying regions that can be considered similar (cluster identification). In general, the learning process of human beings is simulated through experience (training): ML algorithms enable machines to create generalized rules from empirical data and, based on rules that have been learned, to make estimates for future data. This tool has positively demonstrated its potential in Cybersecurity [2], Financial Trading [3], and Healthcare [4], as well as in geological mapping [5-7].

However, to the authors' knowledge, ML has never been used in archaeology. This strongly motivates our approach for speeding up stratigraphic unit identification by providing digital images of the excavation site. It is for this reason that unsupervised learning has been considered among the different types of ML tasks. Its main aim is to find patterns and relationships within data (e.g. pixels in a digital image), and in this case there are no training examples. One of the key approaches of unsupervised learning is clustering: similar data points are grouped together, and these groups differ meaningfully one from the other. Here, we have chosen the k means algorithm for clustering images of archaeological excavation sites into k regions. This algorithm attempts to enhance the color similarity and to keep the colors separate from one another as far as possible. This approach could help archaeologists in

identifying stratigraphic unit during an excavation campaign; moreover, it would reduce the time spent for digitalizing the results.

The work is organized as follows: Section II describes the images considered for color clustering, the k-means algorithm, and the texture analysis performed on the clusters. Section III describes the experiments performed on the images and the results using the k means and texture analysis. Lastly, Section IV presents the conclusions as well as the future perspective.

II. MATERIALS AND METHODS

A. Images for color clustering

In this work, a mockup that simulates an excavation site and two actual archaeological sites have been considered for color clustering. Texture analysis was also used to characterize these clusters in terms of Haralick features such as energy.

The mockup (Fig. 1) was prepared in such a way as to simulate excavation sites characterized by different colors and textures. A green plastic dustsheet was spread under a wooden box; the box was filled with soil (A) and peat (B1-B6). On the soil layer (A), a pebble circle was introduced so as to simulate the case of a uniform soil background in which the anthropic environment is characterized by areas with different colors and textures (pebbles). Six different areas were prepared on the peat background so as to simulate different combinations of texture and color.

Several images were photographed using a digital camera placed at different heights from the soil (1, 4, 7 m), with a 1280x960 resolution. Moreover, the effect of different illumination conditions was considered. The same photos were taken in the morning and in the afternoon: it was observed that the shadows were limited in the photos taken in the morning at a height of 4 m, and were therefore preferable for the current work.

Another set of images was taken during two excavation campaigns in Italy. On the first site, two stratigraphic units were well recognized: one of them was a portion of a wall, in the background there were bricks that probably came from the collapse of the roof, and the excavation limit was also visible. On the second site, the texture of the background appeared smoother than on the first site, and the stratigraphic unit could be clearly distinguished from the different colors. The images of the first and second sites were 1100x1120 px and 1368x912 px, respectively.

B. Archaeological texture

The surface of an archeological site is generally not uniform. It may contain minute variations in color, texture, composition and hardness, some of them are of tactile and some of visual nature. Archaeological texture is considered as surface attributes having visual or tactile variety, that may characterize its appearance.

In image processing, texture describes the amplitude patterns and quantifies the spatial arrangement of color or intensities in an image or in a selected portion of it.

One of the most effective tools for quantifying the perceived texture of an image is based on the gray-level co-occurrence matrix (GLCM, [8]).

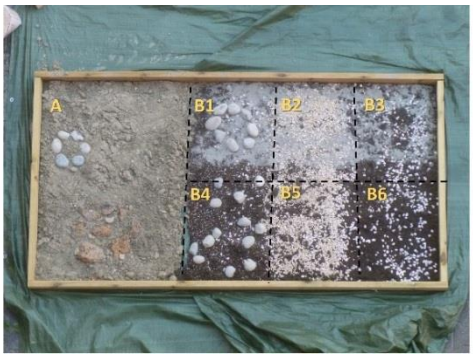

Fig. 1. Mockup prepared to simulate different excavation sites. A) six pebbles have been laid down on an almost uniform soil layer. A peat layer has been partially covered by: B1) sand and a circle of 8 pebbles, B2) sand and a large quantity of gravel, B3) sand and a small quantity of gravel, B4) randomly-placed pebbles, B5) a large quantity of gravel, B6) a small quantity of gravel.

