Available online at www.sciencedirect.com
SCIENCE DIRECT
Remote Sensing of Environment 99 (2005) 425 – 433
Remote Sensing of Environment
www.elsevier.com/locate/rse

# Optimization of sampling schemes for vegetation mapping using fuzzy classification

R. Tapia [1], A. Stein *, W. Bijker

*International Institute for Aerospace Survey and Earth Sciences (ITC), PO Box 6, 7500 AA Enschede, The Netherlands*

Received 17 November 2004; received in revised form 8 September 2005; accepted 17 September 2005

**Abstract**

This paper considers the design of an optimal sampling scheme for a multivariate fuzzy-*k*-means classifier. Fuzzy classification is applied to delineate vegetation patterns from remote sensing data. The confusion index distinguishes subareas with high uncertainty due to class overlapping from those with low uncertainty. These subareas govern allocation of sample points. A simulated annealing approach minimizes the mean of shortest distances between samples. Optimization was done by prioritizing the survey to areas with high uncertainty. The methodology is tested on a site located in the Amazonian region of Peru. It resulted into an almost equilateral triangular scheme at those parts of the area where uncertainty was highest. The study shows that optimal sampling can be successfully combined with fuzzy classification, using an appropriate weight function.
© 2005 Elsevier Inc. All rights reserved.

*Keywords:* Sampling; Fuzzy-*k*-means; Simulated annealing; Vegetation; Mapping; Amazon forest; Peru

## 1. Introduction

Mapping of natural vegetation, as well as that of other natural resources, is a complex activity. The number of classes, the size of units represented in a two-dimensional map and a precise content description may depend on existing ontologies and subjective choices. To some degree, they depend as well upon the intensity of observations in the field and the desired resolution of the map. Mapping typically considers many variables, whereas the number of possible sampling locations is infinite, from which only a sample of a limited size can be collected. Often, the required sample size to estimate population parameters with an acceptable precision, e.g. using simple random sampling, will be too large to be operationally feasible (Webster & Oliver, 2001). To overcome these problems, model-based sampling, using a known relation, may include spatial dependencies and available observations (de Gruijter, 1999; de Gruijter & ter Braak, 1990). So far model-based sampling has only been carried out for crisp units, whereas in vegetation studies we typically encounter gradual transitions, requiring a fuzzy classification.

Fuzzy classification is a well-established technique to classify multivariate units emerging in various vegetation, soil and forestry studies (Burrough et al., 2000, 2001). Here we focus on the classification being unsupervised, which we consider in this paper to be equivalent to a stratification of the study area. From the start, we do not know the position of the units and their sizes, nor the degree of overlapping between them (Burrough & McDonnell, 1998). To overcome this, we classify thematic images corresponding to vegetation related variables. Such a classification creates spatial units, including their sizes and their positions. A unit is defined as a part of an image with a membership value that is larger for one specific class, surrounded by a fuzzy boundary. All emerging units need to be labeled, and hence to be sampled, at least in principle. The success of sampling relies on an appropriate classification model, and hence on the proper selection of variables for classification. In earlier studies, Corsi et al. (2000) sampled species distribution based on a model for species–environment relationships. Sampling is then carried out to characterize the units, as well as to reduce the uncertainty at the fuzzy borders between these units.

Fuzzy membership functions have been used in the past to determine optimal spacing between samples from soils studies,

* Corresponding author. Tel.: +31 534 874444; fax: +31 317 483554.
*E-mail address:* stein@itc.nl (A. Stein).
[1] Present address: University of Lausanne, Faculty of Geography, Bâtiment Amphipôle, CH-1015, Switzerland.

0034-4257/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2005.09.013

showing the limitation of the lag distance as an optimal parameter (Odeh et al., 1990). van Groenigen and Stein (1998) presented the Spatial Simulated Annealing (SSA) as a method to optimize spatial sampling schemes in two-dimensional mapping. All studies so far considered crisp classes, and none addressed the inherent ambiguity in definition of classes.

The objective of this research is to determine an optimal sampling procedure using fuzzy classification of thematic images obtained by remote sensing techniques. The uncertainty from the classification is used to prioritize subareas for sampling. Simulated annealing procedures ensure an even spread of points. The methodology is applied to analyze vegetation distribution in the Yanachaga-Chemillen national park in Peru, a chain of mountains to the east of the Andes.

## 2. Methods and materials

### 2.1. Gradients models

To apply a model-based sampling, we correlated distribution of vegetation types to environmental conditions, following the (plant) community unit approach. According to Kent and Coker (1992), a plant community is "the collection of plant species growing together in a particular location that show a definite association or affinity with each other." Association implies that certain species occur together under certain environmental conditions more often then could be expected by chance. The variation of species abundance in response to a single environmental factor is called an environmental gradient.

Although scientists mostly agree on the influence of environmental conditions on species abundance and the existence of environmental gradients, views on plant communities differ. Historically, viewpoints ranged from the monoclimatic theory of Clements to the individualistic concept of Gleason (1926). The monoclimatic theory of Clements (Kent & Coker, 1992, referring to Clements, 1926, 1928) assumes clearly recognizable plant communities, which repeat themselves with great regularity in a region. There is little fuzziness at the boundaries between the plant communities. According to the individualistic concept of Gleason, everything is fuzzy, all points are transitions between other points and plant communities only exist as a group of species occurring together at a certain location, but not as combinations of associated species repeated in space. Gleason bases this view on his assumption that all species are distributed as a continuum and have their own individual distribution range for a given environmental gradient (Kent & Coker, 1992, referring to Gleason, 1917, 1926, 1939).

Also Townsend (2000) follows the concept of Gleason (1926) by proposing that patterns in the distribution of individual species may vary widely according to environmental constraints, disturbance history and competitive interactions. Some species, though, may be regularly distributed along those gradients and therefore form identifiable associations. The latter differs from the concept of Gleason and is more in line with the community unit theory, describing the vegetation as a mosaic, based on views of Whittaker (1953) and Whittaker and Levin (1977) as described by Kent and Coker (1992). They describe a climax pattern where repeating combinations of environmental factors and biotic pressures will lead to similar vegetation types. The area of gradual change from one vegetation type to another is called a transitional area or ecotone.

