Evaluating Principal Components Analysis of Particular Spatial Statistical Models

Mauro Mazzei, Armando Luigi Palma
National Research Council
Institute of Systems Analysis and Computer Science
Rome, Italy
e-mail: mauro.mazzei@iasi.cnr.it, palma@arpal.it

***Abstract* — This work is based on an analysis of the main components derived from particular patterns of spatial statistical data. The reference models of spatial statistical analysis are extracted only from the data of bi-temporal aerial photographs. This methodological approach introduces a significant improvement in the evaluation of changes in the territorial scenery, providing a wider interpretation of the problems of the area studied and encouraging a more analytical reading of complex environmental phenomena. In order to improve reading and analysis of the territorial changes it is necessary to compare the same geographical space in two different moments that enclose a well-defined period of time.**

Keywords – GIS; cartography; spatial data analysis; spatial statistical model.

I. INTRODUCTION

The techniques of analysis of multi-temporal remote sensing images are currently based on the recognition of the diversity of spectral indices of the two images observed. This methodology is applied by using software products that classify the content of the images. This classification is based on the similarity of the local spectral radiance. Classes of pixels are classified by evaluating the evolution of the state, from beginning to end [1]. The classification is made on the basis of spectral responses of the surfaces, based on the concept of similarity between pixels. This clustering is applied to the pixels of each image. This methodology requires the definition of prototypes based on a comparison of a minimum set of pixels representative of the class. There are basically two different methods for automatic classification of digital images [2]:

- Supervised Classification: this is used for a quantitative analysis of remote sensing images. This methodology can be summarized in two phases. The first phase provides a definition of the legend specifying the set of cover. The second phase provides an identification of the spectral signature on the ground. These two phases are useful for providing information on the software used to carry out the classification of the area [3][4].
- Unsupervised Classification: this uses the concept that produces similar spectral responses. A specific knowledge about the extent, type, and class descriptions is not required with these systems. These properties are based on observations of clusters formed in the spatial features. The main peculiarity is that the classes are identified by cluster compact and it is easy to distinguish between them. These clusters do not require knowledge of the features or of their nature. The entire space is split into spectral classes according to criteria of proximity or similarity. Only after the process of identification of spectral classes is finished, will the analyst associate these properties in relation to the knowledge of the territory [5].

In this paper, we propose a method that belongs to the techniques of unsupervised classification. With each cluster, in addition to the spectral properties, proximity and similarity, there are also recognized: perimeter, area, their relationship, their moments of inertia Jx and Jy, etc. Each cluster extracted from the digital image is described by its radiometric, geometric, and inertial properties, which are stored in a database. Applying the principal component method [9] to clusters extracted from digital images and described by their properties, as mentioned above allows us to identify, for each of the new factorial axes, similar classes of cluster characterized predominantly by only variables that have a high correlation value (factor loadings) between variable and factor.

The clusters, in the new reference by each main component, retain the property of having an average equal to 0 and variance equal to 1. The graphic rendering, for each main component of the clusters of positive coordinates, allows the segmentation of the original image. This shows classes of clusters that often are not detectable in the original image; therefore, the comparison between the segments of multi-temporal images obtained in this manner is facilitated and more readily available.

The paper is structured as follows. In Section 2, we illustrate the location of the study area, the material used, the organization of the digital data collection and preparation of digital data. In Section 3, we describe the method for the analysis of the data. In Section 4, we describe the type of statistical analysis used. In Section 5, we propose the data model used for the analysis, we provide examples in order to

ISBN: 978-1-61208-326-1

show our proposal. Finally, in Section 6, the discussion and conclusion are given.

II. Data Organization

The area of study is located in the south of Italy, in the province of Taranto, and is located in the northern Gulf of Taranto within the first basin of the Little Sea within which flows the river Galeso.