The elements of the co-occurrence matrix measure the number of times different combinations of pixels pairs of a specific gray level occur in an image for various directions (θ) and different distances (d) [9]. Given an M×N neighborhood of an input image with G gray levels, let $f(m,n)$ be the intensity at pixel (m,n) of the neighborhood; then, the element (i,j) of the GLCM is defined as follows:

$$P(i,j|\Delta x,\Delta y) = \frac{1}{(M-\Delta x)(N-\Delta y)} \sum_{n=1}^{N-\Delta y} \sum_{m=1}^{M-\Delta x} A \quad (1)$$

where

$$A = \begin{cases} 1 & \text{if } f(m,n)=i \text{ and } f(m+\Delta x, n+\Delta y)=j \\ 0 & \text{elsewhere} \end{cases} \quad (2)$$

and

$$d = \sqrt{\Delta x^2 + \Delta y^2}, \theta = arctg\frac{\Delta y}{\Delta x}. \quad (3)$$

In other words, the matrix element P contains the second-order statistical probability values for changes between gray levels i and j at a particular displacement distance d and at a particular angle θ.

In this work, 256 gray levels images and 1 as displacement were considered. To avoid dependency of direction, a normalized symmetrical matrix was computed by summing up the four matrices θ=0°,45°,90°,135° and normalized by dividing each entry

by the total number of pixel pairs. Hence, the normalized co-occurrence values lie between 0 and 1, and this enables them to be thought of as probabilities.

A number of textural features, can be extracted from the co-occurrence matrices. In this work, we focused on energy defined as follows:

$$Energy = \sum_i \sum_j P_{i,j.}^2 \quad (4)$$

It reflects the grayscale distribution homogeneity of images and measures the textural uniformity of an image. In other words, it supplies information on the randomness of the spatial distribution. Energy assumes its highest value when gray level distribution has either a periodic or a constant pattern. In a homogeneous image, very few dominant gray-tone transitions are expected. The corresponding co-occurrence matrix has fewer entries of larger magnitude, thus resulting in a large value for the energy feature. Since this feature is generally useful for highlighting geometry and continuity, it has been considered for this work [10, 11].

The size of the neighborhood partly determines the success of a texture-based image analysis. If the window size is too large, it could overlap different features and introduce spatial errors [12]; on the contrary, if it is too small, not enough spatial information can be extracted to distinguish between different features. In this work, the neighborhood area has been chosen with the same aspect ratio of the original image. For each neighborhood size selected for this work, ten calculations chosen randomly from different areas were performed and their mean values was considered. The associated errors were estimated as maximum deviations.

C. *K means color clustering algorithm*

The design of algorithms that enable computers to develop types of behavior based on empirical data (such as from databases or sensors) is commonly referred to as ML. It represents a powerful tool for a variety of problems, from pattern recognition to the visualization of high-dimensional and cluster identification. The aims of ML research are to learn automatically to identify valid and potentially useful patterns and to make intelligent decisions based on data.

In this work, we have considered unsupervised ML to model the hidden structure or distribution in the unlabeled data in order to learn more about the data.

One of the main approaches of unsupervised learning is clustering. It assigns a set of inputs into subsets called clusters, so that each subset ideally shares some common characteristic and is able to place any new input within the appropriate cluster. Clustering is therefore suitable for the identification of different patterns in data. In image processing, it can be used to divide a digital image into different regions for border detection or object recognition. In particular, it has been applied for different purposes in medicine [13-17], biology [18, 19], document clustering [20, 21], agriculture [22, 23], geophysics [24, 25], remote sensing [26, 27], security and crime detection [28], marketing and consumer analysis [29, 30], and also automatic image annotation [31]. K means is one of the simplest unsupervised algorithms that can be used to solve a clustering problem in digital images.

