Application of Goodall's Affinity Index in Remote Sensing Image Classification

A. Altobelli*, E. Feoli *, B. Boglich Perasti*
*Department of Biology, University of Trieste, Italy.

Abstract

The affinity index of Goodall is introduced to integrate remotely sensed data with multisource data (generally called ancillary data) to improve land cover classification. The classification through Goodall's affinity index is performed using "r.affinity", a computer program offered in the GRASS-GIS environment which can deal with quantitative maps (such as image or DTM data), qualitative data (such as land cover maps), ranked data (such as a slope class map). Furthermore this program does not require normal distribution, unlike other popular classifiers used within a GIS context. In this paper "r.affinity" is applied to a set of data from a plateau-area surrounding Trieste and the results are compared with those obtained from maximum likelihood (carried out using GRASS). The highest overall accuracy is achieved by Goodall's affinity index; in particular, while the classified image obtained by the other methods look exceedingly "smoothed" (over-smoothed), the one obtained by "r.affinity" is much more reliable (meaning "near to the real land cover situation") and therefore more suitable and efficient for ecological research.

Introduction

The aim of this work is to test whether the classification of remote sensed images can be improved by integrating remote sensed data with other multisource data.

In particular a comparison is shown between two classifiers: Goodall's affinity index, performed by "r.affinity" and based on non-parametric methods, and the Maximum Likelihood, based on parametric methods.

Goodall's affinity index, proposed for supervised numerical classification (Goodall, 1968; Goodall et al., 1991; Goodall, 1993) is an evolution of the similarity index, applied in the past to phytosociological research in the Quantitative Ecology Division of the Biology Department of the University of Trieste.

This classifier had already been suggested in Remote Sensing Analysis by Altobelli et al. (1995) and Feoli and Zuccarello (1996).

What makes the two classifiers different is the possibility, offered by "r.affinity", to handle qualitative (including binary), ordinal (ranked) and quantitative (ratio and/or interval) data, while the Maximum Likelihood procedure can handle directly only data measured by the ratio or interval scales and it requires normal distribution.

Goodall's affinity index allows the integration of the spectral description of the pixel with descriptions available through other means (e. g. elevation, aspect, inclination, type of soil, etc).

Study area

For the application, the karst plateau was chosen and in particular the area surrounding Trieste.

The land cover situation is here very diversified, being formed by deciduous woods (more or less mature), anthropogenic pinewoods (*Pinus nigra*) and some particular typologies called "dynamic vegetation states", i. e. bushy areas where the land cover appears in standing evolution.

The classification of these particular zones exclusively through spectral signature always presented a wide margin of error and unconformity to the real land cover situation.

Therefore it is useful to find an algorithm which is capable of integrating remotely sensed data with other multisource data, in order to improve the classification.

Methods

The algorithm makes use of probabilistic procedures, since it orders the undifferentiated pixels according to their probability to be similar to the pixels belonging to a recognised cluster (training cluster).

The less similar the observed pixel is among those of the undifferentiated group, the greater will be its affinity to the recognised cluster (the recognised cluster corresponds in this case to the training areas, chosen for the supervised classification).

Since the original Goodall publication on the affinity index (Goodall, 1968) is not readily available, the procedure for its calculation is reported here in its full text (see **Appendix**).

A further example will help understanding.

For each attribute, a value termed norm is recognised as that most representative of the cluster.

In a quantitative attribute (in this example, attribute A) we take the norm as the mean, and all possible values of the attribute are arranged in order of their distance from the norm; values closer to the norm are more similar.

In Table 1 is shown the recognised cluster, of which the norm calculated for quantitative attribute is:

norm = μ = (16+18+18+20+26+25+25+30)/8

In Table 2 we see the pixels belonging to the undifferentiated cluster and the values assumed for the attribute A (quantitative) and attribute B (qualitative); the probabilities for each different value are also reported; for example, value 30 shows the maximum distance (minimal affinity) from the norm.

In Table 3 and 4 are reported the calculated probabilities for attribute A and B.

The probability of 0.3 indicates that it is very unlikely to find a value characterised by a greater distance from the norm. The probability value of 1, on the contrary, indicates that is very likely (it is sure) to find a value further from the norm, since 21 is the nearest value.

