Information Analysis of a Spatial Database for Ecological Land Classification

Frank W. Davis
Department of Geography, University of California, Santa Barbara, CA 93106
Jeff Dozier[*]
Center for Remote Sensing and Environmental Optics, University of California, Santa Barbara, CA 93106

ABSTRACT: An ecological land classification was developed for a complex region in southern California using geographic information system techniques of map overlay and contingency table analysis. Land classes were identified by mutual information analysis of vegetation pattern in relation to other mapped environmental variables. The analysis was weakened by map errors, especially errors in the digital elevation data. Nevertheless, the resulting land classification was ecologically reasonable and performed well when tested with higher quality data from the region.

INTRODUCTION

AN ECOLOGICAL LAND CLASSIFICATION is a system for subdividing land into sub-regions that are relatively homogeneous with respect to one or more ecological variables (Mabbutt, 1968; Rowe and Sheard, 1981). Often the objective is to stratify land surfaces for sampling, resource mapping, habitat assessment, or impact analysis (e.g., Stocker *et al.*, 1977; Strahler, 1981; Carleton *et al.*, 1985). Land classifications are also used in ecosystem modeling over large regions to identify relatively homogeneous sub-regions within which model parameters can be estimated more precisely (e.g., Band and Wood 1988).

There are many methods of land classification ranging from reconnaissance and qualitative survey to quantitative analysis of site data from plots or transects scattered over a study region. Maps of discriminant terrain variables such as vegetation, topography, and soils are often used to predict the distribution of land classes in unsampled areas. Classification systems have also been developed and refined by analyzing correspondence between maps of terrain variables (e.g., Phipps, 1981; Bailey, 1983; Forman and Godron, 1986).

The development of ecological land classifications is increasingly supported by digital satellite and terrain data that can be used to map land classes over large areas (e.g., Morissey and Strong, 1986; Franklin, 1987), and by geographic information systems (GIS) for digital map overlay and spatial modeling (Burrough, 1986; Berry, 1987). A potential advantage of GIS-based mapping of ecological land classes is that maps of terrain variables can readily be weighted and combined to display new or refined classification systems. Such flexibility is important because no single land classification is optimal for all ecological applications, especially when those applications span a range of spatial and temporal scales and include widely divergent purposes (Rowe and Sheard, 1981).

Many quantitative methods have been applied to land classification (e.g., Legendre and Legendre, 1983). A review of these approaches is beyond the scope of this paper. Our purpose here is to demonstrate the usefulness of one method that we have found to be especially well suited to GIS-based cartographic modeling (Berry, 1987) and land classification. This method, which we refer to as mutual information analysis (Michaelsen *et al.*, 1986), was introduced by Phipps (1981) under the name PEGASE (Partition d'un Ensemble Géographique pour l'Analyse Spatiale Ecologique) as a means of determining ecological relationships among a set of overlaid chloropleth maps. The method is hierarchical, divisive, and predictive in the sense that land classes are developed by combining maps to produce a pattern that most closely resembles that of a dependent mapped variable.

We have applied mutual information analysis to classify terrain in a heterogeneous region of southern California based on spatial correspondence between digital maps of natural vegetation and environmental variables including geology, elevation, slope, seasonal insolation, and hillslope position. Maps of these variables were sampled to develop the classification. To measure the impact of map errors on measured association between ecological variables, we tested the predictive capability of the map-based land classification system using a smaller test data set acquired from aerial photographs, topographic maps, and site visits.

METHODS

MUTUAL INFORMATION ANALYSIS

Mutual information analysis is a method for grouping samples that share a set of attributes, based on the association of those attributes with a categorical dependent variable. For example, the technique could be used to classify species habitats defined by vegetation type and elevation zone based on the association of these variables with species sighting data. Similarly, vegetation environments defined by geology, soil type, and slope class could be identified based on the co-occurrence of these terrain variables with mapped vegetation classes. The fact that map elements are classifed with reference to a specific decision variable is consistent with the purposive nature of classification. The use of categorical variables facilitates the analysis of thematic maps (e.g., geology, land cover, soil type), but also means that continuous variables such as elevation must be divided into classes. The divisive hierarchical classification structure is consistent with the notion of scale-dependence and interaction of ecological factors underpinning ecological patterns (e.g., Mabbutt, 1968; Gauch and Whittaker, 1981; Naveh and Lieberman, 1984).

