An Unsupervised Machine Learning Approach in Remote Sensing Data

Mauro Mazzei(✉)

National Research Council, Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", via di Taurini, 19, 00185 Rome, Italy
mauro.mazzei@iasi.cnr.it

Abstract. The analysis, image recognition, the classification in remote sensing data always up to date is for the insiders a very useful system for the management, planning, protection of our territory. Technological evolution the succession of applications and tools in GIS environment, have refined more and more techniques allowing us to computerize much of the territory bringing an enrichment of geographical data distributed more or less updated, remote sensing, digital orthoimage help us in the areas where the local computerization is not always updated. In this paper I want to demonstrate that an approach based on machine learning applied to territorial analysis using dynamic matrices, provides for the definition of numerical indices assigned to each recognized pattern. The numerical indices will be the elements of recognition where the territorial computerization is insufficient. The data from remote sensing or digital orthoimage are variable matrices where the numerical indices have to express their potential. The goal is to educate the recognition in matrix form to obtain a dynamic topology where the computerization has not been updated.

Keywords: Data mining · Spatial data mining · Spatial data analysis · Spatial statistical model

1 Introduction

The method of territorial analysis using dynamic matrices involves the definition and processing of the numerical indices assigned to the information layer (remote sensing images) examined. The numerical indices will be the elements of recognition where the territorial computerization is insufficient. The data from remote sensing or from digital orthoimage are the variable matrices where the numerical indices must express their potential.

The aim is to create a self-learning system to obtain a dynamic topology able to learn the evolutionary changes of territorial contexts and/or landscapes. Some of the most common numerical, spatial and connectivity indexes for territorial analysis can be summarized in these types.

- An index or numerical indicator is a mathematical expression (sometimes an algorithm) capable of transforming a set of data into a synthetic numerical value (in a given scale) attributable to an emergent property of a landscape.

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11621, pp. 435–447, 2019.
https://doi.org/10.1007/978-3-030-24302-9_31

- The distance indices express in terms of space or time the distance between two or more objects within a n-dimensional space.
- The connectivity indices express the degree of geographical relationship between the different objects placed in a geographical or process dimension. In particular the graphs are able to express in numerical form the relationships between the different objects and at the same time to characterize the spatial patterns.
- Spatial indexes are able to describe the shape of patches, evaluate their individual and collective complexity expressed in an environmental mosaic. Border irregularities, area width, concentration and interspersion are fundamental parameters for the characterization of a landscape.
- The spatial arrangement of the patches is measured through irregularity indices, while the tissue structure of the mosaic requires indices based on adjacency ratios.

The methods of landscape analysis can be summarized in some types:

- Numerical, spatial, multiscale analysis and modeling are the main approaches used in landscape analysis.
- The landscape analysis can be conducted on an organism, patch, mosaic or entire landscape scale.
- Shape, size, irregularity, spatial arrangement, morphological diversity and typological diversity are some of the characteristics of the objects identified in a landscape that are analyzed with the use of numerical indices and spatial indices.
- Spatial statistics, fractal mathematics, neutral models are some of the approaches in the study of the complexity of landscapes.

In this work a method of landscape analysis is presented with the aim of instructing the system to automatically recognize the emerging characteristics of the territory or landscape, through the self-learning of the elements that compose it. The elements used for the analysis are the data obtained from the satellite images that make up the variety of elements to be analyzed.

1.1 Numerical Indices

By numerical indexes i mean mathematical expressions able to capture the information emitted by the emergent properties of a system, or in this case, of a landscape. These indices actually measure the variety of objects, their abundance and their spatio-temporal dynamics. Where Ai is the abundance of the object I [22];

$$A = \sum A_i \tag{1.1}$$

The Geographical Distance Index represents a fundamental element of the landscape. The distance in a Euclidean space is given by the application of the Pythagorean theorem, given two points $X'Y'$ and $X''Y''$ their distance is given by the following:

$$d = \sqrt{(X' - X'')^2 + (Y' - Y'')^2} \tag{1.2}$$

the distance between two points is generally calculated as the shortest spatial interval, but while this is true for isotropic surfaces, in the presence of anisotropic surfaces such as environmental mosaics the minimum distance between two points is generally not a straight line.