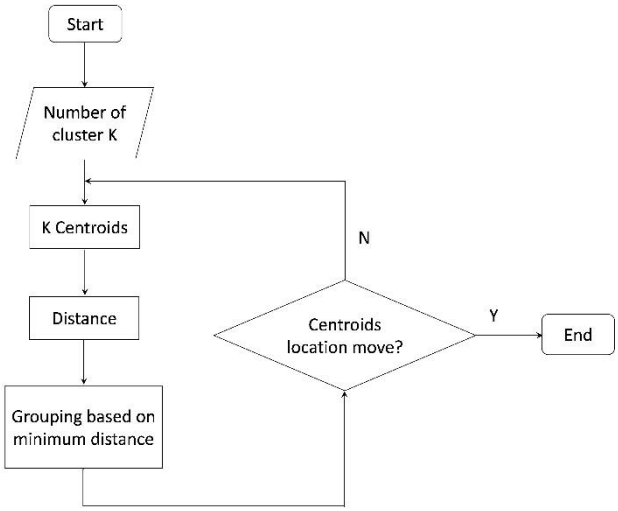

Fig. 2. Block diagram of a k means algorithm for color clustering.

The procedures used in this work follows a simple and fast way to cluster a given dataset through a certain number of clusters established a priori. A block diagram of the algorithm has been outlined in Fig. 2. For each cluster, the main idea is to define a centroid (barycenter) in an ingenuous way. As different locations of the centroids may cause different results, the best choice is to place them far away from each other as much as possible. The next step is to consider each point in the dataset and to associate it with the nearest centroid. This step is concluded and an early groupage is completed when no point is pending. New k centroids are then recalculated as the barycenters of the k clusters obtained in the previous step. The centroids are calculated according to the Euclidean distance between the color dimensions and the centroids. Once new k centroids have been calculated, a new linking has to be established between the same dataset and the nearest new centroid (minimum distance). The iteration stops when k centroids location no more change.

As described above, clustering is the algorithm that enables dataset points to be grouped with some similarity along a dimension, while the points that differ from each other are kept further apart. In the case of digital images, the dimension used is generally color, because the human vision system chooses color, rather than shapes and texture, as the main discriminant feature.

In this work, the RGB images of excavation sites have

been clustered using k means in such a way that the different regions of the image are marked by k colors and the boundaries are revealed by separating the different regions. The outputs of the algorithm are k images in which all the non-zero pixels represent the object in the cluster. By assigning an 8-bit number to each pixel in a cluster, a composite image (in a false color) is then produced: this helps the drawing of the stratigraphic unit contour using standard edge-detection techniques.

III. RESULTS AND DISCUSSIONS

One of the main issues of the k means algorithm is the number of clusters: an underestimation of the k number may result in a poor color clustering and hence a poor stratigraphic unit identification. Some preliminary tests were performed in order to establish the best number of colors.

In Fig. 3 the color clustering obtained on a portion of area A with different k values is shown. The images have been presented in matrix fashion: the first row represents the initial image (1000x1000 pixel); from the second to the fifth rows, the color clustering obtained with k=2, 3, 4, and 5, respectively. For each row, a gray image that represents the composite image has also been inserted into the last column. It is interesting to observe the two colored images obtained for k=2: the wooden box and soil were considered as belonging to the same color cluster, and both the plastic dustsheet and the pebbles circles, to the other. In the corresponding composite gray image, the contour of the wooden box is hardly visible; hence, the underestimation of the colors number is reflected in a mistaken identification of the contour. This also holds true for k=3 and k=4. A suitable number of colors for the best clustering is achieved with k=5. In this case, the anthropic environment (pebbles) is clearly separated from the background (dustsheet and soil); the contour of pebbles circle and of the wooden box can thus be accurately drawn.

Once the color clusters have been obtained, it is interesting to characterize them in terms of mean texture. In particular, the energy (Haralick feature) has been calculated with different neighborhood sizes in the dustsheet and soil regions. The results are summarized in Fig. 4. The curves with lower neighborhood sizes show a downward trend, and with higher neighborhood sizes, the mean energy values tend to be constant. These values could be considered as the mean texture of those colors clusters and can be used to characterize the different regions and, hence, the stratigraphic unit.

It is interesting to note that the mean energy of the dustsheet is always higher than that of the soil. This corresponds to a higher degree of uniformity in the image and, therefore, to a smoother texture.

Analogous tests were performed on the other areas of the mockup: for B areas, the best color clustering was achieved with k=4. In Fig. 5 the composite images obtained for six portions of the B areas with k=4 have been reported. The analyzed areas are indicated as in Fig. 1.

In the B1 case the color clustering has clearly separated the anthropic environment from the background. This represents a demanding case for color clustering, because there is no strong color difference between sand, gravel background, and the pebbles circle as there is in the A and B4 areas.