Conceptually, the method presumes that land surfaces are spatially ordered due to ecological interdependence among terrain variables. For example, geology, topography, and vegetation are often closely coupled in natural landscapes, recurring in a limited number of combinations each possessing characteristic ecosystem properties of energy and mass transfer (e.g., Jenny, 1980; see Phipps (1981) for fuller treatment of the theoretical basis for the method). To apply the method, each terrain variable is treated as a mosaic (i.e., chloropleth map) with no

[*]Also affiliated with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109.

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 56, No. 5, May 1990, pp. 605-613

0099-1112/90/5605-605$02.25/0

gradation between mosaic elements or patches. This representation will obviously be more or less artificial depending on the study region and scale of analysis. The complexity of a mosaic is a function of both the number of patches and the number of patch types. A measure of mosaic complexity is the entropy statistic

$$H = -\sum_{j=1}^{u} p_j \ln p_j \tag{1}$$

where p_j is the proportion of the mosaic in patch type j, $j = 1,\ldots u$. When an area is jointly classified by two ecological variables, say x and y, unless x and y are perfectly correlated, a more complex composite mosaic will result whose joint entropy is

$$H(x,y) = -\sum_{j=1}^{u}\sum_{k=1}^{v} p_{jk} \ln p_{jk} \tag{2}$$

where p_{jk} is the proportion of the mosaic where x and y are in states j and k, respectively.

The joint entropy statistic is maximized when x and y are independent, that is, when they exhibit no spatial correspondence. The joint entropy observed when a composite mosaic is formed from overlaying maps of two ecological variables is generally less than this theoretical maximum because the state of one variable is constrained by the state of another (e.g., vegetation types confined to specific soil types or elevational zones). The difference between the maximum and observed entropy serves as a measure of the degree of correspondence between the two variables. This difference has also been called the mutual information (i.e., mutual entropy) shared between variables x and y (Orlóci, 1978).

The mutual information between x and y may be estimated by analyzing a contingency table of jointly classified observations of both variables taken at random over the study region. Spatial attributes such as patch size, shape, and neighborhood are generally not considered (although in principle these attributes could be encoded by location). For a large sample size N, the mutual information can be estimated from cell frequencies f_{jk} (the number of samples where $x=j$ and $y=k$), marginal frequencies $f_{j\cdot}$ (the number of samples where $x=j$) and $f_{\cdot k}$ (the number of samples where $y=k$) as

$$I(x,y) = N \ln N + \sum_{j=1}^{u}\sum_{k=1}^{v} f_{jk} \ln f_{jk} - \sum_{j=1}^{u} f_{j\cdot} \ln f_{j\cdot} - \sum_{k=1}^{v} f_{\cdot k} \ln f_{\cdot k}. \tag{3}$$

If x and y are independent of each other, $I(x,y) = 0$. If they are perfectly associated, $I(x,y)$ equals the product of the sample size and the entropy of x or y (because $H(x)=H(y)=H(x,y)$). $2I$ is approximately χ-squared distributed with $(u\text{-}1)(v\text{-}1)$ degrees of freedom (Kullback 1959).

To develop a land classification for a dependent variable y (e.g., vegetation), the mutual information is measured between y and terrain variables $x_m, m=1,2,\ldots$(e.g., geology, elevation zone). Samples are stratified based on the variable in x that has the highest mutual information with y, and the analysis is repeated within each stratum using the remaining variables. This divisive hierarchical procedure allows certain terrain variables to take precedence in conditioning the response of y to the remaining variables.

The method produces sample strata that are more homogeneous with respect to y. By stratifying a sample based on the variable with the highest mutual information with y, one maximizes the reduction from the sample entropy ($H(y)$) to the total entropy remaining in the sample strata. The total entropy at a level of stratification T is

$$H_T(y) = \frac{1}{N}\sum_{s=1}^{S} f_s H_s(y) \tag{4}$$

where N is the number of observations, S is the number of strata at level T, and f_s is the number of observations in stratum s. Thus, $H_T(y)$ is the sum of subset entropy measures weighted by stratum size.