In this study, we will assume that identifiable associations of plant species (community units or vegetation types) exist, which are related to environmental gradients and separated by ecotones. In this case, we define gradients of environmental variables from the combination of data derived from remote sensing. Among these we include elevation, slope and NDVI as a measure of total photosynthetically active matter.

### 2.2. Fuzzy-k-means classification

To classify the area, a fuzzy classification is applied. The motivation for this choice is manifold. First, no fixed objects can be identified, as the concept of plant communities is inherently vague. Therefore no clear, quantitative profiles exist. Second, transition between the units is supposedly gradual. Third, input variables come from different sources, hence requiring an advanced image fusion approach.

A fuzzy classification of an area is done using a fuzzy-*k*-means algorithm applied on pixel values of a Landsat image, elevation and slope. Depending upon the degree of fuzziness, specified by the fuzziness parameter $\phi$, and the number of classes $k$, this procedure yields a set of units, identified by the class with the highest membership value. In this study, considering $N$ data, that will be done on the basis of the maximum partition coefficient $F$

$$F = \frac{F' - 1/k}{1 - 1/k} \tag{1}$$

where

$$F' = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{k} (m_{ic})^2 \tag{2}$$

and the entropy parameter $H$, defined as

$$H = \frac{H' - 1 + F}{\log K - 1 + F} \tag{3}$$

where

$$H' = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{k} m_{ic} \log(m_{ic}) \tag{4}$$

where $m_{ic}$ is the membership value of pixel $i$ to class $c$, $c = 1,\ldots k$ (Burrough & McDonnell, 1998; Burrough et al., 2000). In fuzzy classification, the number of classes $k$ and a fuzziness parameter $\varphi$ have to be chosen (see Section 3.2). Fuzzy-*k*-means classification also yields the confusion index, being equal to the ratio between the first and the second highest membership values. A confusion index quantifies the uncertainty and is likely to be high, for example, in the vicinity of borders between units.

### 2.3. Optimal sampling with simulated annealing

Simulated annealing (SA) (Geman & Geman, 1984; Kirkpatrick et al., 1983) is an optimization method based on an analogy taken from thermodynamics (Michalewicz & Fogel, 2000). A mathematical description of the SA-algorithm is given by the theory of finite Markov chains. Related algorithms have been applied to optimize spatial sampling and to restore degraded images because of its insensitivity to local extremes (van Groenigen et al., 1999).

A central concept in SA is the fitness function $\phi$ that has to be optimized, being a continuous, quantitative function. In Spatial Simulated Annealing (SSA), the fitness function $\phi(S)$ depends upon the sampling scheme $S$ (van Groenigen & Stein, 1998). It is optimized with respect to the coordinates of the sample points. To include fuzziness in optimizing sampling, we focus on the confusion index. The area $A$ considered for sampling is split into a subarea $A_C$ with a low confusion index, and a subarea $A_F$ where the confusion index is high. In both $A_C$ and $A_F$ sampling is optimized, but a careful division of samples is applied to independently optimize sampling to those parts of the area where uncertainty is highest, commonly the border area between units, i.e. and that part of the area where uncertainty is lower, what may be termed the hard units.

In this paper, we take for $\phi(S)$ the minimization of the Mean of Shortest Distances (MMSD) criterion. It aims at regular spreading of all sampling points over the sampling region, i.e. minimizing the expectation of the distance of an arbitrarily point within the region to the nearest sampling point. This leads to the following minimization function:

$$\phi(S) = \int_{A_C} ||\vec{x} - V_C(\vec{x})||\mathrm{d}\vec{x} + \int_{A_F} ||\vec{x} - V_F(\vec{x})||\mathrm{d}\vec{x} \qquad (5)$$

where $\vec{x}$ is a location vector, $V_C(\vec{x})$ is the location vector of the nearest sampling point $\vec{x}_i \in S$, restricted to the subarea $A_C$ and $V_F(\vec{x})$ is the location vector of the nearest sampling point $\vec{x}_i \in S$, restricted to the subarea $A_F$. We aim to determine $S^*$ which minimizes $\phi(S)$. For actual data, the integral is replaced by the average value. Both subareas $A_C$ and $A_F$ are covered with $n_{e_C}$ and $n_{e_F}$ evaluation points, respectively, denoted by $\vec{x}_j e_C \in A_C$ and $\vec{x}_j e_F \in A_F$, located on a fine-meshed grid. A fixed number $n_C$ of observation locations $\vec{x}_1, \ldots, \vec{x}_{n_C}$ is randomly distributed over the area $A_C$, yielding the sampling scheme $S_0^C = \{\vec{x}_1, \ldots \vec{x}_{n_C}\}$. Similarly, $n_F$ observation locations $\vec{x}_1, \ldots, \vec{x}_{n_F}$ are randomly distributed over the area $A_F$, yielding the sampling scheme $S_0^F = \{\vec{x}_1, \ldots \vec{x}_{n_F}\}$. Eq. (5) is estimated by replacing the integral with the finite sum:

$$\hat{\phi}(S) = \frac{1}{n_{e_C}} \sum_{i=1}^{n_{e_C}} ||\vec{x}_i - \vec{x}_j e_C||^2 + \frac{1}{n_{e_F}} \sum_{i=1}^{n_{e_F}} ||\vec{x}_i - \vec{x}_j e_F||^2. \qquad (6)$$

Starting with $S_0^C$ and $S_0^F$, improved sampling schemes $S_1^C$ and $S_1^F$ are obtained, by randomly moving individual points, with changes taken place according to common simulated annealing requirements (van Groenigen et al., 1999). This procedure is repeated until convergence occurs.