The material used for the analysis of this data is in raster digital format, with reference to an aerial photo of the Military Geographical Institute of 1940. The analyses were made available in digital format thanks to the collaboration of the Laboratory of ancient topography and photogrammetry of the University of Salento. The Geoportal's Web Map Service (WMS) of the Campania Region, which has kindly allowed us to use the orthophotos of the same site obnserved in 2010, as well as a portion of a topographic map in scale 1:10.000, used for georeferencing of both aerial photos.

The raster digital format was acquired through a photogrammetric scanner; georeferencing [6][7], was then used. The system used is the Universal Transverse Mercator (UTM) - European Datum 1950 (ED50).

The georeferencing of the Regional Technical Map requires a value of Root Mean Square (RMS); the graphics below show the graphical representation error with respect to the scale representation (1:10000 scale is 0.2 mm, 2 meters in real scale) [6][7].

This technique allows us to identify homologous points in the aerial photo for the georeferencing of the same space, without resorting to tools of GPS positioning, in order to detect in situ. In this way we obtain a minimum margin of error for the measurements of the spatial analysis for the case study.

The recognition of homologous points between technical mapping and aerial photo requires an accuracy for the identification of the exact location and a good distribution of support points; in this scenario, five control points are identified, which gave good results for a correct georeferencing of the aerial photo.

The transform algorithm applied to the matrix is of the type polynomial affine [6].

The recognition of homologous points in the aerial photo of 1940 is more difficult, since the change in territory does not permit an easy identification of common elements with aerial photos of the Military Geographical Institute.

In this case, we first identify large objects such as the geomorphological aspects, then small elements, for example, buildings and any other element which allows a good distribution of the points of support for the application of the georeferencing method.

The picture of 1940 is not a calibrated aerial photo and therefore the picture is distorted and requires an appropriate transformation. A transformation involves a mapping of locations of points in one image to new locations in another. Image Transformations used to align two images may be global or local. A global transformation is given by a single equation which maps the entire image. Examples are the affine, projective, perspective, and polynomial transformations. Local transformations map the image differently depending on the spatial location and are thus much more difficult to express succinctly. In the image of 1940 a local transformation was used [6].

The Pixels of the aerial photos positioned in their correct geographic space have been a simply of the nearest/neighbor interpolation.

III. Data Analysis

The two orthophotos from Figure 1 and 2 show the Northern arc of the first Sinus of the Mar Piccolo, into which flows the river Galeso, dating back to 1940 and 2010, respectively.

An initial examination of comparative urbanization can be seen in this frame of 2010, which mainly affected the western arc of the first Sinus of the Mar Piccolo. Variations seem to be able to detect an industrial plant related to “Cantieri Navali”, located north of the Sinus, and then, along the course of the river that flows near Galeso, the same shipyards that are no longer used.

Our curiosity extended far beyond a first visual comparison of the two images from which you can detect, even with the naked eye, the profound changes that occurred over a period of seventy years, especially in the first western Sinus of the Mar Piccolo.

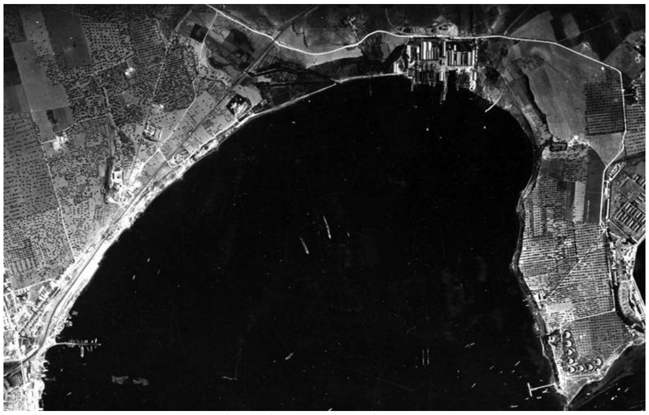

Figure 1. Northern arc of the first Sinus of the Mar Piccolo - Orthophoto Institute Geographical Military (IGM) 1940