$H_T(y)$ can be compared to the starting sample entropy as a measure the redundancy between the stratification based on variables in x and the dependent variable y, i.e.,

$$r_T = 1 - H_T(y/H(y). \tag{5}$$

In other words, r_T measures the fraction of the information in y that is accounted for by the stratification.

If x is a continuous variable, it can be converted to a binary variable by identifying the cutpoint at which $I(x,y)$ is maximized. Because $2I$ is approximately χ-squared distributed, a convenient stopping rule is to stop splitting the data at the point where $2I$ does not exceed some chosen level of significance. Another approach is to develop a very large classification tree and then prune it back by removing classes that do not improve the skill of the classification in predicting the state of new or subsampled observations (Michaelsen *et al.*, 1986). In the example that follows we have retained classes based on a χ-squared significance of 0.05.

Study Area

We applied mutual information analysis to map ecological land units in a 73 km^2 region of northern Santa Barbara County, California (Figure 1). A brief description of the area is provided here. More detailed descriptions of the area can be found in Davis *et al.* (1988) and Ferren *et al.* (1984).

The study region encompasses two distinctive physiographic regions, the Burton Mesa and the Purisima Hills (latitude 34° 42′ N, longitude 120° 27′ W). The local climate is Mediterranean, having a strong maritime influence, cool summers, and mild winters. Over 90 percent of the 36 cm average annual precipitation falls between November and April.

The Burton Mesa is a marine terrace covered with Orcutt sandstone (Figure 2), 0.5 to 40 m of weakly cemented quaternary aeolian sands (Dibblee, 1950). Level uplands from 100 to 120 m above sea level are dissected by streams that have formed wide valleys with short steep slopes. Soils range from deep excessively drained sand to poorly drained shallow sand overlying iron or clay pans (Shipman, 1972). Most of the valley bottoms are filled with quaternary alluvium that is developed or under cultivation. Upland vegetation is mostly "sandhill chaparral," composed of shrubby, multi-stemmed coast live oaks (*Quercus agrifolia*) scattered among evergreen chaparral shrubs including chamise (*Adenostoma fasciculatum*), ceanothus (*Ceanothus ramulosus*), and manzanita (*Arctostaphylos* spp.) (Plate 1). Many upland areas have burned recently, or have been grazed or cleared, so that the sandhill chaparral is now one element in a regional mosaic of vegetation types that also includes introduced grassland, coastal scrub, chaparral, and oak woodland (Ferren *et al.*, 1984). Coast live oak forest occurs near streams and steep north-facing slopes. Southern exposures are dominated by coastal scrub or chaparral species.

The Purisima Hills border the Burton Mesa to the north and are a northwest-southeast trending anticline of late Tertiary marine sedimentary rocks (Figure 2). Widespread surficial geological formations in the study area include the Paso Robles conglomerate, the Careaga sandstone, and the Sisquoc diatomite and diatomaceous shale. The Upper Careagea (Graciosa member) is a loose medium- to coarse-grained sandstone and the Lower