The real value of the index for each pixel is calculated through the following formula:

$$\chi^2 = -2 \Sigma \ln p_{ij}$$

which considers the probabilities for every pixel and the values assumed (by the pixel itself) for each attribute. The two calculated χ^2 refer to the first individual of the undifferentiated group (which assumes value 20 for attribute A and 2 for attribute B) and for the last individual of the same group (which assumes value 15 for attribute A and 3 for attribute B).

χ^2 = -2 (ln 0.8 + ln 0.3) = 2.8 for the first individual in the undifferentiated group
χ^2 = -2 (ln 0.5 + ln 0.5) = 1.38 for the last individual in the undifferentiated group

GRASS 4.3 was used to construct a GIS with a Landsat TM (image of 31/07/97) with the digital elevation model (DEM).

The classification by Maximum Likelihood was performed using TM bands 3, 4, 5, 7, considered as quantitative data (Fig. 1).

The classification by Goodall's affinity index (Fig. 2) was carried out using "r.affinity", a program written in C language and available in the GRASS context (http://geog.uni-hannover.de/grass/).

The TM bands 3, 4, 5, 7 and the Fractal Dimension map (calculated starting from the NDVI index) were considered quantitative data; a slope class map was considered as ranked data; a qualitative map indicates the rural and urban zone.

9 classes were identified:

1. pinewood
2. thicket
3. bushy area
4. wood
5. mosaic vegetation
6. urban area
7. karstic grassland
8. grass
9. bare soil

Results and Conclusions

The efficiency of both classifications was tested by checking the overall accuracy by confusion (or error) matrices (Lillesand and Kiefer, 1994).

The highest accuracy was achieved by Goodall's affinity index (overall classification accuracy: 83.40%), while by the Maximum Likelihood method only 68.70% was obtained (Table 5 and Table 6).

The map obtained by Goodall's affinity index is less homogeneous, but much more reliable, near to the real land cover situation and therefore suitable for ecological research.