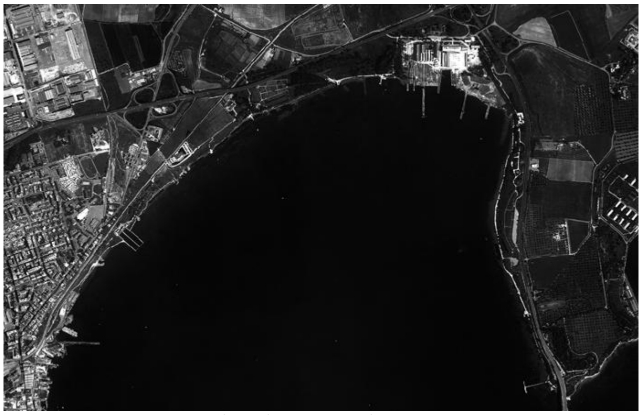

Figure 2. Northern arc of the first Sinus of the Mar Piccolo - Orthophoto Web Map Service (WMS) 2010

ISBN: 978-1-61208-326-1

We, therefore, subjected the two images, in view of their automated processing, with a scanning step of 20 microns, obtaining their digitization by a mosaic in which each tile was represented by its coordinates x, y and by the value of 8 bits of its gray level.

For each image, the mosaic thus obtained, we proceeded with an automatic extraction of “patch” with the criterion of “similarity” between tiles assigned to satisfying the criteria of closeness and proximity of the relative levels of gray.

“Patch” in ecology means the structural unit of an environmental system heterogeneous, identified on the basis on the differences that appear within the area itself.

We have thus converted each image into a set of “patches” each of which were calculated fifteen numeric attributes related to the geometric properties of the patch (perimeter, delta-X and delta-Y, area, etc.) and others to its inertial properties (Jx, Jy, Rx, Ry, etc.).

From the image of 1940 were extracted 341 patches, whereas 438 will be extracted from the image of 2010, about a third more of the patches contained in the image of 1940. This first result shows how the environment has increased the fragmentation of the examined area.

The two analog images, when converted into two distinct sets of patches, 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].

IV. STATISTICAL ANALYSIS

The evaluation of the variance between the two aerial photos examined is based on the classification of patches 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 $Z=(Z_1,Z_2, \ldots, Z_k)$ in terms of k variables $Y_1,Y_2,\ldots Y_k$, linear combinations of the Z_j. It has:

$$Y_i = \sum_j b_{ij} Z_j \ (i=1,2,\ldots,k) \qquad (1)$$

where b_{ij} are constants to be determined. Y_i 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:

$$\mathrm{cov}(Y_i,Y_j)=0 \quad (i \neq j) \qquad (2)$$

$$V(Y_1) \geq V(Y_2) \geq \ldots \geq V(Y_k) \qquad (3)$$

where *cov* is covariance and *V* is variance. Without loss of generality we can assume that the variables Z_i 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:

$$Z_j = (X_j - \mu_j)/\sigma_j \qquad (4)$$

Also, impose the condition that the overall variance of Z_j is equal to that of $Y_{i,}$ i.e.:

$$\Sigma_i V(Y_i) = \Sigma_i V(Z_i) = k \qquad (5)$$

At last, suppose that the vectors

$$b_i = (b_{i,1}, b_{i,2},\ldots, b_{i,k}) \qquad (6)$$

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

$$\Sigma_j b^2_{ij} = 1 \ (i=1,2,\ldots, k) \qquad (7)$$

On account of this, the vectors b_i that maximize the variance of Y_1, of Y_2, …, to Y_k with the constraints (3) and (4), are the eigenvectors of the matrix C of the coefficients of correlation between the variables Z_j, which correspond to the eigenvalues λ_1, $\lambda_{2,}\ldots$, λ_k of **C**, sorted by non-increasing value. We then have:

$$|C - \lambda I| = 0 \qquad (8)$$

$$b_i (C - \lambda_i I) = 0 \qquad (9)$$

where I is the unit matrix. The matrix C is symmetric and positive definite for which the solutions λ_i of the (8) are non-negative and such that their sum (trace of the matrix C) is equal to k. We then have:

$$\Sigma_i \lambda_i = k \ \ (i=1,2,\ldots, k) \qquad (10)$$

The variance of the i-th component is:

$$V(Y_i) = \lambda_i \qquad (11)$$

And the contribution of Y_i to the overall variance is:

$$P_i = V(Y_i) / k = \lambda_i / k \qquad (12)$$

V. Spatial Statistical Models

This procedure allowed us to calculate the correlation coefficients between the 15 variables adopted to describe each patch, 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.

