KNOWLEDGE-BASED SYSTEM FOR AUTOMATIC GENERALIZATION OF SATELLITE-DERIVED THEMATIC MAPS

Stefania Goffredo
Institute for Remote Sensing Applications, Environmental Mapping Unit,
European Commission, Joint Research Centre, 21020 Ispra (Varese) Italy
e.mail: stefania.goffredo@jrc.it
and
University of Leicester, Dept. of Geography, Leicester LE1 7RH, UK

Abstract

This paper presents an original automatic processing chain for the spatial and thematic generalization of satellite-derived maps to be introduced in a Geographical Information System (GIS) data base. The prototype system uses image understanding and artificial intelligence techniques, combined with fundamental cartographic generalization principles. The aim for this research was to show that a completely automatic and objective map-making processing chain could be realized, taking the advantages from both Remote Sensing and Cartography disciplines while reducing the negative side-effects. A GIS represents the base on which to organize a comprehensive integration of the two disciplines. The prototype system has been developed to automatically integrate remotely sensed data with the CORINE Data Base, and the prototype has been organized in a 2-level generalization procedure (low-level, high-level). The generalization prototype system works individually on objects, closed polygons detected by Edge Detection technique from the original satellite image. The input polygons are pre-processed to automatically prepare the necessary knowledge to run the low-level generalization. The input to be generalized consists of satellite-derived thematic products in ERDAS format: in this case, an Artificial Neural Network classification. The classification is then simplified, by the spatial and thematic simplification of each polygon during the low-level procedure, and from the thematic abstraction during the high-level procedure. The output consists of a map of main homogeneous patches, each of which is individually identified by a unique CORINE land cover class.

1 Generalization

Remote Sensing can be considered one of the most powerful tools for Earth monitoring and mapping. However it is known that classifications (even if obtained by the most innovative pixel-based image processing algorithm) may not be directly applicable to general land cover applications because they are too rich in detail [1]. The role of generalization is to raise the classification level of abstraction to a higher one directly comparable to maps stored in Geographical Information Systems (GIS).

Up to now, remotely sensed data have been considered by cartographers, and other Earth sciences experts, as auxiliary data to be referenced in land cover mapping using visual-manual approaches. Remotely sensed data however can be used in a more significant way to create a ready to use data base. It has been demonstrated in literature that the integration of Remote Sensing and GIS is very attractive and powerful in environmental managing, where Earth monitoring and mapping and decision-making are fundamental activities [2, 3]. However this integration is not completely explored and realized. Too many human interactions are still involved to connect the two activities in map-making processes. Generalization is the necessary intermediate step to allow a completely autonomous and automated remote sensing and GIS integration.

In the last two years a prototype of an Automatic Spatial Generalization System has been created and it is now at the final stage of development. The Generalization System prototype consists of an automatic processing chain capable of transforming a satellite derived thematic product into a spatially and thematically generalized map to introduce into a GIS data base. The fundamental characteristics of this raster generalization system are the combination of basic principles of data integration and cartographic generalization.

2 Method

As shown in the scheme in Figure 1, an initial integration of data, independently and automatically obtained from the original satellite image, a Multi-spectral LANDSAT TM, represents the starting point of the Generalization System.

The Generalization System is divided in two main procedures: low-level generalization and high-level generalization.

The Low-Level Generalization includes the following subprocesses:

- Automatic Detection and Treatment of Slivers within each Polygon
- Rule-Based Polygon Merging
- Polygon Thematic Smoothing: combination of filtering and region growing techniques

The typical pixel-by-pixel scanning procedure of the raster image processing is not appropriate during the generalization procedure, therefore an abstraction of the image content unit concept has been made, considering an entire polygon being the image content unit. Considering the different shape, size and contents of each polygon, it is highly improbable to find objective thresholds to control and stop an iterative algorithm, such as the typical raster filtering thresholds: reached number of iterations or reached percentage. Iterations during the automatic smoothing activity, are not then available in the prototype.