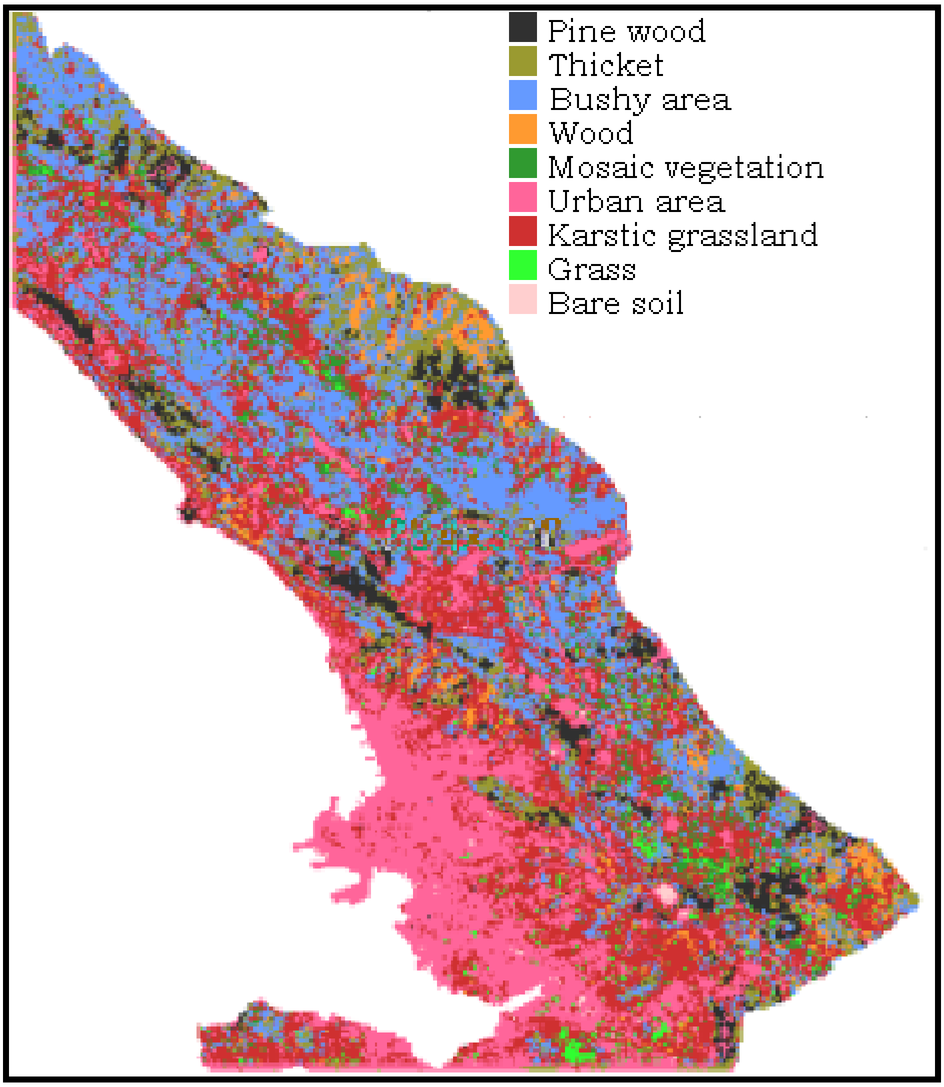

Fig. 1 The map obtained by Maximum Likelihood.

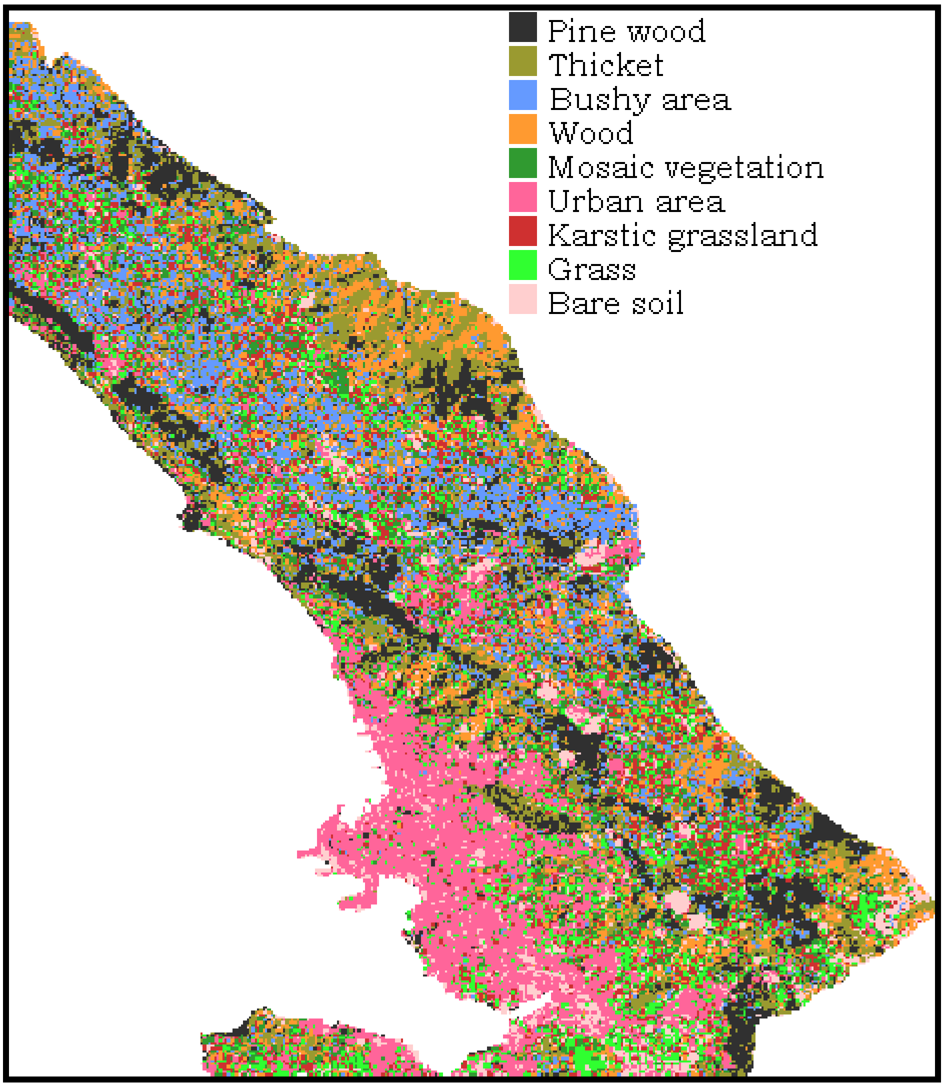

Fig. 2 The map obtained by Goodall's affinity index.