TABLE I. VARIABLES OF THE MODEL USED

Variable	Description
Nz	Z coordinate - Average depth
Nt	N.ro of pixels of the object
Area	Attributes of the object:-Area
Perimeter	Perimeter of the object (Edge detection – Sobel)
DeltaX	Xmx-Xmn
DeltaY	Ymx-Ymn
IdealArea	Ideal Area = DeltaX * DeltaY
Gx	- barycentre X
Gy	- barycentre Y
Jx	- moment of inertia with respect to the X axis
Jy	- moment of inertia with respect to the Y axis
Rx	- radius of inertia X
Ry	- radius of inertia Y
AreaRect	- area of the circumscribed rectangle.
RapportAAR	- relationship between area and area of the circumscribed rectangle.

By using the factor analysis of the first array of data - about the image of 1940 – there emerged five factors that explained 96% of the total variance of the system. Of these five factors, the first alone explained 49% of the variance while the other four factors explained, neatly, 22%, 13%, 7% and 5%.

Table II reports the values of composition of each factor in function of the original variables. The weights of the variables were calculated on each factor (factor loadings).

These weights may also be interpreted as the correlation coefficients between variable and factor.

TABLE II. ROTATED FACTOR MATRIX (FACTOR LOADINGS) - IGM 1940 - NUMBER FACTOR 5

Weight of the variables	F1	F2	F3	F4	F5
Nzm is correlated with factor 3	0.3294	-0.5032	0.7269	-0.1063	-0.1220
Nt is correlated with factor 2	0.0225	0.9623	0.0025	-0.0441	0.2119
Area is correlated with factor 2	0.0225	0.9623	0.0025	-0.0441	0.2119
Perimetro is correlated with factor 1	-0.8277	0.1478	-0.4097	0.1351	-0.0239
DeltaX is correlated with factor 1	-0.9129	0.2140	-0.0695	-0.2047	0.2345
DeltaY is correlated with factor 1	-0.7447	0.1449	0.2987	0.1453	0.4479
AreaIdeale is correlated with factor 1	-0.9084	0.2367	-0.0920	-0.2057	0.2198
Gx is correlated with factor 4	0.0999	0.0407	0.1019	0.9694	0.1508
Gy is correlated with factor 3	0.2159	0.0369	0.9192	0.2047	0.1815
Jx is correlated with factor 2	-0.3072	0.9288	-0.1034	0.0636	-0.0637
Jy is correlated with factor 2	-0.3076	0.9286	-0.1037	0.0636	-0.0635
Rx is correlated with factor 1	-0.9788	0.0787	-0.1247	-0.0472	0.0154
Ry is correlated with factor 1	-0.9795	0.0756	-0.1216	-0.0396	0.0155
AreaRett is correlated with factor 1	-0.9475	0.0306	-0.1927	0.0140	-0.0814
RapportoAAR is correlated with factor 5	-0.1775	0.1713	0.0751	0.1507	0.9249

We reduced the size of the area of the patch definition, the image of 1940, from 15 to 5, after which we reconstructed the images with the new standardized patch values greater than average, that is greater than zero - based on each major component. These are illustrated in Figure 3, Figure 4, Figure 5, Figure 6 and Figure 7.