When we are in the presence of a group of points in a Euclidean space, for example a group of trees or other objects that are scattered, the distance between the points is not the measure of each but the standard distance that represents the dispersion of the whole. This distance can be calculated with a quadratic average of the distance from the center of gravity;

$$d_l = \sqrt{\left(\sum^{n/i=1}\left(d_{ic}^{2/n}\right)\right)} \tag{1.3}$$

d_{ic} is the distance between each observation i is the middle center of all points c, n the number of points.

1.2 Spatial Indices

By spatial indexes i mean those indexes that describe the spatial characteristics of the objects i encounter in an environmental mosaic. These characteristics are both topological (size, shape) and chorological (position with respect to other objects of the same type or different). Spatial indexes are able to describe the shape of the patches, evaluate their individual complexity and that expressed collectively in an environmental mosaic. Irregularity of the edges, width of the area, interdispersion and contagion are fundamental parameters for the characterization of a landscape.

The shape indices are based on the difference between a geometric figure giving (circles, square) and the patch inscribed within this figure, assuming the maximum regularity for geometric figures and whose ratios between appropriately treated area and the respective perimeter is about 1.

$$Y^1 = (2\sqrt{\pi A})\Big/P \tag{1.4}$$

Where A is the area and P the perimeter of the patch with g $1 \simeq 1$ for circular patches and g1 < 1 for shapes other than circle or based on polygons. Or simplified area perimeter report for forms represented in raster format we have:

$$Y^3 = A/p^3 \tag{1.5}$$

The perimeter is calculated in raster data through the Sobel operator which is an algorithm used to process digital images in particular to perform contour recognition (edge detection).

The operator applies two 3×3 kernels, i.e. two convolution matrices to the original image to calculate approximate values of the derivatives - one in the horizontal direction, and one in the vertical direction. If we call A the source image, and Gx and

Gy the two images whose points represent respectively the approximate values of the horizontal and vertical derivatives, the operation is described by:

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A \text{ e } G_y \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \tag{1.6}$$

1.3 Connectivity Indices

The connectivity indices express the degree of "geographical relationship" between the different objects placed in a geographical or process dimension. In particular, the graphs are able to express in numerical form the chorological relations between the different objects and at the same time to characterize their spatial patterns.

The connectivity indices express the maximum of the distances between the node i from each of the other nodes j. Each graph can be converted by a matrix where $d_{ij} = 1$ if it exists;

$$K_i = \max d_{ij} \tag{1.7}$$

Accessibility index is given by the following:

$$A_1 = \sum^{n/i=1} d_{ij} \tag{1.8}$$

Where d_{ij} is the number of nodes encountered to arrive at the chosen node. The value of the accessibility index is inversely proportional to the accessibility of the network node.

2 Data Organization

The starting data are remote sensing images obtained from sensors with high spatial and radiometric resolutions (QuickBird). The radiometric resolution therefore expresses the number of gray levels (dynamic range) used to represent the data collected by the sensor. In general, the detection systems are designed to record the greatest possible number of signal levels to minimize the loss of information relative to the observed scene, but the upper limit of the instrument's dynamic range is defined by the signal-to-noise ratio. The acquired QuickBird image was provided in 16-bit mode, so the intensity of the energy arriving at the detector is divided into discrete brightness ranges that are converted to 2^{16} (= 65536) gray levels. The acquired QuickBird satellite image is provided in two modes: multispectral and panchromatic. The multispectral image is recorded in four bands, partly overlapping, in the blue, green, red and near infrared channels, and therefore allows to trace the spectral response of the elements present on the ground, while the panchromatic integrates the signal received in a single wide band

of 140 nm. The spectral resolution in this case is therefore rather low, but this allows to optimize the spatial resolution of the sensor, which is able to detect the signal coming from a surface on the ground of only 70 cm per side [6, 7] (Fig. 1).



Fig. 1. Input remote sensing image.

The image in the preprocessing phase is corrected radiometrically and geometrically, after which we pass to the data insertion phase through a reading function of each individual pixel within the database. This phase allows data to be stored in an RDBMS for direct access without having to store a huge amount of data. The process used starting from the input data to get to the final output is shown (Fig. 2).