The High-Level Generalization performs the final generalization activity on the output of the low-level procedure, converting the input spectral classes into map classes. Expert System (ES) and Artificial Neural Network (ANN) approaches are used in this final procedure. ANN determines a correspondence between the low-level (spectral) classes hierarchy and the Corine classes hierarchy. The ES converts the low-level product into the final map.

The prototype has been designed and developed respecting objectivity and the fundamental cartographic principles. The prototype involves directly: already existing processes (if suitable) modified processes (when necessary) and newly created processes.

Comparing advantages and disadvantages of already existing smoothing procedures, Iterative Majority Filtering (IMF) [4, 5] has been considered the easiest algorithm to implement, and the most powerful. However, the original IMF algorithm has been modified to obtain a better match to the generalization constraints. The main IMF disadvantage is its over-crossing, during its application, of the natural regional borders clearly delineated in the original satellite image and kept after the classification. A cartographer "visually" fixes main regional borders before applying, within these borders, smoothing activities. The simulation of this fundamental cartographic action, is necessary.

Automatic Edge Detection is the solution to this problem of "over-crossing" at the actual prototype status. An already existing algorithm, Significant Edge Detection [6] have been chosen. SED detects irregular closed polygons from the original satellite image. Generally, segmentation algorithms can detect very detailed objects. SED,.in fact, detects polygons at different shape and size, in a range from a minimum of 4-5 pixels perimeter up to a maximum of hundreds pixels perimeter. In our specific test image, 256x256 size, 2032 polygons have been automatically extracted by SED. However, too many detected polygons are and they are too small to render significative the smoothing algorithm within their borders.

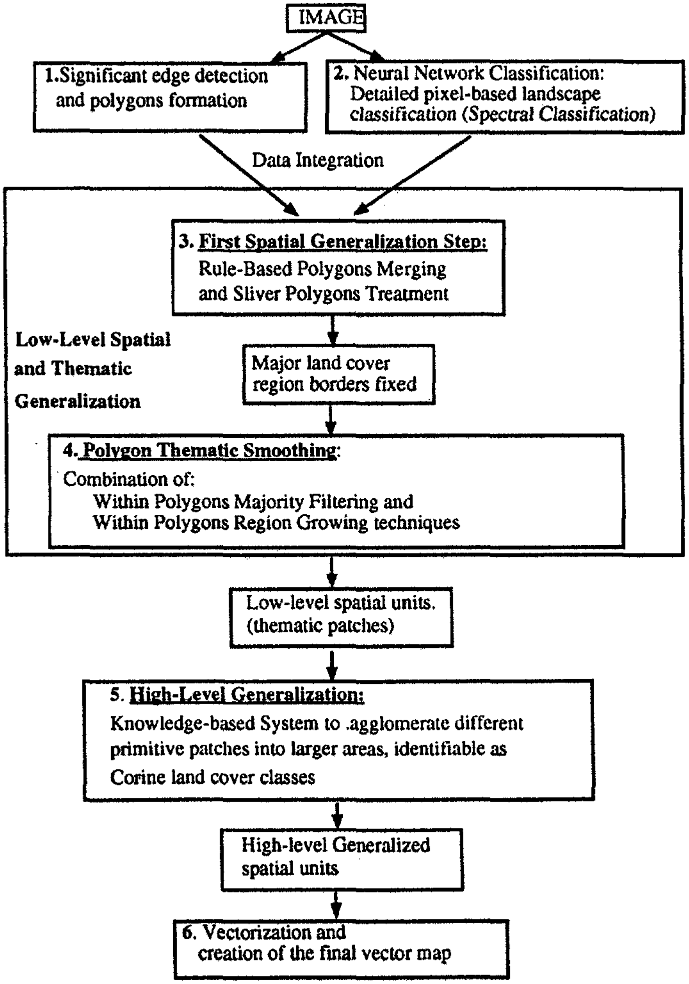
IMAGE
1.Significant edge detection and polygons formation
2. Neural Network Classification: Detailed pixel-based landscape classification (Spectral Classification)
Data Integration
Low-Level Spatial and Thematic Generalization
3. First Spatial Generalization Step: Rule-Based Polygons Merging and Sliver Polygons Treatment
Major land cover region borders fixed
4. Polygon Thematic Smoothing: Combination of: Within Polygons Majority Filtering and Within Polygons Region Growing techniques
Low-level spatial units. (thematic patches)
5. High-Level Generalization: Knowledge-based System to .agglomerate different primitive patches into larger areas, identifiable as Corine land cover classes
High-level Generalized spatial units
6. Vectorization and creation of the final vector map