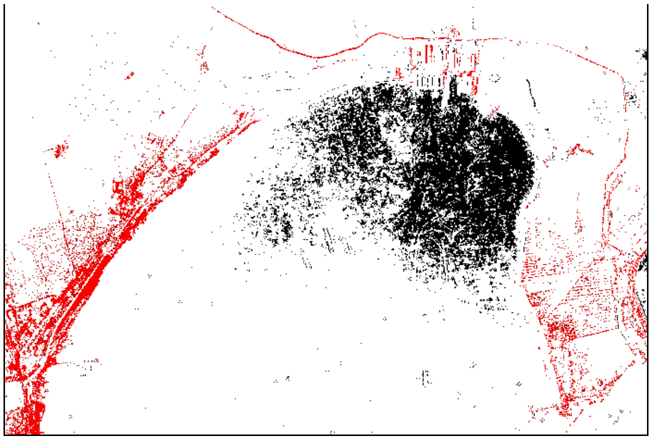

Figure 3. The reconstructed image with the I main component – 1940

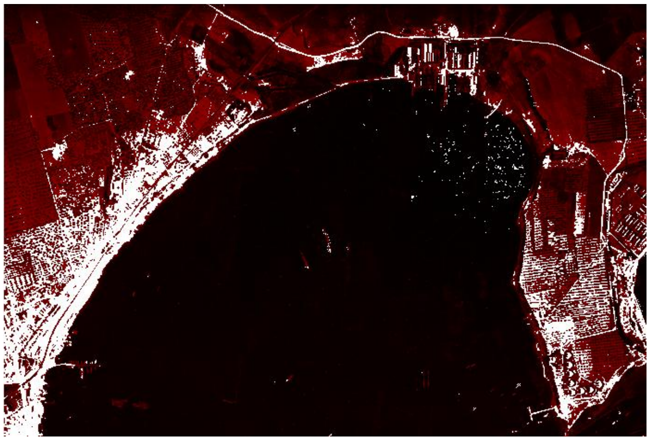

Figure 4. The reconstructed image with the II main component – 1940

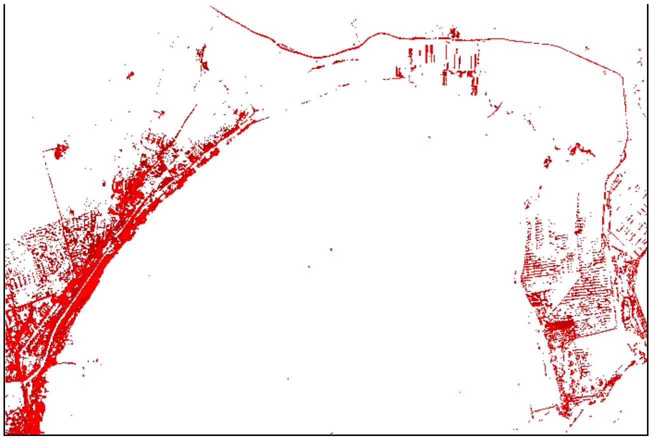

Figure 5. The reconstructed image with the III main component – 1940

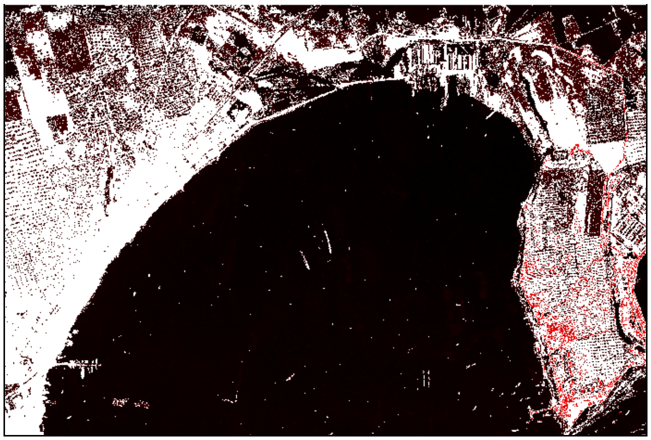

Figure 6. The reconstructed image with the IV main component – 1940

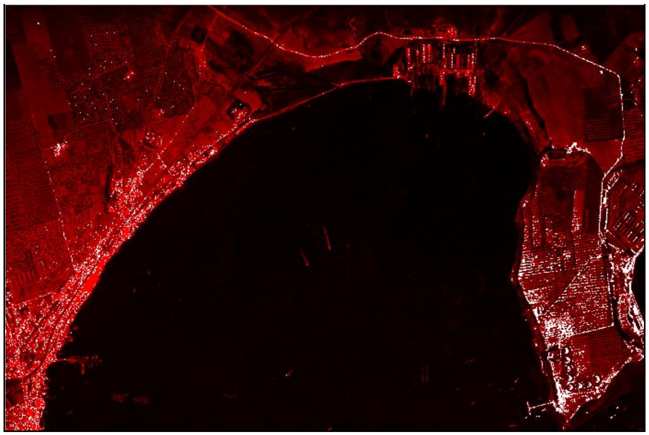

Figure 7. The reconstructed image with the V main component – 1940

Similarly, we proceeded to the WMS image taken in 2010. From this processing four main components (or factors) emerged that explained 90% of the overall variance of the system of the 438 patch (each described by 15 variables), while the first factor explained only 52% of the overall variance, the second factor 24%, 8% on the third, and the fourth 6%.

It can take the following list of factors emerged according to the percentage of variance explained: 50% = excellent, 40% = very good, 30% = good, 20% = sufficient, 10% = poor, <= 10% by ignore [15][16].

Table III reports the values of composition of each factor in function of the original variables. The weights of the variables were calculated on each factor (factor loadings).