A critical issue concerns the sizes of the individual division $n_C$ and $n_F$ of sampling locations to be allocated to the two subareas $A_C$ and $A_F$, respectively. In fact, subarea $A_F$ is a combination of classes on a two by two basis, as two membership values are considered to obtain the confusion index. For $N$ classes on the hard area, we could approximate this by defining at least $N \times (N-1)/2$ new classes or intragrades (not yet classified) based on the combination of the hard classes. This yields a proportion equal to $n_C/n_F = N/N \times (N-1)/2$ of samples to be allocated to the areas $A_C$ and $A_F$, respectively.

### 2.4. Study area

The study site is located at the western border of the Yanachaga-Chemillen national park, near the Santa Barbara Mountains, in the department of Pasco, province of Oxapampa, Peru. There the Santa Cruz River descends on a deep canyon till its junction with the Huancabamba River at an altitude of 700 m. From here flows the Pozuzo River, which joins with other tributaries of the Amazon River basin.

The Park is a protected area of 122,000 ha extension, established in 1986 with the objective to preserve a world record of diversity in flora and fauna. Its isolated location has played an important role from times back to the Pleistocene period (1.8 million years to 10,000 years AC), where the last glaciations took place, eliminating species not adapted to big changes. This chain of mountains near the Amazonian plateau acted as a refuge for old species that are still surviving on micro-climes created by the wide range of elevations that goes from the 400 to 3800 m above sea level.

This range of altitudes has generated differentiated forest ecosystems, including a transition forest (700–2000 m), cloud forest (2000–3400 m) and exposed meadows (>3400 m) (Vasquez & Rojas, 2003). The temperature and height changes promoted events of plants and animal adaptation, leading to forming of new species. The transition forest collects humidity for the Amazonian forest. The cloud forest is important for plants specialized in living from the humidity in the air and includes a vast surface covered by different sorts of mosses. It also holds large varieties of brackens, orchids, cans, bushes and grasses and it provides fodder for the fauna migrating through the ecosystems. Finally, the exposed meadows are the highest ecosystems, harboring species adapted to extreme conditions and many wild relatives of crops.

Although the Park's fauna and flora have been little studied to date, species registered during preliminary surveys included 2584 plant species, many with scientific value for medicine and genetics. The park further includes 85 mammal species, 427 bird species, 16 reptile species, 2 amphibian species, and 31 fish species. Threatened species found within the park include the harpy eagle (*Harpia harpyia*), the jaguar

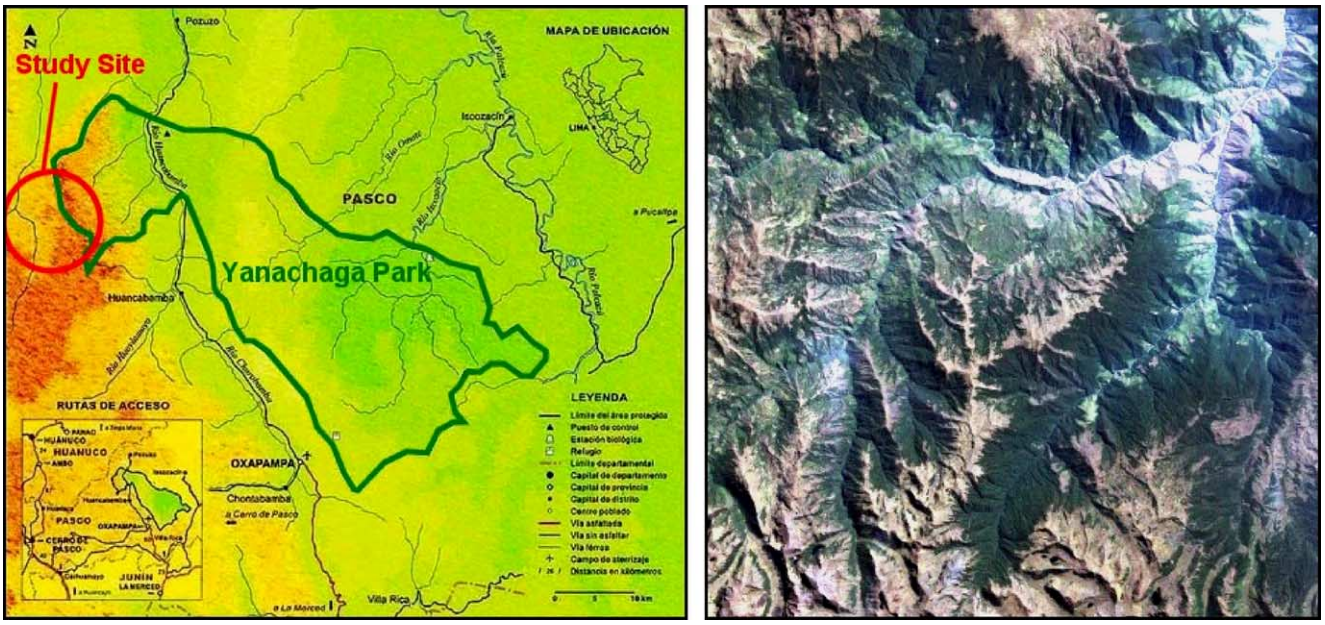


Fig. 1. Location of the study area at the border of the Yanachaga-Chemillen National park (left) and Landsat image for the study area (right).

(*Panthera onca*), and the giant otter (*Pteronura brasiliensis*) (Parks in Peril, 1998). Collection of plants and animals is done for scientific purposes, but it is still difficult to interpret the influence of the environment on species distribution without a proper predictive model and a survey schema associated to it. Remote sensing data could help by providing the necessary inputs to build such a model, in particular elevation data.

This remote area is especially suitable for studies of ecosystems diversity due to its large ranges of elevation and different densities of vegetation cover (Vasquez & Rojas, 2003). Nevertheless, the intensive fog present at any time of the year in the park poses difficulties in finding cloud-free satellite images. This is reflected as well in the national topographic map, on which the park area appears blank. Also, no complete elevation data set of the Park exists in the national topographic charts. For these reasons, a location in the vicinity near the western border was selected for this research and a cloud-free subset of a Landsat image taken on August 5, 1999 was selected (Fig. 1). Elevation data with contour lines of a 50-m interval were available. Although this area might not have the same density of species, it is located within the buffer zone of the park. It comprises altitudes ranging from 1400 to 4000 m a.m.s.l., and shows distinctive land covers. For these reasons, we consider it suitable for modeling species distribution and to test the proposed methodology.