Individuals	1	2	3	4	5	6	7	8
attribute A	16	18	18	20	26	25	25	30
attribute B	1	1	2	3	4	4	3	1

Table 1. An example of recognized cluster.

Individuals	1	2	3	4	5	6	7	8	9	10
attribute A	20	20	15	21	25	30	21	30	30	15
attribute B	2	2	3	1	2	4	1	4	4	3

Table 2. An example of undifferentiated group to be classified.

value	**distance from the norm**	**frequency in the undifferentiated group**	**probability**
30	∣30-22,25∣ = 7.75	3	3/10 = 0.3
15	∣15-22,25∣ = 7.25	2	(2+3)/10 = 0.5
25	∣25-22,25∣ = 2.75	1	(1+2+3)/10 = 0.6
20	∣20-22,25∣ = 2.25	2	(2+1+2+3)/10 = 0.8
21	∣21-22,25∣ = 1.25	2	(2+2+1+2+3)/10 =1

Table 3. Calculated probabilities for attribute A.

value	**frequency in the recognised cluster**	**frequency in the undifferentiated cluster**	**probability**
2	1	3	3/10 = 0.3
3	2	2	(2+3)/10 = 0.5
4	2	3	(3+2+3)/10 = 0.8
1	3	2	(2+3+2+3)/10 =1

Table 4. Calculated probabilities for attribute B.

MAXIMUM LIKELIHOOD

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**user's accuracy**
1	31	.	.	.	.	.	.	.	.	31
2	.	20	10	.	.	.	1	.	.	31
3	.	.	15	1	.	.	.	.	.	16
4	.	2	.	24	.	.	.	.	.	26
5	.	.	.	.	5	.	9	.	.	14
6	.	.	.	.	.	57	.	.	.	57
7	.	.	.	.	11	.	12	.	.	23
8	.	.	.	.	.	.	13	1	.	14
9	.	.	.	.	.	30	.	.	4	34
producer's accuracy	31	22	25	25	16	87	35	1	4	

Table 5. Confusion matrix for the Maximum Likelihood

tot. samples = 246
correctly classified number of pixels = 169
overall classification accuracy = 169/246 = 68.70%

GOODALL'S AFFINITY INDEX

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**user's accuracy**
1	31	.	.	.	.	.	.	.	.	31
2	1	29	1	.	.	.	.	.	.	31
3	1	1	6	8	.	.	.	.	.	16
4	1	7	.	17	1	.	.	.	.	26
5	.	.	.	.	7	.	7	.	.	14
6	.	.	.	.	.	52	.	.	5	57
7	.	.	.	.	7	.	16	.	.	23
8	.	.	.	.	1	.	.	13	.	14
9	.	.	.	.	.	.	.	.	34	34
producer's accuracy	34	37	7	25	16	52	23	13	39	

Table 6. Confusion matrix for Goodall's affinity index

tot. samples = 246
correctly classified number of pixels = 205
overall classification accuracy = 205/246 = 83.40%

References

Altobelli A., Feoli E., Milesi C., & Zuccarello V., 1995. CLID, una procedura di analisi di immagini satellitari basata su metodi non parametrici., AIT VII National Meeting Telerilevamento GIS e Cartografia al Servizio dell'Informazione Territoriale, (A. Zaghi e M.Gomarasca, editori), Chieri (Torino), 255-259.

Feoli E., & Zuccarello V., 1996. Spatial pattern of ecological processes: the role of similarity in GIS applications for landscape analysis, Spatial Analytical Perspectives on GIS, .(M. Fisher, H. Scholten and D. Unwin, editors), Taylor and Francis, London, 175-185.

Goodall D.W., 1964. A probabilistic similarity index, Nature, 203: 1098.

Goodall D.W., 1966. A new similarity index based on probability, Biometrics, 22: 882-907.

Goodall D.W., 1968. Affinity Between an Individual and a Cluster in Numerical Taxonomy, Biometrie Praximetrie, IX, I: 53-55.