Figure 1. Block Diagram of Automatic Generalization Procedure

2.1 Low-Level Generalization

The aim for the low level generalization procedure is to smooth the classified image eliminating, not only noise and misclassifications (responsible for the typical not actractive scatter appearance of the classifications), but also to reduce the details. The detailed information has the aspect of (small) spatially insignificant thematic patches which can easily be included into (larger) spatially more significant adjacent areas.

The term Low-Level associated to this step, is due to the fact that the generalization involved is related to the spatial re-adjustment of the input spectral thematic patches. No change of abstraction level is involved at this stage. There is the transition from the pixel-per-pixel point of view to the new polygon-per-polygon point of view. To reach the aim, new procedures have been developed and already existing ones have been modified to match the requirements of the new polygon vision.

This first step in the generalization system is based on the integration of different types of data automatically and independently extracted from the same original LANDSAT image: 1) Closed Polygon Map, obtained by an Edge Detection Procedure, and 2) Classified Image obtained by an Artificial Neural Network Classifier [7]. The integration follows the fundamental principles adopted by the cartographic generalization.

Raster image processing can be so powerful to produce very detailed outputs. Due to the pixel-per-pixel point of view, these outputs can show very small particulars in the scene. The cartographer eye, guided by the experience and by the knowledge *stored* in her/his brain, selects automatically the important information excluding the rest from it. The automatic extraction of objects or borders performed by segmentation and edge detection techniques, is an attempt to simulate this human activity. However, the raster activity gives outputs which are not immediately suitable for general purpose use. For example, the raster edge detection used for the prototype produces a polygons map of 2032 objects (closed polygons) over a small subset image of 256x256 pixels size.

The first and fundamental abstraction needed in a generalization system is obviously related to the definition of a new *spatial unit* in the image: a spatial entity considered as a single unit, even if it is formed by groups of other elements. In the prototype system, this entity is the *polygon*, formed by a border and a thematic content. The border consists of a set of connected edges oriented in the four different cardinal orientations: NS, NE, NW, WE. Each border's edge is a set of connected pixels, for each of which the (x,y) coordinates are recorded. The polygon's content consists of a set of pixels values (the themes) associated to each pixel belonging to the area surrounded by the polygon's border.

A brief scheme of the involved procedures is:

- DATA ACQUISITION AND REORGANIZATION OF:
  - per-pixel borders description of each polygon, found in the Closed Polygon Map; extraction and organization of the polygons attributes: perimeter, border gradient value, adjacents, shared borders among all adjacents etc.
  - classified image
  - overlaying of the two maps and consequent updating of the polygons attributes: content (spectral classes contained within each polygon), statistics on the content themes etc.
- POLYGONS MANAGING(): processing, statistically and thematically, each polygons by rule-based activities, in order to obtain larger polygons; within the borders of which the smoothing algorithm is applied. The input data are the original Closed Polygon Map and the Classified Image. The output of each procedure is the input for the subsequent one.
  - *First_Merging_Rule()*: iterative procedure. Application of the Rule R1 to each polygon and its adjacents.

R1: IF the current polygon contains a class occupying at least the 90% of the polygon's area, *THEN* the polygon is considered containing that unique class.

R1bis: take the current polygon and all of its adjacents forming polygons pairs as

(curr_poly, adjac*1*);

...

...