The data available for this site include:

- Landsat TM 5 image, with path-row 7-067, taken on 1999-08-05, georeferenced to UTM coordinates. A subset from the image was selected with a size of $800 \times 800$ rows and columns, covering an area of 576 km$^2$ (Fig. 1).

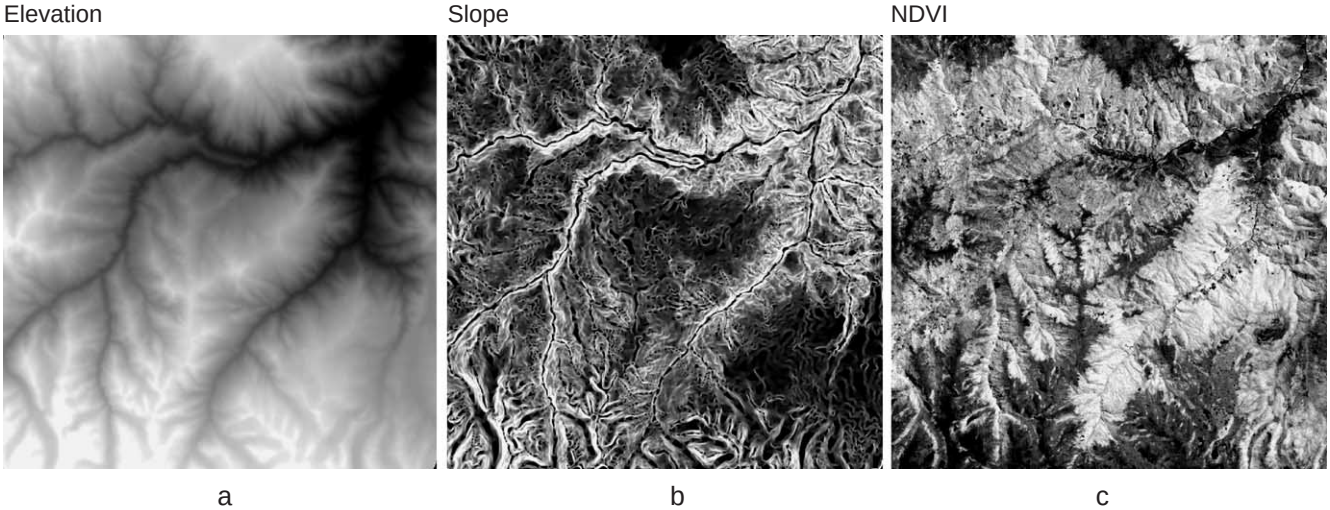


Fig. 2. Thematic maps composing the gradient model: (a) elevation, (b) slope and (c) NDVI.

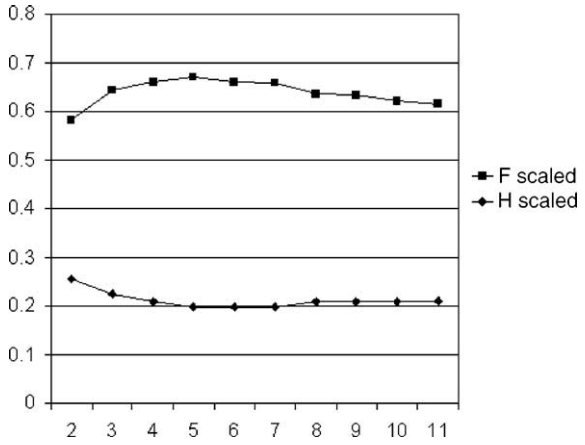


Fig. 3. $H$ and $F$ for $\varphi = 1.5$.

- Contour lines each 50 m digitalized from national topographic maps at the 1:100,000 scale.

## 3. Results

### 3.1. Gradient model

To determine the units for sampling, we classified a multiband image composed of thematic maps corresponding to landform parameters (De Bruin & Stein, 1998) and vegetation indices (Maselli et al., 1998). The range of values for each theme was scaled equally to 8 bit format (a 0–255 quantization) giving the same weight to each theme during classification. Landform parameters *elevation* and *slope* were derived from a DEM with a pixel resolution of 30 m. This DEM was in turn obtained from contour lines digitized from a national topographic map on a scale 1:100,000 with a contour interval of 50 m NDVI served as an indicator for amount of vegetation. It was derived from the Landsat image, keeping a resolution of 30 m (Fig. 2a).

### 3.2. Selection of values for k and φ

Values for the number of classes $k$, ranging from 2 to 11, and the fuzziness parameter $\varphi$ were selected by applying the fuzzy-$k$-means classification to the multiband gradient model. The partition coefficient $F$ (Eq. (1)) and the entropy parameter $H$ (Eq. (3)) were used as comparison parameters for the number of classes $N$ ranging from 2 to 11 (Burrough et al., 2001; van der Wel et al., 1997). Fig. 3 shows values for $F$ and $H$ for a common choice of $\varphi = 1.5$. Clearly, a larger value of $\varphi$ results in less variation in $H$ and $F$.

Fig. 3 shows that the maximum partition coefficient $F$ is reached with the number of classes equal to $k = 5$. If we consider the entropy $H$, the best number of classes (where $H$ is minimal) occurs for $k = 5$ to $k = 7$ classes.

With the results of this analysis and considering that the number of classes corresponding with different vegetation types we expect to identify with the survey is well above 3, we chose a fuzzy classification with $k = 5$ classes. Fig. 4 shows a fuzzy classification considering for $k = 5$ classes. This value was combined with a fuzziness coefficient $\varphi = 1.5$ to obtain a group of transition classes or intragrades. This fuzziness of the transition classes can result from the gradient model and the data upon which it was based, or from the fact that these areas belong to ecotones, transition zones between vegetation types. Fig. 4 displays the confusion index map.