TABLE III. ROTATED FACTOR MATRIX (FACTOR LOADINGS) – WMS 2010 - NUMBER FACTOR 4

Weight of the variables	**F1**	**F2**	**F3**	**F4**
Nzm is correlated with factor 4	0.4374	-0.4517	0.1415	0.5824
Nt is correlated with factor 2	-0.0667	0.9765	0.0660	0.0436
Area is correlated with factor 2	-0.0667	0.9765	0.0660	0.0436
Perimetro is correlated with factor 1	-0.7377	0.2125	-0.1163	-0.5314
DeltaX is correlated with factor 1	-0.9537	0.1984	-0.1589	-0.0500
DeltaY is correlated with factor 1	-0.9272	0.1492	-0.0867	0.1140
AreaIdeale is correlated with factor 1	-0.9469	0.2052	-0.1419	-0.0903
Gx is correlated with factor 3	0.4177	0.0521	0.7174	-0.2025
Gy is correlated with factor 3	0.0733	0.1317	0.9084	0.1544
Jx is correlated with factor 2	-0.2183	0.9404	0.0656	-0.1875
Jy is correlated with factor 2	-0.2189	0.9402	0.0654	-0.1875
Rx is correlated with factor 1	-0.9750	0.1039	-0.1355	-0.0577
Ry is correlated with factor 1	-0.9774	0.0929	-0.1023	-0.0244
AreaRett is correlated with factor 1	-0.9468	0.0773	-0.1085	-0.1710
RapportoAAR is correlated with factor 2	-0.1778	0.5313	-0.3406	0.4860

As can be seen in Figure 8, Figure 9, Figure 10 and Figure 11, there are shown standardized values - greater than the average - calculated on each main component.