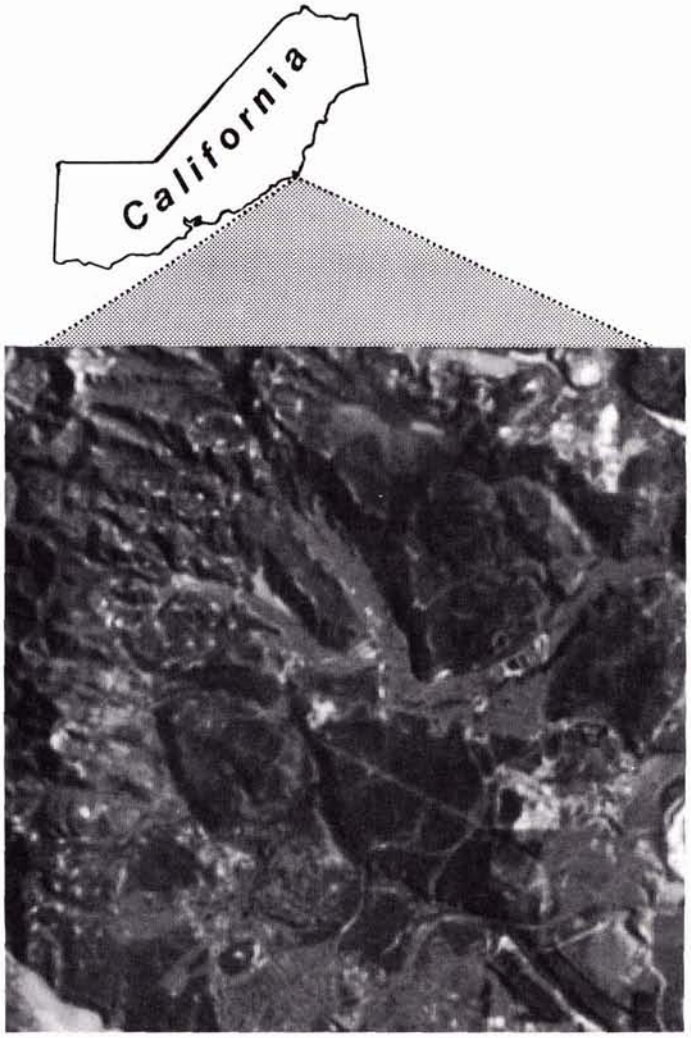

FIG. 1. Location map of the study region with inset showing Thematic Mapper image (December, 1985) of the study area. Image is in UTM projection, oriented north, 9 by 8.1 km. Dots in the upper right hand portion of the image are oil platforms in the Purisima Hills. Bright areas on the Burton Mesa (lower left portion of image) are residential and agricultural areas.

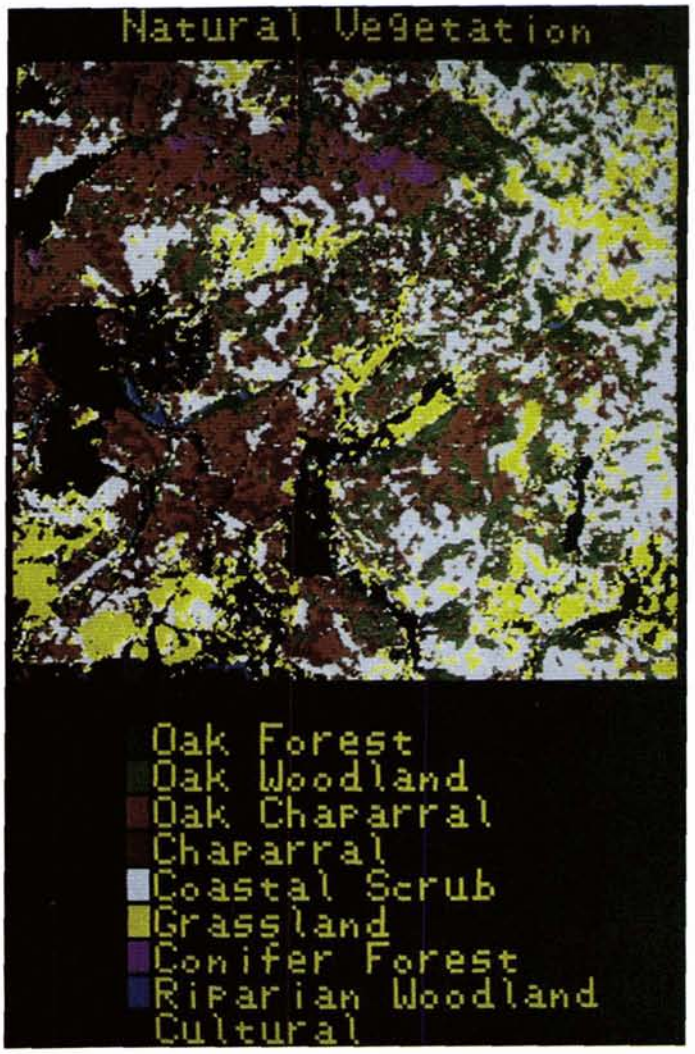

PLATE 1. Vegetation classification for the study area based on Thematic Mapper Simulator Data from July, 1985. Image covers the same area as Figures 1 and 2.