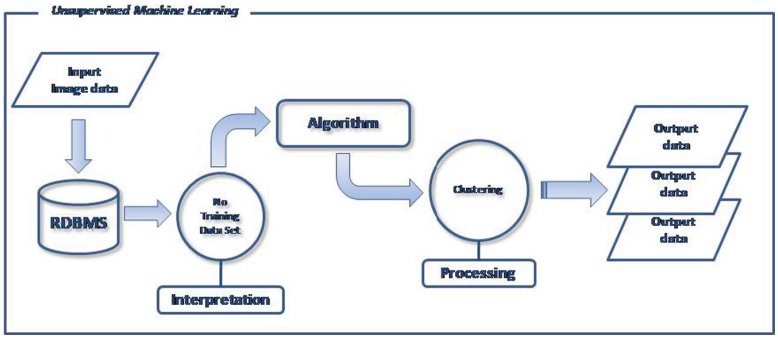

Fig. 2. The process unsupervised machine learning.

2.1 Data Analysis

Digital images are two-dimensional arrays, each element of which contains the pixel value to be displayed. To calculate the partial derivatives of this discrete representation, discrete operators will be required. This technique consists in scrolling on the original image a square matrix containing elements. For each position the scalar product is calculated between the covering matrix and the covered area and then all the elements

of the result are added. The value obtained will constitute the value to replace the pixel located in the middle of the examined area.

For each image, the mosaic thus obtained, we proceeded with an automatic extraction of "object" with the criterion of "similarity" between tiles assigned to satisfying the criteria of closeness and proximity of the relative levels of radiometric value. The automatic extraction was performed through algorithms and techniques of image processing, smoothing operators, laplace operators, threshold operators, edge extraction, form extraction.

"Object" are the structural unit of an environmental system heterogeneous, identified on the basis on the differences that appear within the area itself.

Each image was converted into a set of "objects" each of which were calculated fifteen numeric attributes related to the geometric properties of the object (perimeter, delta-X and delta-Y, area, etc.) and others to its inertial properties (Jx, Jy, Rx, Ry, etc.).

The image, when converted into two distinct sets of objects, each characterized by fifteen attributes, were subjected to a factor analysis procedure with the method of principal components (Hotelling) [8, 9, 11].

The Principal Components Analysis (PCA) is a linear transformation that transforms the data into a new coordinate system; the new set of variables, the principal components, are linear functions of the original variables and are uncorrelated. The greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. In practice, this is achieved by computing the covariance matrix for the full data set. Next, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue. One can see that the PCA's bias is not always appropriate; features with low variance might actually have high predictive relevance; this depends on the application [10].

Given a set of p observations for each variable of a complex of m variables, the principal component analysis is proposed to determine new variables linearly related with the given variables, but in a lesser number of these latter, so that we can represent the variability expressed by the original variables. If it is not possible to meet these conditions, it is not possible, to represent the variability of the original variables with less than m, t and he principal component analysis is limited to an acceptable extent represent the majority of this variability with less than m variables. The problem of the analysis of the components is therefore related to the reduction of the number of descriptive variables m of p objects, regardless of the ability to identify new variables; Such identification must be decided in each particular case, generally, without any reference to the statistics, and in the field of the phenomena involved in the study [12, 13, 17, 19, 20].

2.2 Statistical Analysis

The evaluation of the variance between the two aerial photos examined is based on the classification of objects extracted and subjected to statistical methodology of the analysis of the main components (Hotelling).

The principal component analysis is a multivariate statistical technique that explains the variability of a statistical variable in k dimensions Z = (Z1, Z2, …, Zk) in terms of k variables Y1, Y2,…Yk, linear combinations of the Z j. It has:

$$\mathrm{Yi} = \sum \mathrm{j}\ \mathrm{bij}\ \mathrm{Zj}\ (\mathrm{i} = 1, 2, \ldots, \mathrm{k}) \tag{2.1}$$

where bij are constants to be determined. Yi are called the main components of the variable Z and assuming they are not related to each other ordered by importance, in the explanation of the variability of Z we have:

$$\mathrm{cov}(\mathrm{Yi}, \mathrm{Yj}) = 0(\mathrm{i} \neq \mathrm{j}) \tag{2.2}$$

$$\mathrm{V}(\mathrm{Y1}) \geq \mathrm{V}(\mathrm{Y2}) \geq \ldots \geq \mathrm{V}(\mathrm{Yk}) \tag{2.3}$$

where cov is covariance and V is variance. Without loss of generality we can assume that the variables Zi are standardized, with mean equal to 0 and variance equal to 1, so as to eliminate the influence of the origin and the unit of measurement data, so that it results the following expression:

$$\mathrm{Z\,j} = (\mathrm{X\,j} - \mu\mathrm{j})/\sigma\mathrm{j} \tag{2.4}$$

Also, impose the condition that the overall variance of Zj is equal to that of Yi, i.e.:

$$\Sigma\mathrm{i}\ \mathrm{V}(\mathrm{Yi}) = \Sigma\mathrm{i}\ \mathrm{V}(\mathrm{Zi}) = \mathrm{k} \tag{2.5}$$