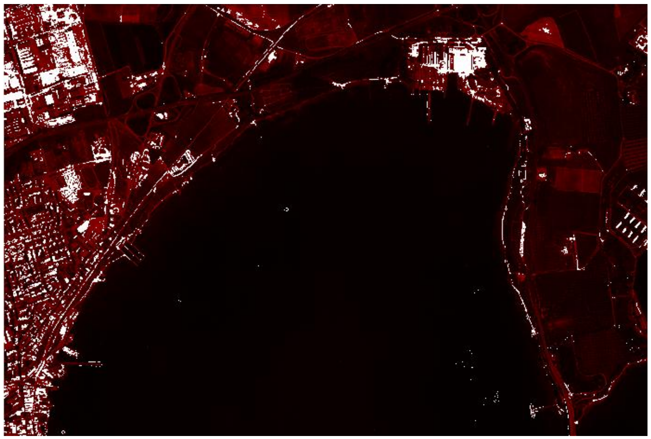

Figure 9. The reconstructed image with the II main component – 2010

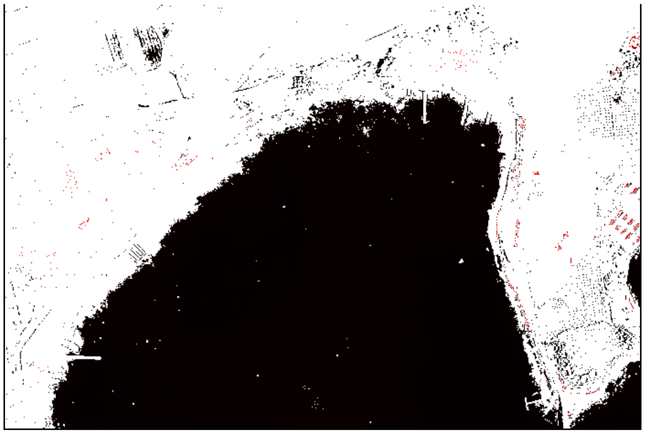

Figure 10. The reconstructed image with the III main component – 2010

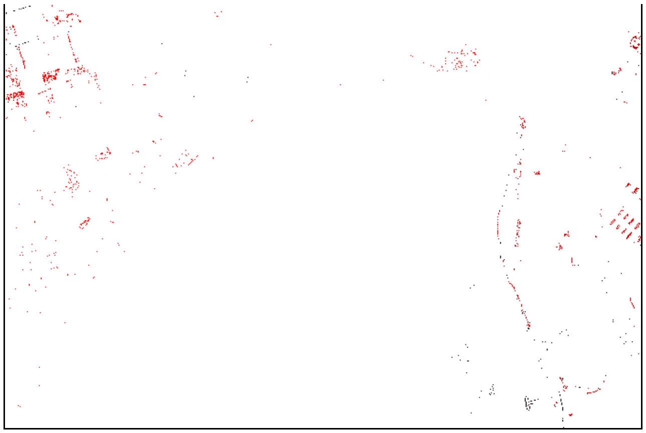

Figure 8. The reconstructed image with the I main component – 2010

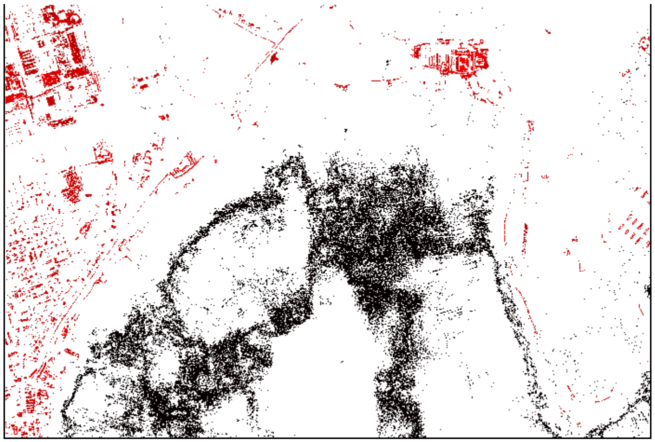

Figure 11. The reconstructed image with the IV main component – 2010

VI. Conclusion and future work

Factor analysis applied, as has been described, the first two images of the Sinus of the Mar Piccolo, respectively in the aerial photo of the Military Geographical Institute that dates back to 1940 and the aerial photo of the Web Map Service (WMS) in 2010, has enabled us to obtain the decomposition of the two analog images into layers that describe numerically the evolutionary peculiarities of the area examined. When considering only the main components characterized by an explanatory contribution to the overall variance of the system more than 10%, it can be seen that the landscape of the first Sinus of the Mar Piccolo of Taranto has suffered fragmentation from 1940 to 2010, a fact that is evidenced by an increase in patches extracted from each of the images taken between 1940 and 2010. In addition, the composition of the loading factors between variables and factors remains generally unchanged between 1940 and 2010, which is meant to signify that there was not a material change in the landscape between 1940 and 2010, with the exception of the arc, where western coastal residential developments grew on significant extensions of territory. It is very interesting to note the comparison of the image reconstructed with the first principal component orthophoto of 1940, (see Figure 3), and the reconstructed image with the fourth main component of the image of the Web Map Service (WMS) 2010 (see Figure 11). Both of these digital images, obtained from the analysis of the main components, seem to indicate what was a significant numerical evolution of the body of water in the first Sinus of the Mar Piccolo between 1940 and 2010. Perhaps it would also be useful to search for physical causes of this evolution.

Acknowledgment

We thank the laboratory of Topography and Photogrammetry of the University of Salento and the Web Map Service (WMS) offered by the geoportal of the Region of Puglia.

References

[1] S. M. Niladri, G. Susmita, and G. Ashish, Fuzzy clustering algorithms incorporating local information for change detection in remotely sensed images Applied Soft Computing, vol. 12, iss. 8, August 2012, pp. 2683-2692.

