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.

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 ϕ, 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 φ 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.