### 3.3. Confusion index map

The confusion index delineates the subareas where classes overlap from subareas that are less ambiguous (Fig. 4b). To avoid a large degree of fragmentation, an image generalization sequence was applied until more integrated areas were obtained. The resolution of the generalized image was $100 \times 100$ rows and columns, corresponding to a pixel size of 240 m (Fig. 5a). By such a generalization, the minimum size of the units to sample will have a support of 240 m, as such avoiding a large number of small polygons, i.e. a large degree of fragmentation. To distinguish between hard and overlapped (fuzzy) areas, the CI image was classified using different thresholds. In Fig. 5b, the threshold value corresponds to the median value for the confusion index image after generalization. The histogram for the CI image (Fig. 6) shows a median

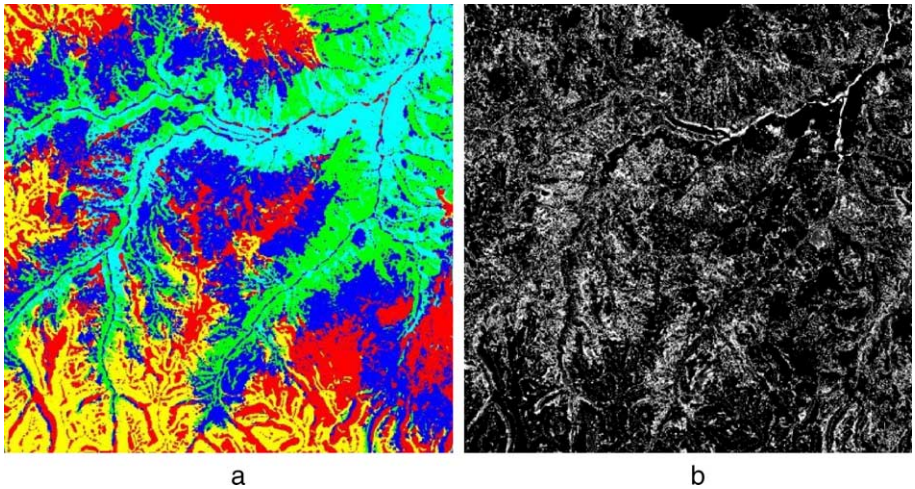


Fig. 4. A fuzzy classification with $\varphi = 1.5$ into $k = 5$ classes (a) and corresponding confusion index (b). Different colors refer to different classes.

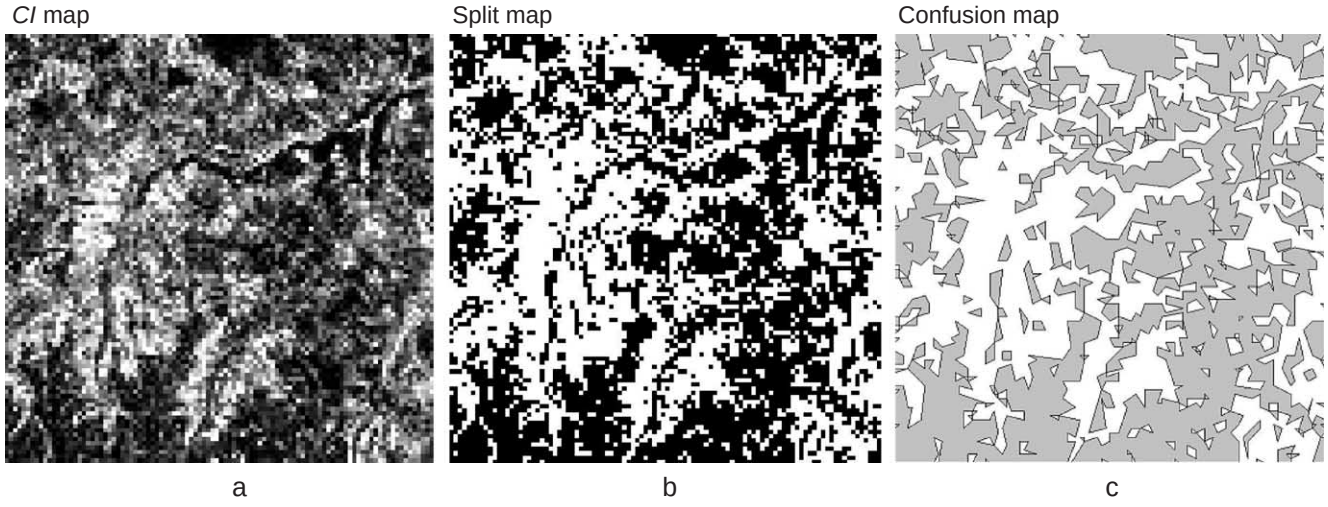


Fig. 5. Confusion index map for $N=5$ and $\varphi=1.5$ (a), split map applying the median CI value as a threshold (b) and map showing white areas showing high confusion ($A_F$), grey areas showing low confusion ($A_C$) (c).

value at 50 in the range from 0 to 192, corresponding to 0.26 in a 0–1 scale. This threshold determines equal areas on the image. Assigning equal areas to the hard and uncertain units avoids adding bias in favor of either of the two groups, whereas prioritization of sampling to the uncertain area is done by assigning a higher number of samples to it.

Next the image was converted from raster to vector format. A polygon map was built using the ILWIS software. Polygons representing uncertain (overlap) classes were assigned an ID=1 label, whereas polygons representing hard classes received an ID=0 label (Fig. 5c). Some resolution is lost in the process from the initial image of CI to the final polygon map, but the generalized view maintains the distribution of units and will facilitate computations to allocate samples.

### 3.4. Sampling scheme

A critical issue concerns the sizes of the individual division $n_C$ and $n_F$ of sampling locations to be allocated to the two subareas $A_C$ and $A_F$, respectively. In fact, subarea $A_F$ supposedly consists of pairwise combinations of classes, as two membership values are considered to obtain the confusion index. For $N$ classes on the full $A_C$, we could approximate the number of points in $A_F$ by defining at least $N\times(N-1)/2$ new classes or intragrades based on the combination of the $N$ classes in $A_C$.