At last, suppose that the vectors

$$\mathrm{b\ i} = (\mathrm{bi}, 1,\ \mathrm{bi}, 2, \ldots, \mathrm{bi}, \mathrm{k}) \tag{2.6}$$

have unit length, i.e., they fulfill the condition:

$$\Sigma\mathrm{jb2\ ij} = 1(\mathrm{i} = 1, 2, \ldots, \mathrm{k}) \tag{2.7}$$

On account of this, the vectors bi that maximize the variance of Y1, of Y2, …, to Yk with the constraints (2.3) and (2.4), are the eigenvectors of the matrix C of the coefficients of correlation between the variables Zj, which correspond to the eigenvalues λ1, λ2,…, λk of C, sorted by non-increasing value. We then have:

$$|\mathrm{C} - \lambda\mathrm{I}| = 0 \tag{2.8}$$

$$\mathrm{b\ i}(\mathrm{C} - \lambda\mathrm{i\ I}) = 0 \tag{2.9}$$

where I is the unit matrix. The matrix C is symmetric and positive definite for which the solutions λi of the (2.8) are non-negative and such that their sum (trace of the matrix C) is equal to k. We then have:

$$\Sigma i \lambda i = k(i = 1, 2, \ldots, k) \tag{2.10}$$

The variance of the i-th component is:

$$V(Yi\ ) = \lambda i \tag{2.11}$$

And the contribution of Yi to the overall variance is:

$$Pi = V(Yi\ )/k = \lambda i/k \tag{2.12}$$

2.3 Spatial Statistical Models

This procedure allowed us to calculate the correlation coefficients between the 15 variables adopted to describe each object (See Table 1), with the aim of drastically reducing the number of variables, thereby explaining the overall variability of the system with a smaller number of attributes, each of which appears to be a linear combination of the attributes of departure [1–5].

Table 1. Variable considered

Variable	F1	F2	F3	F4	F5	F6
Nzm - F6	0.0424	−0.0031	0.2495	−0.1517	0.0175	0.9553
Nt - F2	0.0951	0.9901	−0.0063	−0.0058	0.0255	0.0039
Area - F2	0.0951	0.9901	−0.0063	−0.0058	0.0255	0.0039
P - F2	0.071	0.9709	0.0264	−0.0043	−0.0024	0.0064
DeltaX - F1	0.9736	0.0744	0.0292	0.0418	0.0453	0.016
DeltaY - F1	0.9083	0.0265	−0.02	−0.1049	0.0885	0.0208
ArIdl - F1	0.9388	0.1517	−0.0109	−0.0125	−0.0368	0.0116
Gx - F4	−0.0147	0.0068	0.076	−0.9825	0.115	−0.1055
Gy - F3	0.0294	0.005	0.9862	0.0612	−0.0762	−0.127
Jx - F2	0.0428	0.9929	−0.0022	0.0039	−0.0139	−0.0071
Jy - F2	0.0499	0.9926	−0.0027	0.003	−0.0139	−0.0071
Rx - F1	0.9627	0.0388	0.0435	0.0674	0.1535	0.0096
Ry - F1	0.9765	0.0216	0.0139	−0.005	0.1021	0.0108
ArRe - F1	0.9443	0.0876	0.0101	0.0459	−0.0366	0.0068
R/AAR - F5	0.1646	0.0115	0.0585	0.0961	0.9777	0.016

The Table 2 show the percentage of total explained variance that is derived from 6 components. The first extract component has eigenvalue 6.02, explained the 40% of the total variance. The second extract component has eigenvalue 4.43 explained a improve

of 30% of the cumulative variance, they explained a total variance of the 70%. The sixth component explained the 96% of the cumulative total variance [19].