(curr_poly,adjac*n*);

and examine the content of each pair.

IF the adjacent j contains the same class contained into the current polygon, *THEN* the pair is merged together.

In case of ties, the "larger" adjacent is taken for merging.

For each iteration only one merging is allowed. After each iteration the Closed Polygon Map is automatically updated with new attributes, to be the input for the next iteration.

— *Sliver_Treatment()*: recognition and treatment of slivers within polygons; each polygon is independently treated.

— *TP_Merging_Rule()*: iterative procedure; it works with the same behavior of the First_Merging_Rule(), but a different threshold is considered here. The most compatible adjacent of the current polygon, having a Perimenter minor then a certain threshold is merged to the current polygon. At the actual prototype stage, the term "compatible" means that the adjacent contains the same classes of the current polygon. Obviously more sophisticated rules can be used to define the compatibility among polygons, introducing for example the concept of "statistical compatibility"; not only similar contents but further similar percentages.

— *TG_Merging_Rule()*: iterative procedure; the term of comparison here is the average gradient value associated to each border shared between a pair of polygons. A compatible adjacent to the current polygon having Gradient value minor then a certain threshold is merged to the current polygons.

Once the Polygon Managing process is stopped, an updated map of larger Closed Polygons is ready. The underlying classified map has been slightly modified, principally by the First_Merging_Rule(). Both the updated maps are the input for the second process of the Low-Level Generalization.

- PMF_PRCG THEMATIC SMOOTHING
  — *PMF* = non iterative Majority Filtering constrained within the Polygon borders
  — *PRCG* = non iterative Reduced Class Growing constrained by the Polygons borders

This process is a pure thematic smoothing of the classification applied independently within each regional (polygon) border. It consists of a modified version of Majority Filtering and Region Growing procedures. Majority Filtering has been chosen for its strong smoothing power and for the simplicity of its implementation. In the modified version, the filtering is constrained within borders and no iterations are allowed, to avoid the need for human supervision during the application. Region Growing [8] has been chosen as a complement to the majority filtering, having the scope to re-introduce homogeneously extended patches which, although eliminated by the filtering, represent a large part of the polygon. Region Growing has been modified from the original algorithm by allowing no iterations, restricting its application within polygon borders and checking (before the re-introduction) where, within the polygon, the original homogeneous patch was placed.

The output of the Low-Level procedure consists of a map where homogeneous thematic patches are shown, and the main regional borders have been kept. This thematic map is further generalized by the

High-Level generalization procedure.

2.2 *High-Level Generalization*

The aim at this step of the system is the change of abstraction level, passing from the spectral classes scheme, of the input classification, to the more general Corine land cover classes scheme, to use for the final map. The development of an Expert System is the kernel of the high level procedure. The analysis phase is still under examination, however the fundamental structure of the expert system is already clear, and the firsts experiments have already given interesting results.

First it is necessary to determine a common level of "comparison" between the spectral classes scheme and the Corine classes scheme:

- determine which land cover classes can be automatically recognized by image understanding algorithms. Among these are classes such as Transport Units (roads, railways), Airports and Ports, recognizable by shape and context techniques integrating different data sources: SPOT images, aerial photographs already introduced into a GIS data base
- land cover classes not automatically recognizable are substituted by their parents in the hierarchical scheme

The modified Corine land cover classes scheme, that will be used during the high level procedure, is called Pseudo Corine Class Scheme (PCC).

AVAILABLE DATA:

- low-level generalized map
- Corine map
- Agro-Forestal cartographic map
- Cadastral cartographic map

from the Agro-Forestal Map, for example, it is possible to extract information to organize into an appropriate data base for the expert system:

— tangerine forest polygons
— olive forest polygons
— other trees species forest polygons
— vineyards polygons
— non-irrigated culture fields polygons
— irrigated culture fields polygons
— rice fields polygons
— main urban areas
— rural urban areas

this is only an example of the information that may be taken from the available data.