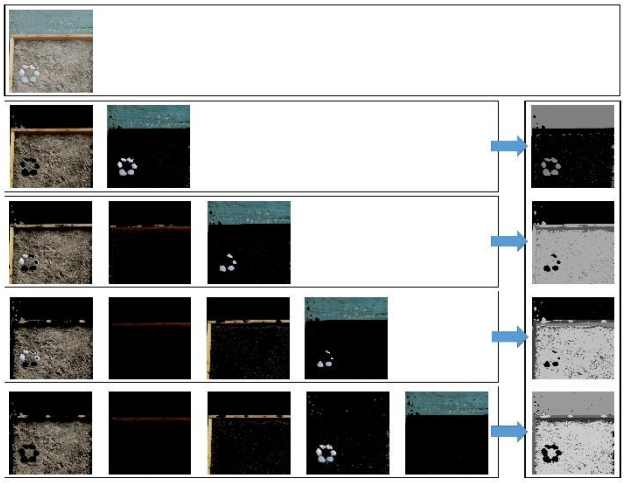

Fig. 3. *Color clustering of a portion of area A. The first row represents the original image; the color clusters for k=2, 3, 4, 5 are presented in rows 2, 3, 4 and 5, respectively. Each row ends with a gray image (composite) that is calculated by assigning an 8-bit number to each color cluster.*

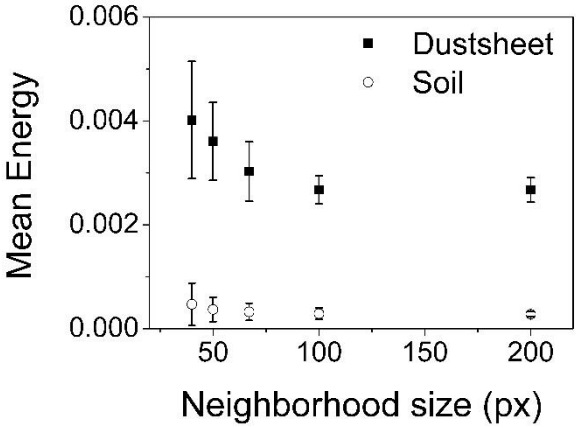

Fig. 4. Mean energy calculated with different neighborhood sizes and the same aspect ratio as in the original image (1000x1000).

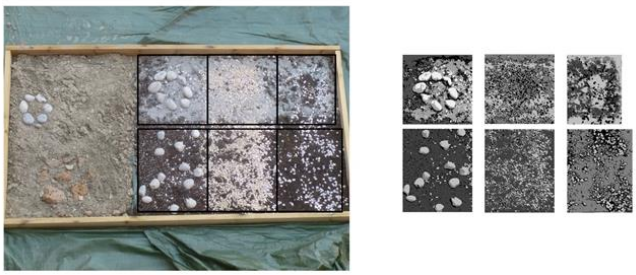

Fig. 5. Color clustering with k=4 of the six portions of area B (left) and composites images (right).

Fig. 6. Excavation sites: on the left is the original image of site 1 with overlaid contours that were obtained with k=5; on the right are the results obtained for the second site with k=4 color clustering.

This suggests that the algorithm could be useful also in the case in which the main differences are in texture rather than in color. In the case of B2, the background is fairly uniform (gravel partially covered by a sand layer): it shows almost the same texture, but with different colors.

The clustering has pointed out the two areas quite well, even if it is not possible to draw a contour. The same holds true for the cases B3, B5 and B6, in which different combination of peat, gravel and sand were considered for the background.

Moreover, color clustering was also performed on images of two excavation sites. Fig. 6 shows the original images with, overlaid, the regional contours obtained by means of edge-detection.

The contours obtained highlight two stratigraphic units for site 1: a portion of a wall (4) and a region with different texture (5) placed between the wall and the excavation limit (3). The stone (1) is also identified. This is interesting, since the algorithm may spread its potentiality to findings that are not actually stratigraphic units, but are naturally of interest to archaeologists. The contour of site 2 highlights a stratigraphic unit (1) that seems to have a coarser texture than that of the background (2).