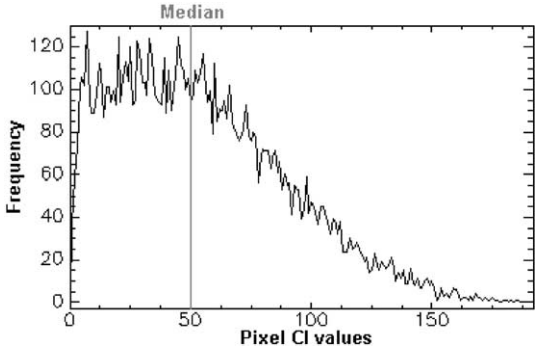


Fig. 6. Histogram of pixel values from the CI map (see Fig. 5a).

This yields a proportion equal to $n_C/n_F=N/N\times(N-1)/2$ of samples to be allocated to the areas $A_C$ and $A_F$, respectively.

Design of the sampling scheme is based on the $N=5$ classes in $A_C$. Hence, a possible overlap of 10 combinations of classes (intragrades) occurs in $A_F$. The number of samples $n_C$ and $n_F$ to be allocated to $A_C$ and $A_F$, respectively, is therefore determined by the $n_C/n_F$ ratio equal to 5:10=1:2. Considering that a preliminary survey on the area will have a budget for a limited number of samples, we fix $n_F$ to be equal to 100, whereas $n_C=50$ samples are assigned to units covering $A_C$ (Fig. 7). The smaller size of the generated polygons indicates the level of disaggregation for areas that are similar in attributes but are spatially discontinuous. In this way, these polygons reflect the accuracy and scale adopted for the study.

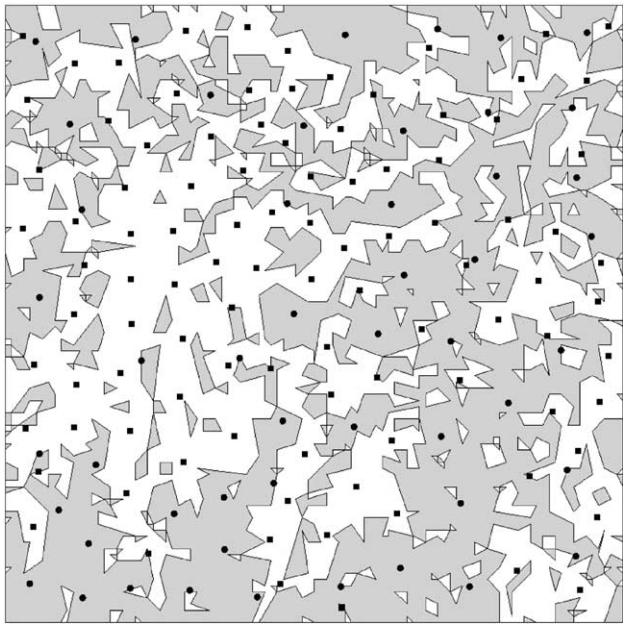

Fig. 7. Location of 150 samples over the study area: 100 samples located in the intragrade subarea $A_F$ (white) and 50 on the classes $A_C$ (grey), using the median of the CI values.

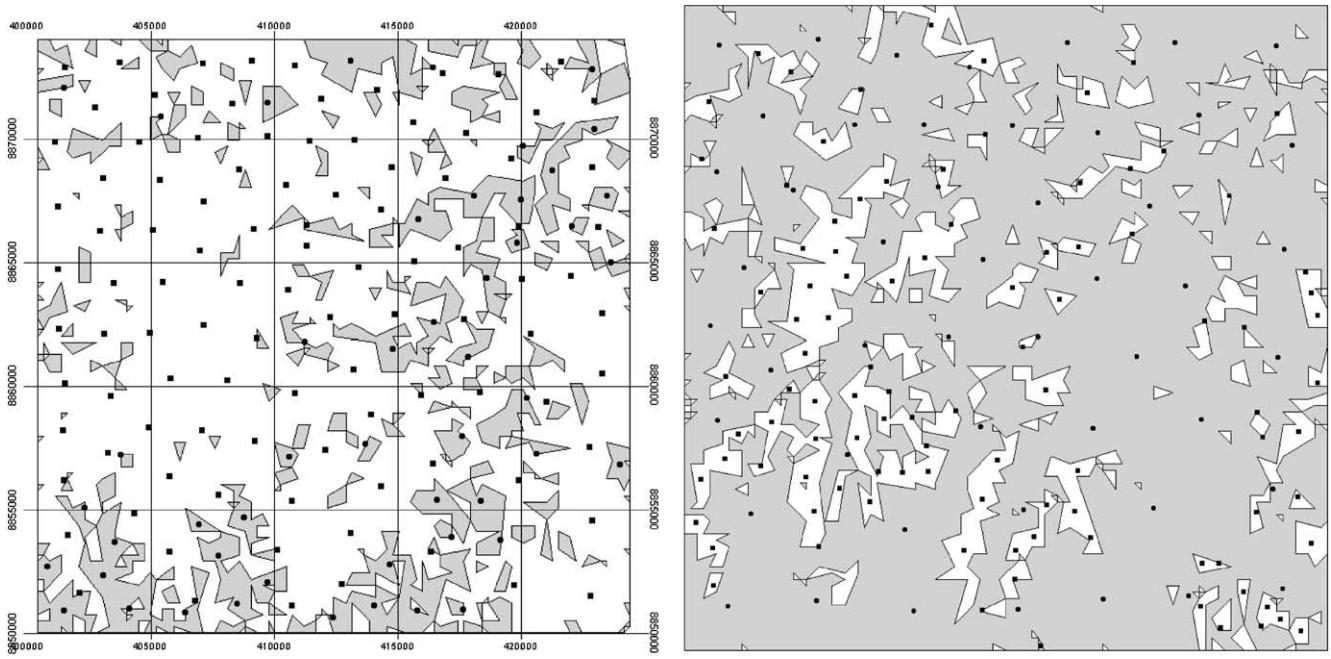


Fig. 8. Location of 150 samples over the study area using CI values equal to the first quartile (left) and the third quartile (right) for ($N=5$, $\phi=1.5$), the white area denoting the intragrade subarea $A_F$.