[2] L. Castellana, A. D'Addabbo, and G. Pasquariello, A composed supervised/unsupervised approach to improve change detection from remote sensing Pattern Recognition Letters, vol. 28, iss. 4, 1 March 2007, pp. 405-413

[3] J. S. Deng , K. Wang , Y. H. Deng & G. J. Qi (2008) PCA based land use change detection and analysis using multitemporal and multisensor satellite data, International Journal of remoote Sensing, 29:16, pp. 4823-4838, DOI: 10.1080/01431160801950162

[4] P. Coppin, I. Jonckheere, K. Nackaerts, and B. Muys, Digital change detection methods in ecosystem monitoring: a review. International Journal of Remote Sensing, 25, pp. 1565-1596, 2004.

[5] J. Byeungwoo, and A. David, Partially supervised classification using weighted unsupervised clustering. IEEE T. Geoscience and Remote Sensing (TGRS) 37(2):1073-1079 1999.

[6] L.G. Browna, Survey of Image Registration Techniques, ACM Computing Surveys, Vol 24, No. 4, December 1992.

[7] K.P. Schwarz, M.A. Chapman, M.E. Cannon, P. Gong, An Integrated INS/GPS Approach to the Georeferencing of Remotely Sensed Data, Photogrammetric Engineering & Remote Sensing, 59(11): 1167-1674, 1993.

[8] H. Hotelling, The generalization of Student's Ratio. Ann. Math. Statist.,vol. 2, pp. 30-378, 1931.

[9] M. Zevi, The matrix calculation in the method of the components princiapli. Faculty of Architecture, Rome 1977.

[10] F. Ricci, Statistics and Statistical Processing of the Information. Zanichelli, Bologna 1975.

[11] S. Saddocchi, manuale di analisi statistica multivariata F. Angeli, Milano, 1993.

[12] A. Mezzetti, Air pollution and vegetation, Edagricole, 1987.

[13] M.A. Fischler, and R.C. Bolles, Random Sample Consensus: a Paradigm for Mode l Fitting with Applications to Image Analysis and Automated Cartography. Comm. ACM, no. 24, pp. 381-395, 1981.

[14] M. Cramer, D. Stallmann, and N. Halla, High Precision Georeferencing Using GPS/INS and Image Matching, Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Alberta, Canada, June 3-6, pp. 453-462, 1997

[15] G. Zuendorf, N. Kerrouche, K. Herholz, and J.C. Baron, Efficient Principal Component Analysis for multivariate 3D Voxel - based mapping of brain functional imaging data sets as applied to FDG - PET and normal aging, no. 18, pp. 13-21, 2003.

[16] M. Mazzei, and A. L. Palma, Spatial Statistical Models for the Evaluation of the Landscape, Lecture Notes in Computer Science, Computational Science and Its Applications, ICCSA June 2013, pp. 419-432, doi 10.1007/978-3-642-39649-6_30.