Table 2. Percentage of total explained variance

Components	Eigenvalues	% Variance	% Cumulatative
1	6.02	40	40
2	4.43	30	70
3	1.14	7	77
4	1.01	7	84
5	0.95	6	90
6	0.87	6	96

I have reduced the size of the area of the object definition, from 15 to 6, after which i have reconstructed the images with the new standardized object values greater than average and minor than average, that is greater than zero and lower than zero - based on each major component. These are illustrated in the Fig. 3 and in the Fig. 4 [14–16].

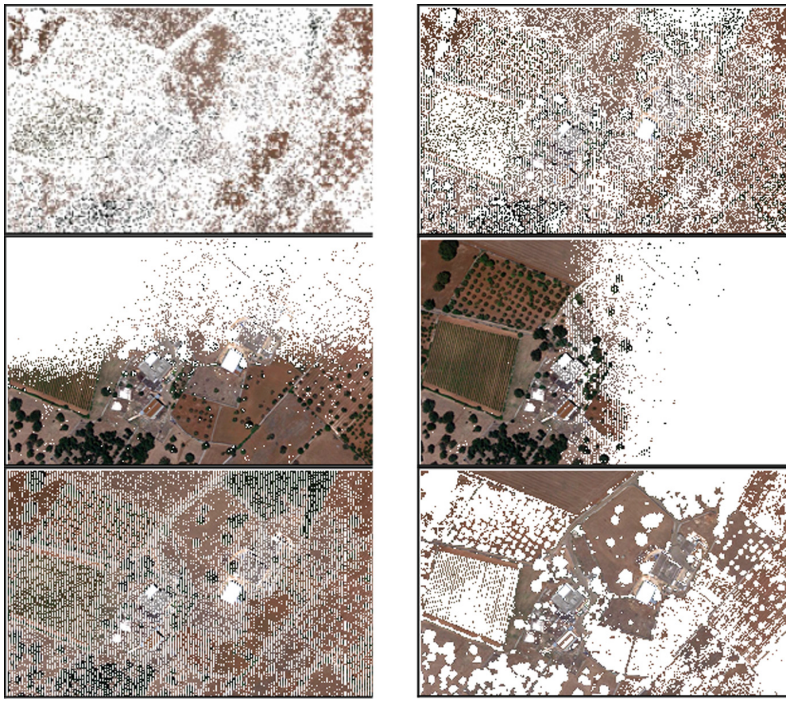

Fig. 3. Results of components from 1 to 6 greater than zero.

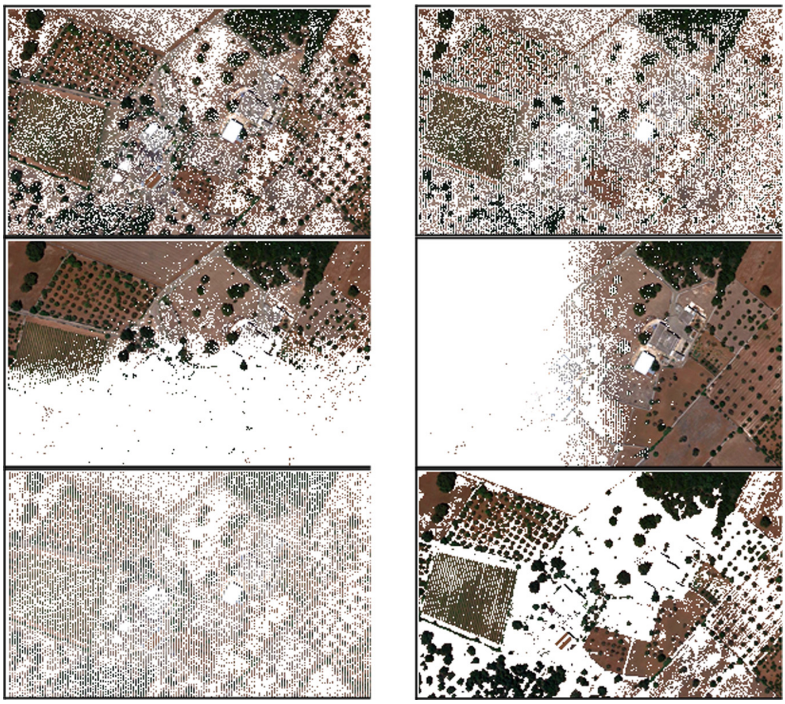

Fig. 4. Results of components from 1 to 6 lower than zero.