We repeated the procedure for two different CI thresholds, the first quartile ($q_{25}$) and the third quartile of the CI values ($q_{75}$), using the same proportion between $n_C$ and $n_F$ (Fig. 8). We notice that different thresholds lead to a different distribution of sampling points. In fact, the median CI value ($q_{50}$) corresponds to the optimum, because almost equal areas are assigned to the $A_C$ and $A_F$ subareas.

To make a quantitative comparison, we repeated the whole procedure for 150 randomly distributed data points within the area and compared the mean CI value ($\overline{CI}$) between the optimal scheme $S_o$ and the random scheme $S_r$ within the $A_C$ and $A_F$ subareas. We first notice that for $S_r$ no optimization occurs in terms of covering areas with more confusion since almost the same number of points fall within $A_C$ and $A_F$ (70 and 80, respectively). In terms of the CI, the $q_{50}$ threshold has similar values in both $A_C$ and $A_F$, $S_o$ reduces the $\overline{CI}$ in both $A_C$ and $A_F$. Applying the $q_{25}$ threshold reduces the $\overline{CI}$ in $A_C$ and $A_F$, whereas application of the $q_{75}$ threshold leads to an increase in $\overline{CI}$ when changing from $S_r$ to $S_o$. In $S_o$, we see that application the $q_{75}$ threshold concentrates the points in a subarea with the highest $\overline{CI}$ but at the same time hot-spots in a bigger $A_C$ area are more likely to be missed, while some clustering is also observed. In that sense, the $q_{50}$ threshold provides a balance of parameters with a notable difference between $A_C$ and $A_F$ in terms of the number of points and at the same time an acceptable spatial distribution of points. Another drawback the random schema is its clustering of points that may eventually pose problems for interpolation.

Table 1 summarizes the items to formulate a sampling scheme for the study of vegetation distribution in our area of interest. After optimization, sampling locations were plotted over the confusion map (Table 2).

Table 1
Sampling scheme for the study area

| |
|---|
| Study area: Western border Yanachaga-Chemillen National Park |
| Purpose of sampling: Mapping of vegetation types and vegetation composition based on a topo-land cover model, model validation |
| Predictive Model: Gradient model: NDVI+Slope+Elevation |
| Fuzzy-classification parameters: $N=5$, $\phi=1.5$ |
| Strategy: Prioritize units with high CI |
| Measurements to be performed: Vegetation composition and predicting variables as soil, microclimate and landcover. |

## 4. Discussion

The procedure defined in this paper leads to the possibility to combine fuzzy classification with optimal sampling. Research of van Groenigen and Stein (1998) was related to that of Sacks and Schiller (1988) for optimizing sampling schemes. This paper shows some major progress in order to modify simulated annealing for optimization of spatial sampling. A generation mechanism, the fitness function, and the cooling scheme were decided upon. The use of a variable $c$, which ensures that inferior solutions are accepted with

Table 2
Average digital numbers for a random scheme and the optimized schemes for different CI thresholds

| | $q_{25}$ | | $q_{50}$ | | $q_{75}$ | |
|---|---|---|---|---|---|---|
| | $A_C$ | $A_F$ | $A_C$ | $A_F$ | $A_C$ | $A_F$ |
| *Optimal scheme* | | | | | | |
| Samples | 50 | 100 | 50 | 100 | 50 | 100 |
| $\overline{CI}$ | 18 | 64 | 30 | 75 | 42 | 101 |
| *Random scheme* | | | | | | |
| Samples | 29 | 121 | 70 | 80 | 116 | 34 |
| $\overline{CI}$ | 83 | 74 | 34 | 81 | 74 | 78 |

decreasing probability as the process evolves, is the most important difference with the proposed algorithms proposed by de Gruijter (1999). The Boolean constraint level which distinguishes uncertain areas from crisp areas is losing some information about uncertainty. On the other hand, it is operational for performing sample point allocation. Here we applied the median confusion index threshold to separate the two areas. This, somewhat arbitrary, choice results into areas with almost the same size, since it represents the median value in the confusion index map histogram, and therefore is not giving preference for the samples assigned due to the amount of area to survey but rather to the uncertainty factor. A further improvement may be obtained by applying a more gradual distinction than the Boolean constraint level. Allocation could be proportional to CI values to focus on position and shape of the boundaries, or inversely proportional to CI values to focus on characteristics of the $A_C$ area. We leave these expansions for future work.

The identification of compact areas of uncertainty using the confusion index in the Yanachaga forest required a small overlap and a large number of classes. This could be achieved by using the large variation in elevation and NDVI values. Since NDVI values saturate at a certain level of green biomass and NDVI loses sensitivity to differences in dense forest, it has limitations as an indicator for different forest types. As an alternative, spectral data transformations or radar data could be used to identify vegetation structures. Models could also be improved considering temporal aspects, like the change in NDVI values between different dates of image acquisition (van Straten, 2003).

As was concluded by Lark (1998), smoothing of class memberships (for the case confusion index) is a useful strategy for the formation of spatially coherent regions by multivariate classification. The MMSD procedure was used to allocate samples considering constrained areas of certainty–uncertainty created by the fuzzy-classification. Other constraints like accessibility and risk could also be included to plan sampling of vegetation distribution. If more constraints are added, however, the sample may become less representative for mapping and interpolation.

A further choice has to be made with other sampling procedures. A transect design, for example, may lose precision against equilateral triangular grids obtained with the constrained-MMSD procedure, in favor of an easier, more accessible data collection in the field. Nevertheless, drawback of data collected with transect design could be its restrain to extrapolation and regional analysis. The procedure applied in this study could be further extended and improved. In particular, the delineated units $A_C$ may benefit from a further optimized sampling procedure.

