Software development for unsupervised approach to identification of a multi temporal spatial analysis model

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

Abstract — This software development 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 according to image processing, pattern recognition and filter algorithm. 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 radiances. 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 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 Geopoint’s Web Map Service (WMS) of the Puglia Region, which has kindly allowed us to use the orthophotos of the same site observed 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 [14]. 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 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.

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 into pixel.

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

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.

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 into pixel.

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

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.

Digital images are two-dimensional arrays, each element of which contains the pixel value to be displayed. In particular for black and white images this value goes from 0 (Black) to 255 (White). To calculate the partial derivatives of this discrete representation, discrete operators will be

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
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 gray. 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.

We have thus converted each image 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.).

From the image of 1940 were extracted 341 objects, whereas 438 will be extracted from the image of 2010, about a third more of the objects 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 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 study statistics, and in the field of the phenomena involved in the study [12][13][17][19][20].

IV. 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 Zᵢ. It has:

$$Y_i = \sum b_j Z_j (i=1,2,...,k)$$  

(1)

where bᵢ 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:

$$\text{cov}(Y_i,Y_j)=0 \ (i \neq j)$$  

(2)

$$V(Y_i) \geq V(Y_j) \geq \ldots \geq V(Y_k)$$  

(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:

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

(4)

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

$$\sum V(Y_j) = \sum V(Z_j) = k$$  

(5)

At last, suppose that the vectors

$$b_1=(b_{i,1}, b_{i,2},\ldots, b_{i,k})$$  

(6)

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

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

(7)

On account of this, the vectors bᵢ that maximize the variance of Yᵢ, of Y₂, ..., to Yᵦ with the constraints (3) and (4), are the eigenvectors of the matrix C of the coefficients of correlation between the variables Zᵢ, which correspond to the eigenvalues λ₁, λ₂, ..., λᵦ of C, sorted by non-increasing value. We then have:

$$|C - \lambda I| = 0$$  

(8)

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

(9)

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

$$\sum \lambda_i = k \ (i=1,2,\ldots, k)$$  

(10)

The variance of the i-th component is:

$$V(Y_i) = \lambda_i$$  

(11)

And the contribution of Yᵢ to the overall variance is:

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

(12)

V. SPATIAL STATISTICAL MODELS

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

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, nearly, 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

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

We reduced the size of the area of the object definition, the image of 1940, from 15 to 5, after which we reconstructed the images with the new standardized object values greater than average, that is greater than zero - based on each major component. These are illustrated in Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8.
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 object (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][18].

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>
<thead>
<tr>
<th>Weight variables</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name is correlated with factor 4</td>
<td>0.4374</td>
<td>-0.4517</td>
<td>0.1415</td>
<td>0.5824</td>
</tr>
<tr>
<td>Nt is correlated with factor 2</td>
<td>-0.0667</td>
<td>0.9765</td>
<td>0.0660</td>
<td>0.0436</td>
</tr>
<tr>
<td>Area is correlated with factor 2</td>
<td>-0.0667</td>
<td>0.9765</td>
<td>0.0660</td>
<td>0.0436</td>
</tr>
<tr>
<td>Perimetro is correlated with factor 1</td>
<td>-0.7377</td>
<td>0.2125</td>
<td>-0.1163</td>
<td>-0.5314</td>
</tr>
<tr>
<td>DeltaX is correlated with factor 1</td>
<td>-0.9537</td>
<td>0.1984</td>
<td>-0.1589</td>
<td>-0.0500</td>
</tr>
<tr>
<td>DeltaY is correlated with factor 1</td>
<td>-0.9272</td>
<td>0.1492</td>
<td>-0.0867</td>
<td>0.1140</td>
</tr>
<tr>
<td>Areaideale is correlated with factor 1</td>
<td>-0.9469</td>
<td>0.2052</td>
<td>-0.1419</td>
<td>-0.0903</td>
</tr>
<tr>
<td>Gx is correlated with factor 3</td>
<td>0.4177</td>
<td>0.0521</td>
<td>0.7174</td>
<td>-0.2025</td>
</tr>
<tr>
<td>Gy is correlated with factor 3</td>
<td>0.0733</td>
<td>0.1317</td>
<td>0.9084</td>
<td>0.1544</td>
</tr>
<tr>
<td>Jx is correlated with factor 2</td>
<td>-0.2183</td>
<td>0.9404</td>
<td>0.0656</td>
<td>-0.1875</td>
</tr>
<tr>
<td>Jy is correlated with factor 2</td>
<td>-0.2189</td>
<td>0.9402</td>
<td>0.0654</td>
<td>-0.1875</td>
</tr>
<tr>
<td>Rx is correlated with factor 1</td>
<td>-0.9750</td>
<td>0.1039</td>
<td>-0.1355</td>
<td>-0.0577</td>
</tr>
<tr>
<td>Ry is correlated with factor 1</td>
<td>-0.9774</td>
<td>0.0929</td>
<td>-0.1023</td>
<td>-0.0244</td>
</tr>
<tr>
<td>AreaRett is correlated with factor 1</td>
<td>-0.9468</td>
<td>0.0773</td>
<td>-0.1085</td>
<td>-0.1710</td>
</tr>
<tr>
<td>RapportoAAR is correlated with factor 2</td>
<td>-0.1778</td>
<td>0.5313</td>
<td>-0.3406</td>
<td>0.4860</td>
</tr>
</tbody>
</table>
As can be seen in Figure 9, Figure 10, Figure 11 and Figure 12, there are shown standardized values - greater than the average - calculated on each main component.

VI. CONCLUSION AND FUTURE WORK

Factor analysis summarized very much the starting information without reconstructing a significant pattern normally used in supervised methods. Figures 13-14 show the correlations between factors and their spatial distribution in a multi-dimensional space.

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 objects 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. In 2016, the Italian navy has mapped a large amount of anthropogenic materials into the small sea of Taranto. The waste has been geolocalized, video-recorded and cataloged with tags. The Navy hydrographic institute now has a complete database, a submarine map of the seabed of Mar Piccolo and the waste thrown into the sea.
These results can confirm the anomalies highlighted in this work. The data confirm that in the time interval between 1940 and 2010 there are disturbing elements that can’t be classified as part of the surrounding territory. The methodological approach and the developed software result to be a very useful tool for investigating territorial differences.

REFERENCES