Goodall D.W., 1993. Probabilistic indices for classification - Some extensions. Abstracta Botanica, 17: 125-132.

Goodall D.W., Ganis P., Feoli E., 1991. Probabilistic Methods in Classification: A Manual for seven Computer Programs: Computer Assisted Vegetation Analysis (Feoli, E. and Orloci, L., editors), Kluwer Academic Publishers, Dordrecht, 453-467.

Lillesand T. M., & Kiefer R.W., 1994. Remote sensing and image interpretation. John Wiley & Sons, New York, 750 pp.

Appendix

Since the original Goodall's publication on affinity index (Goodall 1968) is not easy to find, the procedure for its calculation is reported here in its full text:
"For each attribute, a value termed norm is recognized as that most representative of the cluster. All possible values of the attribute are then arranged in order of their similarity to the norm; these concepts are defined differently for each category of attribute. The probability of each value is then estimated from the data for individuals outside the cluster, and these are then summed to give the cumulative probability of the given value or any other more similar to the norm. The probabilities for the different attributes are then combined to give the overall probability that an individual whose attribute values were a random sample from those in the undifferentiated group would have as great an affinity for the cluster as the observed individual.
In a qualitative attribute, we take the norm for cluster as the mode, and the similarity of any particular value to the norm as being in the order of the frequency of these values within the cluster. If we write Aij for the affinity for the cluster in respect of attribute *i* (with r_i alternatives) shown by an individual with value *j* for this attribute, and n_{ij} is the number of individuals in the cluster with this value, then

$$n_{ij} < n_{ik} \supset A_{ij} < A_{ik} \qquad (1)$$

and

$$n_{ij} = n_{ik} \supset A_{ij} = A_{ik} \qquad (2)$$

If *mij* is the number of individuals in the undifferentiated group with this value, the probability that a particular individual will have this or a greater degree of affinity is estimated by

$$Pij = \sum_{k\varepsilon S} m_{ik} / \sum_{k=1}^{ri} m_{ik}, \qquad S = \{k{:}A_{ik} \geq A_{ij}\} \qquad (3)$$

For a ranked attribute, the cluster norm may be taken as the median, and if two values differ from the median in the same direction, that closer to the norm is more similar. Two values on opposite sides of the norm are ordered for similarity according to the number of cluster individuals included in the respective “tails”. This follows a rule analogous to that used in the probabilistic similarity and deviant indices already described (Goodall, 1964, 1966).
If the median class has a value *l*, and the number of individuals in the “tail” of the cluster beyond the j^{th} class be represented by t_{ij}, i.e.

$$t_{ij} = \sum_{k=1}^{j} n_{ik} , j \leq l \tag{4}$$

$$t_{ij} = \sum_{k=1}^{r_i} n_{ik} , j > l \tag{5}$$

Then

$$t_{ij} = t_{ik} \supset A_{ij} = A_{ik} \tag{6}$$

$$t_{ij} < t_{ik} \supset A_{ij} < A_{ik} \tag{7}$$

The probability of the observed or a greater degree of affinity is then again estimated by (3) above. For a quantitative attribute, the cluster norm is the mean, values closer to the norm are more similar, and values deviating equally from the norm in opposite directions are ordered on their “tail” frequencies in the cluster. If the cluster mean is μ_i and the individual value is x_{ij}, then

$$| xi_j - \mu_i |<| x_{ik} - \mu_i |\supset Ai_j > A_{ik} \tag{8}$$

$$\left(| xi_j - \mu_i |=| x_{ik} - \mu_i |\right) \wedge \left(t_{ij} < t_{ik}\right) \supset Ai_j < A_{ik}, \tag{9}$$

$$\left(| xi_j - \mu_i |=| x_{ik} - \mu_i |\right) \wedge \left(ti_j = t_{ik}\right) \supset Ai_j = A_{ik}, \tag{10}$$

The probability of a given value x_{ij} or any evidencing greater affinity is then, as before, estimated by the expression (3).

When the p_{ij} have been computed for a attributes, the combined probability p_k (on the assumption that they are independent) is then obtained by computing

as a 2 variable with 2α degrees of freedom, or by the more exact method described elsewhere

$$\chi_2 = -2\sum_{i=1}^{\alpha} \ln p_{ij}$$

(Goodall, 1966)."