The predictive models used in this study are not based on a precise analysis of vegetation distribution. They were used mainly to test operational procedures for sampling in a fuzzy environment. Therefore, if the procedures were to be applied for vegetation studies, more attention should be given to selecting and modeling the appropriate environmental gradients governing vegetation distribution in that particular region. Another point of attention is the difference between fuzziness or confusion resulting from a transition area or ecotone in the vegetation and uncertainty resulting from the gradient models and the data upon which they are based. For this, a detailed set of field data of plant species and their location is needed.

## 5. Conclusions

Uncertainty of classes resulting from fuzzy-*k*-means classification was expressed by the confusion index. This could be used to indicate constrained areas for sampling in multivariate processes with discontinuous spatial dependency as the regional distribution of vegetation. The confusion index was used to distinguish between a hard area $A_C$, containing relevant units, and an uncertain area $A_F$, representing possible overlapping between units. Such a distinction proved to result into a division of available sampling, where emphasis was given in this study to properly characterize the $A_F$ subarea. As such, this approach is an extension of current optimal sampling procedures with a focus on crisp classes.

Optimization of sampling allocation was achieved using simulated annealing by minimization of the mean shortest distances. This yielded an even distribution of sampling points, reflecting an equidistant grid arrangement. In such a way, a better set of data will become available for interpolation and mapping purposes.

Uncertainty in fuzzy-*k*-means classification depends on the amount of information contained in the model, and hence in the selection of variables, and in choices made for number of classes and fuzziness. Considering a small overlap ($\phi = 1.5$), the optimal number of classes could be shown to be most likely equal to 5.

This study is the first to combine optimal sampling with fuzzy classification, where we focused on increased sampling efforts where uncertainty is largest. Other objectives could be used as well, leading to possibly different sampling schemes.

## References

Burrough, P. A., & McDonnell, R. A. (1998). *Principles of geographical information systems*. Oxford University Press.

Burrough, P. A., van Gaans, P. F. M., & Mac Millan, R. A. (2000). High-resolution landform classification using fuzzy-*k*-means. *Fuzzy Sets and Systems*, *113*, 37–52.

Burrough, P. A., Wilson, J. P., van Gaans, P. F. M., & Hansen, A. J. (2001). Fuzzy *k*-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone area, USA. *Landscape Ecology*, *16*, 523–546.

Clements, F. (1926). Plant succession. An analysis of the development of vegetation. *Publication*, *vol. 242*. Washington, DC: Carnegie Institute, Washington.

Clements, F. (1928). *Plant succession and indicators*. New York: H.W.Wilson.

Corsi, F., De Leeuw, I., & Skidmore, A. K. (2000). *Research techniques in animal ecology* (pp. 389–434). New York: Columbia University Press.

De Bruin, S., & Stein, A. (1998). Soil-landscape modelling using fuzzy-*c*-means clustering of attribute data derived from a digital elevation model, DEM. *Geoderma*, *83*, 17–33.

de Gruijter, J. J. (1999). *Spatial statistics for remote sensing* (pp. 211–242). Kluver Academic Publishers, ch. 13, Spatial sampling schemes for remote sensing.

de Gruijter, J. J., & ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, *22*(4), 407–415.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gleason, H. (1917). The structure and development of plant association. *Bulletin of the Torrey Botanical Club*, *43*, 463–481.

Gleason, H. A. (1926). The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club*, *53*, 7–26.

Gleason, H. (1939). The individualistic concept of the plant association. *American Midland Naturalist*, *21*, 92–110.

Kent, M., & Coker, P. (1992). *Vegetation description and analysis—a practical approach*. New York: John Wiley and Sons.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lark, R. (1998). Forming spatially coherent regions by classification of multivariate data: An example from the analysis of maps of crop yield. *International Journal of Geographical Information Science*, *12*(1), 83–98.

Maselli, F., Gilabert, M. A., & Conese, C. (1998). Integration of high and low resolution NDVI data for monitoring vegetation in Mediterranean environments. *Remote Sensing of Environment*, *63*, 208–218.

Michalewicz, Z., & Fogel, D. B. (2000). *How to solve it: Modern heuristics*. Berlin Heidelberg-Germany: Springer-Verlag.

Odeh, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1990). Design of optimal sample spacings for mapping soil using fuzzy-*k*-means and regionalized variable theory. *Geoderma*, *47*(1–2), 93–122.

Parks in Peril. (1998). Yanachaga Chemillen National Park. http://parksinperil.org/docs forms/yanachagacomp.doc

Sacks, J., & Schiller, S. (1988). *Statistical decision theory and related topics IV, Volume 2* (pp. 385–399). New York: Springer Verlag.

Townsend, P. A. (2000). A quantitative fuzzy approach to assess mapped vegetation classifications for ecological applications. *Remote Sensing of Environment*, *72*, 253–267.

van der Wel, F. J., van der Gaag, L. C., & Gorte, B. G. H. (1997). Visual exploration of uncertainty in remote-sensing classification. *Pergamon*, 335–343.

van Groenigen, J. W., Siderius, W., & Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, *87*, 239–259.

van Groenigen, J. W., & Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, *27*, 1078–1086.

van Straten, O. (2003). Changing woodland ecosystems: Post-disturbance woody species succession dynamics and spatial trends. Master's thesis, ITC.

Vasquez, R., & Rojas, R. (2003). Sinopsis de la caracterizacion de la vegetacion en las areas naturales protegidas de la selva central. http://www.jbmperu.org/yvegetacion.htm

Webster, R., & Oliver, M. A. (2001). Geostatistics for environmental scientists. *Statistics in practice*, (first ed.). John Wiley and Sons, Ltd.

Whittaker, R. (1953). A consideration of climax theory: The climax as a population and pattern. *Ecological Monographs*, *23*, 41–78.

Whittaker, R., & Levin, S. (1977). The role of mosaic phenomena in natural communities. *Theoretical Population Biology*, *12*, 117–139.