For the expert system usability, the PCClasses are grouped into 5 different general types of land use: Agricultural, Natural, Urban, Man-Made Ecological and Artificial.

RULES SET 1: The expert system decides the general land use type of the area under study depending on the percentages of the PCClasses present.

First the general land use type is established on the subset of the Corine map corresponding to the satellite derived thematic map; the general land use type on the low-level generalized map will then be established. If the two general land use types match then the process can go on.

RULES SET 2: automatic re-labeling of the Spectral Classes into Pseudo Corine Classes. This is a "simple" but important point. To establish that the spectral class 'cereals' is equivalent to the Pseudo Corine Class 'non-irrigated arable land' is trivial for a human expert in land cultivations, but it is very difficult "explaining" to a computing program that these two different sets of words have the same meaning. If the problem cannot be rapidly solved in a 'semantic' way, a syntactic solution can be considered. For this purpose a type of Dictionary has been created for both the classes scheme: a Dictionary for the Spectral Classes (SCDictionary) and a Dictionary for the Pseudo Corine Classes (PCCDictionary), for each entry class name, a list of synonyms, commonly used in land cover context, is associated.

A syntactic procedure for the automatic re-labeling of the Spectral Classes into PCC has been considered: for each spectral class, each related term stored in the SCDictionary is syntactically compared to each term relative to each PCC stored in the PCCDictionary. When a match is found, then the Spectral Class is re-labeled as the PCC owner of the matching term.

RULES SET 3: analysis of the remaining Spectral Classes. If no direct re-labeling is possible, it means that these Spectral Classes must be associated to others in order to form a PCC.

The comparison of all the available data is fundamental. It could be relevant to analyze the neighbors for each PCC, trying to establish "objective" relationships, for example high-density urban never has rural neighbors (agricultural fields, forest, rice field etc).

It is under analysis the efficacy of Artificial Neural Networks (ANN) during the establishment of these neighbors relationships. Initially it has been thought to use ANN to establish relationships among combinations of spectral classes and Corine land cover classes.

Overlaying the low-level generalization up to the correspondent scene in the Corine map, it has been possible to detect which spectral classes, and in which percentages, are contained within each Corine polygons. Percentages have been extracted and an attempt to study them has been made. The data extracted were not completely satisfactory. Because of the Corine polygons' size, often extremely large, the subscene represented by the low level generalized map was completely contained into one unique Corine polygon. No conclusions could then be inducted. This should be obvious if we consider that Corine polygons are manually delineated and labeled.

A similar attempt has been made overlaying the whole (1024x1024) classified image to the whole Corine map. Because of the abundance of details present in the classification, Corine polygons labeled with the same land cover class, contained completely different combinations of spectral classes, not only in percentages, but, above all, in terms of classes. Basically all the spectral classes recognized by the classification, were contemporary present within each Corine polygon. For example in the Corine map there were 27 polygons labeled with the land cover class 111 'continuous urban fabrics'. The scheme in Figure 2 shows spectral classes combinations relative to 6 of the 27 Corine polygons.

Only the Corine polygon 4 could represent "what" a human expects to find in a high density urban area: highest percentage of tiled, concrete, a lower percentage of grass and no agricultural cultivation. Considering the range of the percentages of tiled, concrete, varying from a minimum of 16% to a maximum of 73%, the range of the percentages of agricultural cultivation varying from a minimum of 0% to a maximum of 22, the range of sea, fresh water varying from a minimum of 0% to a maximum of 22%, the range of grassland, weeds varying from a minimum of 20% to a maximum of 78%, the range of forests varying from a minimum of 0% to a maximum of a maximum of 4%, it is very difficult to set significative thresholds. Some other Corine class can have similar ranges of percentages to the ones associated to the land cover class 'continuous urban fabrics' (111). If the fundamental attribute to distinguish the Corine class 111 is the percentage of the spectral class tiled, concrete, then if the threshold for the majority is too high (< 50%), most of the spectral combinations associated to the corine class 111 having tiled/concrete from 16% to 49% will be discarded, accepting, instead, spectral combinations associated to different Corine land cover classes.