3 Conclusions

Factor analysis summarized very much the starting information without reconstructing a significant pattern normally used in supervised methods. The Fig. 5 show the correlations between factors and their spatial distribution in a multi-dimensional space [18, 21].

The distribution of the variables in a multi-dimensional space shows that in the sixth corresponding component in the graph Fig. 5, by identified with *S6* simbology, have a homogeneous distribution in all the variables analyzed with a preponderance in the variable that most characterizes this type of analysis, that is the depth of the pixel expressed by the radiance of the elements belonging to the territory and/or landscape analyzed.

I consider only the main components of the mosaic, belonging to the sixth component that has explained the 96% of the cumulative total variance (See Fig. 6). They highlight all the elements belonging to the analyzed mosaic. The results obtained are the exact perception that the human eye has when observing a landscape from above. The perception of the human eye subdivides the image in the background and foreground, highlighting the emerging texture. This is the same result that I obtained by training the system through machine learning techniques. In the two images below, the

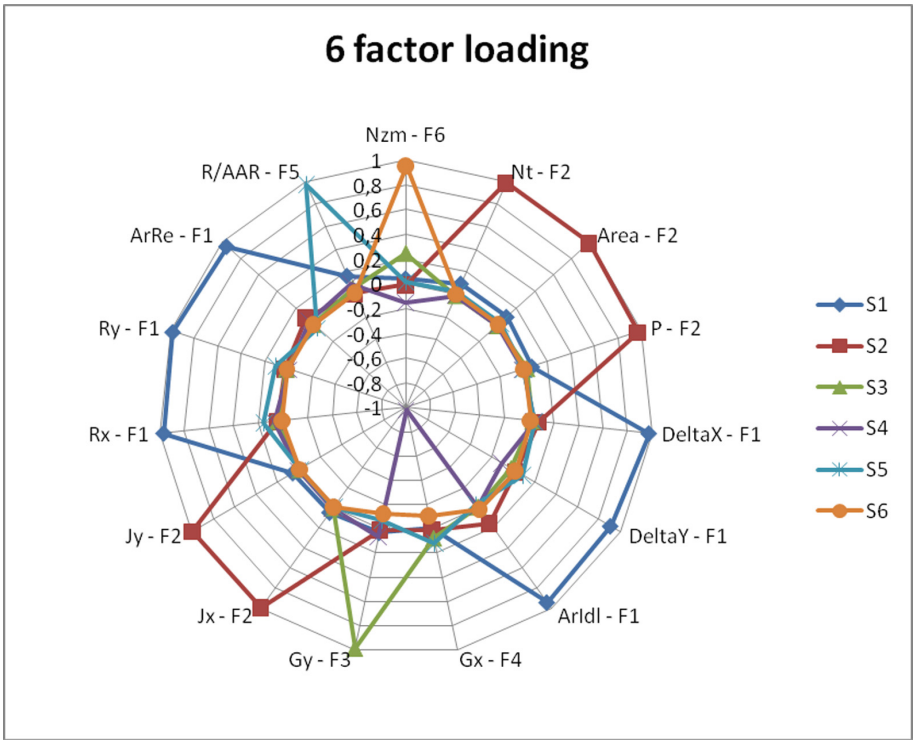

Fig. 5. Results of the distribution of the 6 factor loading.

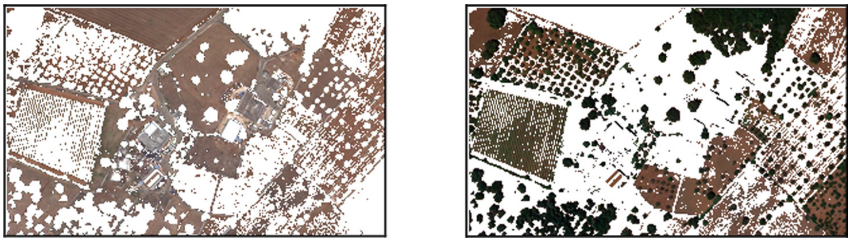

Fig. 6. Results only of component 6 greater and lower than zero.

background of the texture is highlighted in the image on the left with the links to the housing structures, while in the image on the right all the emerging of trees vegetation elements are represented with their respective connective systems.