The color clusters are characterized in terms of texture by calculating the corresponding mean energy. As each neighborhood area should be included in the color clusters, it is not possible to perform a texture analysis with larger neighborhood sizes. For this reason, we considered for the image of site 1 (1100x1120 pixel) an area of 44x45 pixel and for the image of site 2 (1368x912), an area of 68x46 pixel. These two sites are extremely interesting because different combinations of texture and colors can be observed.

The results summarized in Fig. 7 suggest that each cluster could be characterized by different texture values (mean energy). In site 1, the two stratigraphic units (labelled with cluster 4 and cluster 5) have approximately the same texture even if they are chromatically different. Clusters 3 and 5 seem similar in terms of color, but show different textures.

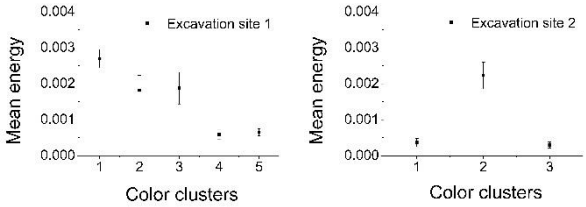

Fig. 7. Mean energy of color clusters of site 1 (left) and site 2 (right).

In site 2, clusters 1 and 3 have similar colors and textures. This suggests that, when combined with color clustering, texture analysis can provide another level of information for interpreting and characterizing stratigraphic units during an archaeological excavation.

IV. CONCLUSIONS

In this work, images of archaeological excavation sites have been analyzed with the use of an unsupervised ML algorithm.

The results obtained with the k means algorithm and edge detection represent the first demonstration that stratigraphic units can be readily and properly identified. Moreover, the textural uniformity calculated in these regions by means of energy proves to be useful for characterizing stratigraphic unit.

The combination of ML and texture analysis can become a good practice for speeding up the documentation work of archaeologists and could open the way towards the creation of an "automated archaeologist".

ACKNOWLEDGEMENT

The present work has been carried out within the framework of the Archeo 3.0 project funded by the Tuscan Region (POR FESR 2014-2020).

REFERENCES

[1] J. A. Barceló, "Towards a True Automatic Archaeology: Integrating Technique and Theory", Proc. of the 35th International Conference on

Computer Applications and Quantitative Methods in Archaeology, 2007, pp. 413-417.

[2] S. Dua, X. Du, “Data Mining and Machine Learning in Cybersecurity”, Auerbach Publications Boston, MA, USA, 2011.

[3] L. Gyorfi, G. Ottucsak, H. Walk, “Machine Learning For Financial Engineering”, Imperial College Press, London, UK, 2012.

[4] D. A. Clifton, “Machine Learning for Healthcare Technologies”, Institution of Engineering and Technology, London, UK, 2016.

[5] M. J. Cracknell, A.M. Reading, “Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information”, Comput Geosci, vol. 63, 2014, pp. 22-33.

[6] A. S. Harvey, G. Fotopoulos, “Geological Mapping using machine learning algorithms”, ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLI-B8, 2016, pp. 423-430.

[7] D. J. Lary, A. H. Alavi, A. H. Gandomi, A. L. Walker, “Machine learning in geosciences and remote sensing”, Geoscience Frontiers, vol. 7, 2016, pp. 3-10.

[8] R. M. Haralick, K. Shanmugam, I. Dinstein, “Textural features for image classification”, IEEE Trans Syst Man Cybern Sys, vol. 3, No. 6, 1973, pp. 610-621.

[9] M. Nixon, A. Aguado, “Feature extraction & Image Processing”,

[10]Press, Oxford, UK, 2008

[11] G. Preethi, V. Sornagopal, “MRI image classification using GLCM texture features”, International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), 2014, pp. 1-6.

[12] B. P. West, S. R. May, J. E. Eastwood, and C. Rossen, “Interactive seismic facies classification using textural attributes and neural networks” The Leading Edge, vol. 21, No. 10, 2002, pp. 1042-1049.

[13] K. D. Toennies, “Guide to Medical Image Analysis: Methods and Algorithms”, Advances in Computer Vision and Pattern Recognition, Springer-Verlag London Limited, 2012.

[14] W. Halberstadt, T. S. Douglas, “Fuzzy clustering to detect tuberculous meningitis-associated hyperdensity in CT images”, Comput Biol Med, vol. 38, No. 2, 2008, pp. 165-170.