The problem to train the ANN with this set of data, was to find other auxiliary information capable to discriminate between similar spectral combinations but associated to different Corine classes.

It is the author's opinion that because of the 'guiding principle' of ANN, to obtain satisfactory results it is necessary to define easily and univoquely the model of the problem we want to solve, in order to avoid ambiguities. With the two experiments described above, too many ambiguities were involved, therefore it was not possible to determine objective relations.

At the actual status of the expert system design, the use of ANN can be differently applied. All the thematic patches present in the low level generalized map, not automatically translated by the Rules set 2, into PCClasses, will be iteratively associated to their adjacent patches forming combinations of spectral patches. Depending on the neighbors, there will be one combination more reliable to represent a Corine class than the other combinations, then this Corine class will be chosen to re-label this combination.

The main difficulty in the conduction of this analysis is to be sure that the objectivity is respected. It is fundamental for the scope of the project to find automatic and objective solutions, in order to utilize the prototype in different mapping fields: forest, soil erosion and other disciplines in the Earth surface study. The prototype must work independently from the mapping context; once provided the appropriate data base, the appropriate dictionaries and the appropriate selection of rules, the prototype should "generalize".

Spectral Classes

SC1 = Tiled, concrete
SC2 = Bare sand
SC3 = Clayey bare soil
SC4 = Sea water
SC5 = Freshwater, lakes
SC6 = Coastal estuaries, lagoons
SC7 = Wheat plantations
SC8 = Barley plantations
SC9 = Maize plantations
SC10 = Rice plantations
SC11 = Vineyards
SC12 = Weeds
SC13 = Garrigue
SC14 = Grassland
SC15 = Deciduous forest
SC16 = Coniferous forest

Corine polygon 1
land cover class 111

SC1 = 19%	
SC2 = 3%	28% tiled, concrete
SC3 = 6%	
SC4 = 0%	
SC5 = 1%	
SC6 = 0%	
SC7 = 1%	
SC8 = 2%	
SC9 = 1%	14% agricultural cultiv
SC10 = 2%	
SC11 = 8%	
SC12 = 18%	
SC13 = 2%	56% grassland, weeds
SC14 = 36%	
SC15 = 0%	
SC16 = 0%	

Corine polygon 2
land cover class 111

SC1 = 1%	
SC2 = 0%	**16%** tiled, concrete
SC3 = 15%	
SC4 = 0%	**22%** fresh water, sea, lagoon
SC5 = 17%	
SC6 = 5%	
SC7 = 3%	
SC8 = 1%	
SC9 = 1%	**12%** agricultural cultiv
SC10 = 5%	
SC11 = 2%	
SC12 = 21%	
SC13 = 0%	**50%** grassland, weeds
SC14 = 29%	
SC15 = 0%	
SC16 = 0%	

Corine polygon 3
land cover class 111

SC1 = 36%	
SC2 = 1%	38% tiled, concrete
SC3 = 1%	
SC4 = 0%	
SC5 = 1%	
SC6 = 0%	
SC7 = 1%	
SC8 = 1%	
SC9 = 1%	8% agricultural cultiv
SC10 = 0%	
SC11 = 5%	
SC12 = 25%	
SC13 = 0%	54% grassland, weeds
SC14 = 29%	
SC15 = 0%	
SC16 = 0%	

Corine polygon 4
land cover class 111

SC1 = 65%	
SC2 = 3%	**70%** tiled, concrete
SC3 = 2%	
SC4 = 0%	
SC5 = 0%	
SC6 = 0%	
SC7 = 5%	
SC8 = 2%	
SC9 = 1%	9% agricultural cultiv
SC10 = 0%	
SC11 = 1%	
SC12 = 9%	
SC13 = 1%	20% grassland, weeds
SC14 = 10%	
SC15 = 0%	
SC16 = 0%	