In the future works it is foreseen to instruct the system to the perception of increasingly complex images such as urban and suburban territories where many discriminating elements insist.

References

1. Niladri, S.M., Susmita, G., Ashish, G.: Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images Appl. Soft Comput. **12**(8), 2683–2692 (2012)
2. Castellana, L., D'Addabbo, A., Pasquariello, G.: A composed supervised/unsupervised approach to improve change detection from remote sensing. Pattern Recogn. Lett. **28**(4), 405–413 (2007)
3. Deng, J.S., Wang, K., Deng, Y.H., Qi, G.J.: PCA based land use change detection and analysis using multitemporal and multisensor satellite data. Int. J. remote Sens. **29**(16), 4823–4838 (2008). https://doi.org/10.1080/01431160801950162
4. Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B.: Digital change detection methods in ecosystem monitoring: a review. Int. J. Remote Sens. **25**, 1565–1596 (2004)
5. Byeungwoo, J., David, A.: Partially supervised classification using weighted unsupervised clustering. IEEE Trans. Geosci. Remote Sens. (TGRS) **37**(2), 1073–1079 (1999)
6. Browna, L.G.: Survey of image registration techniques. ACM Comput. Surv. **24**(4) (1992)
7. Schwarz, K.P., Chapman, M.A., Cannon, M.E., Gong, P.: An integrated INS/GPS approach to the georeferencing of remotely sensed data. Photogram. Eng. Remote Sens. **59**(11), 1167–1674 (1993). https://www.asprs.org/wp-content/uploads/pers/1993journal/nov/1993_nov_1667-1674.pdf
8. Hotelling, H.: The generalization of student's ratio. Ann. Math. Statist. **2**, 30–378 (1931)
9. Zevi, M.: The matrix calculation in the method of the components principali. Faculty of Architecture, Rome (1977)
10. Ricci, F.: Statistics and Statistical Processing of the Information. Zanichelli, Bologna (1975)
11. Saddocchi, S.: Manuale di analisi statistica multivariata F. Angeli, Milano (1993)
12. Mezzetti, A.: Air pollution and vegetation, Edagricole (1987)
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for mode l fitting with applications to image analysis and automated cartography. Comm. ACM **24**, 381–395 (1981)
14. Cramer, M., Stallmann, D., Halla, N.: High precision georeferencing using GPS/INS and image matching. In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Alberta, Canada, 3–6 June, pp. 453–462 (1997)
15. Zuendorf, G., Kerrouche, N., Herholz, K., Baron, J.C.: Efficient principal component analysis for multivariate 3D voxel - based mapping of brain functional imaging data sets as applied to FDG - PET and normal aging. Hum. Brain Mapp. **18**, 13–21 (2003)
16. Mazzei, M., Palma, A.L.: Spatial statistical models for the evaluation of the landscape. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7974, pp. 419–432. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39649-6_30
17. Mazzei, M., Palma, A.L.: Comparative analysis of models of location and spatial interaction. In: Murgante, B., et al. (eds.) ICCSA 2014. LNCS, vol. 8582, pp. 253–267. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09147-1_19
18. Mazzei, M., Palma, A.L.: Spatial multicriteria analysis approach for evaluation of mobility demand in urban areas. In: Gervasi, O., et al. (eds.) ICCSA 2017. LNCS, vol. 10407, pp. 451–468. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62401-3_33
19. Mazzei, M.: Software development for unsupervised approach to identification of a multi temporal spatial analysis model. In: Muller J. (ed.) Proceedings of the 2018 International Conference on Image Processing, Computer Vision, & Pattern Recognition 2018. Computer Science, Computer Engineering & Applied Computing

20. Amato, F., Tonini, M., Murgante, B., Kanevski, M.F.: Fuzzy definition of rural urban interface: an application based on land use change scenarios in Portugal. Environ. Model Softw. **104**, 171–187 (2018)
21. Ranjbar nia, B., Murgante, B., Molaei Qelichi, M., Rustaei, S.: A comparative study employing CIA methods in knowledge-based urban development with emphasis on affordable housing in Iranian cities (case: Tabriz). In: Gervasi, O., et al. (eds.) ICCSA 2017. LNCS, vol. 10407, pp. 485–501. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62401-3_35
22. Farina, A.: Ecologia del paesaggio. UTET, Torino (2004)