[15] L. Liao, T. Lin, B. Li, “MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach”, Pattern Recognit Lett, vol. 29, No. 10, 2008, pp. 1580–1588.

[16] J. Hill, E. Corona, J. Ao, S. Mitra, B. Nutter, “Information Theoretic Clustering for Medical Image Segmentation”, in Advanced Computational Approaches to Biomedical Engineering. Springer, Berlin, Heidelberg, 2014.

[17] S. Fouad, D. Randell, A. G. Hisham, M. G. Landini, “Unsupervised Superpixel-Based Segmentation of Histopathological Images with Consensus Clustering”, Annual Conference on Medical Image Understanding and Analysis, MIUA 2017: Medical Image Understanding and Analysis, 2017, pp 767-779.

[18] V Govindaraj, A Vishnuvarthanan, Ah Thiagarajan, Kannan M, and P. R Murugan, “Short Notes on Unsupervised Learning Method with Clustering Approach for Tumor Identification and Tissue Segmentation in Magnetic Resonance Brain Images”, J Clin Exp Neuroimmunol, vol. 1, No. 1, 2016, pp. 1-10.

[19] E. R. Hruschka, R. J. G.B. Campello, L. N. de Castro, “Evolving clusters in gene-expression data”, Inf Sci, vol. 176, No. 13, 2006, pp. 1898-1927.

[20] Y. Xu, J. Wu, C.-C. Yin, Y. Mao, “Unsupervised Cryo-EM Data Clustering through Adaptively Constrained K-Means Algorithm”, PLOS, 2016.

[21] X. Cai, W. Li, “A spectral analysis approach to document summarization: clustering and ranking sentences simultaneously”, Inf Sci, vol. 181, No. 18, 2011, pp. 3816–3827.

[22] M. Carullo, E. Binaghi, I. Gallo, “An online document clustering technique for short web contents”, Pattern Recognit Lett, vol. 30 No. 10, 2009, pp. 870-876.

[23] P. Papajorgji, R. Chinchuluun, W. S. Lee, J. Bhorania, P. M. Pardalos, “Advances in Modeling Agricultural Systems”, New York, NY, USA, Springer US, 2009

[24] R. Chinchuluun, W. S. Lee, J. Bhorania, P. M. Pardalos, “Clustering and classification algorithms in food and agricultural applications: a survey”, Advances in Modeling Agricultural Systems, vol. 25, 2008, pp. 433-454.

[25] Y.-C. Song , H.-D. Meng, M. J. O' Grady, G. M. P. O'Hare, “The application of cluster analysis in geophysical data interpretation”, Computat Geosci, vol. 14, No. 2, 2010, pp. 263-271.

[26] V. S. Sidorova, “Unsupervised classification of image texture”, Pattern Recognition and Image Analysis, vol. 18, No. 4, pp 693-699, 2008.

[27] Haixia Bi, Jian Sun, Member, IEEE, and Zongben Xu, “Unsupervised PolSAR Image Classification”, IEEE Trans Geosci Remote Sens, vol. 55, No. 6, 2017, pp. 3531-3544.

[28] A. Hassanzadeh, A. Kaarna, T. Kauranne, “Unsupervised Multi-manifold Classification of Hyperspectral Remote Sensing Images with Contractive Autoencoder”, Scandinavian Conference on Image Analysis, SCIA 2017: Image Analysis, 2017, pp. 169-180.

[29] T. H. Grubesic, “On the application of fuzzy clustering for crime hot spot detection”, J Quant

Criminol, vol. 22, No. 1, 2006 pp. 77-105.
[30] Y.-J. Wang, H.-S. Lee, "A clustering method to identify representative financial ratios", Inf Sci, vol. 178, No. 4, 2008 pp. 1087-1097.
[31] J. Li, K. Wang, L. Xu, "Chameleon based on clustering feature tree and its application in customer segmentation", Annals of Operations Research, vol. 168, No. 1, 2009, pp. 225-245.
M Favorskaya Lakhmi, C. Jain, A. Proskurin, "Unsupervised Clustering of Natural Images in Automatic Image Annotation Systems", in New Approaches in Intelligent Image Analysis pp 123-155, Eds. Roumen Kountchev, Kazumi Nakamatsu, 2016.