Corine polygon 5
land cover class 111

SC1 = 18%	
SC2 = 0%	26% tiled, concrete
SC3 = 8%	
SC4 = 0%	
SC5 = 1%	3% fresh water, sea
SC6 = 1%	
SC7 = 1%	
SC8 = 0%	
SC9 = 0%	4% agricultural cultiv
SC10 = 1%	
SC11 = 2%	
SC12 = 35%	
SC13 = 1%	66% grassland, weeds
SC14 = 30%	
SC15 = 1%	4% forests
SC16 = 3%	

Corine polygon 6
land cover class 111

SC1 = 13%	
SC2 = 1%	**18%** tiled, concrete
SC3 = 4%	
SC4 = 0%	
SC5 = 3%	3% fresh water, sea
SC6 = 0%	
SC7 = 0%	
SC8 = 0%	
SC9 = 0%	
SC10 = 0%	
SC11 = 0%	
SC12 = 48%	
SC13 = 0%	**78%** grassland, weeds
SC14 = 30%	
SC15 = 0%	
SC16 = 0%	

Figure 2. Samples of the relationships between combinations of spectral classes and the Corinc land cover class "continuous urban fabrics"

References

[1] Goffredo, S., and Chiuderi, A., 1995. From Data to Maps: the Exploitation of Remote Sensing. In press on Sistema Terra, Remote Sensing and the Earth, May 1995, Editors Laterza

[2] Goffredo, S., Kannellopoulos, I., and Wilkinson, G. G., 1993. Integrazione GIS e telerilevamento nella gestione dell'ambiente in Europa. Alcune attività del Centro Comune di Ricerca della Comunità Europea. Oral presentation at the seminar "Geographic Information Systems (GIS): un indispensabile strumento di supporto alle decisioni". 26 February 1993, place: l'Auditorium Area Ricerca CNR, Via P. Castellino, 111, Napoli Italia, organizers: Consiglio Nazionale delle Ricerche (CNR), Istituto per la Ricerca sui Sistemi Informatici Paralleli - IRSIP; Istituto di Pianificazione e Gestione del Territorio - IPIGET; Università degli Studi di Napoli - Dipartimento di Pianificazione e Scienza del Territorio - DIPIST

[3] Kontoes, C., Wilkinson, G. G., Burrill, A., Goffredo, S., and Mégier, J., 1993. An Experimental System for the integration of GIS Data in Knowledge-Based Image Analysis for Remote Sensing of Agriculture. International Journal of Geographic Information Systems, 1993, vol. 7, no 3, 247-262

[4] Goldberg, M., Goodenough, D., and Shlien, S., 1975. Classification methods and Error Estimation for multispectral Scanner Data. Proc. 3rd. Canadian Symposium on Remote Sensing, September, pp. 125-143

[5] Guo, L. J., and Mc Moore, J., 1991. Post-classification Processing for Thematic Mapping Based on Remotely-Sensed Image Data. Proc. International Geoscience and Remote Sensing Symposium (IGARSS'91), Espoo, Finland, 23-26 May, IEEE Press, Piscataway NJ, pp. 2203-2206

[6] Shoenmakers, R. P. H. M., Wilkinson, G. G., and Schouten, T. E., 1992. Multi-Temporal Image Segmentation on a Distributed Memory parallel Computer. Proc. International Geoscience and Remote Sensing Symposium (IGARSS'92), International Space Year - Space Remote Sensing, Houston, 26-29 May, IEEE Press, Piscataway NJ, Volume 2, pp. 1114-1116

[7] Kanellopoulos, I., Varfis, A., Wilkinson, G. G., and Mégier, J., 1992. Land-Cover Discrimination in SPOT HRV Imagery using an Artificial Neural Network -a 20 class Experiment. International Journal of Remote Sensing, 13, 5, pp. 917-924

[8] Wilkinson, G. G., 1993. The Generalization of Satellite-Derived Raster Thematic Maps for GIS Input. Geo-Information-System, Volume 6, No. 5, October 1993, pp. 24-29.