Careaga (Cebada member) is massive fine-grained sandstone and siltstone. Elevations range from 225 to 450 m, and the topography consists of rolling hills with steep slopes and narrow valley bottoms. Except in valleys underlain by quaternary alluvial deposits, the soils are shallow and rocky over most of the Purisima Hills.

Most valley floors in the Purisima Hills are cultivated and many slopes are actively grazed. Small oil and gas wells are the only activity over the remaining area, where predominant natural vegetation types include coast live oak forest, coast live oak woodland and savanna, Bishop pine (*Pinus muricata*) forest, chaparral, coastal scrub, and grassland (Plate 1). Vegetation here is strongly associated with geology and topography (Wells, 1962; Cole, 1980).

DATABASE CONSTUCTION

Our objective was to identify and map vegetation environments in the study region based on the association of mapped vegetation pattern and physical terrain variables including geology, elevation, slope angle, slope azimuth, clear-sky insolation, and position in a drainage basin. The association of these terrain variables with vegetation in southern California is well documented (e.g. Wells, 1962; Harrison *et al.*, 1971; Steward and Webber, 1981). Although soil maps exist for the study area, we did not use them because the maps for the Purisima Hills are much less detailed than the geologic map and have less predictive value.

Our approach to land classification presumed that actual vegetation cover was a reliable indicator of ecological conditions at a site. We did not account for historical burning, cutting, and grazing, although these exert a strong and persistent influence on vegetation pattern and weaken the association between vegetation and physical site variables (Wells, 1962; Davis *et al.*, 1988).

Vegetation was mapped following the classification system of Paysen *et al.* (1980). This system defines vegetation formations and sub-formations based on physiognomy and vegetation series based on dominant overstory species (Table 1). The classification is well suited to remote sensing applications and to semi-arid shrublands and woodlands, which are frequently dominated by one to several overstory species. We departed from the classification system slightly in distinguishing coast live oak chaparral from chaparral for chaparral vegetation that included

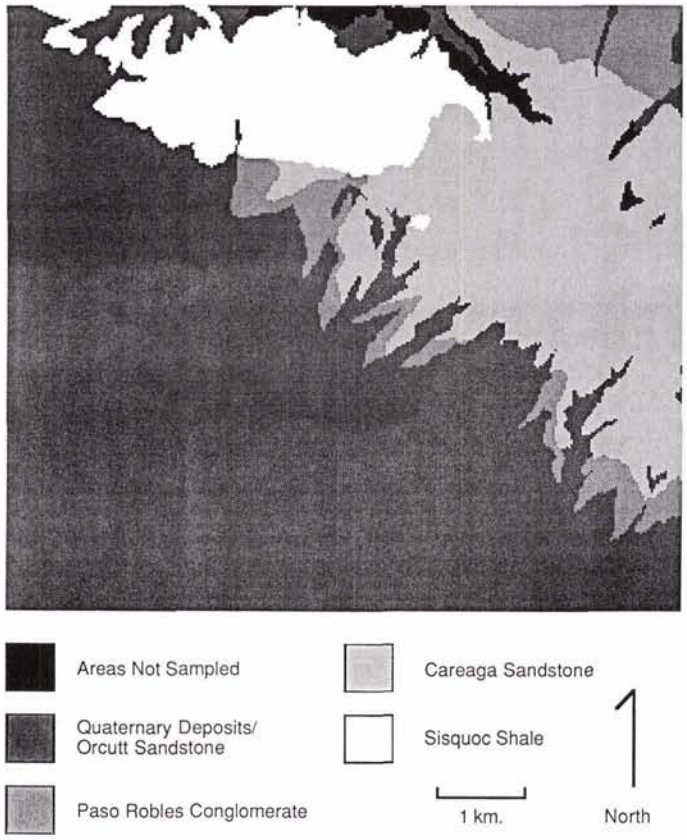

FIG. 2. Surficial geology of the study area (after Dibblee (1950)).

TABLE 1. VEGETATION CLASSIFICATION SYSTEM FOR THE STUDY REGION (NOT INCLUSIVE). MAP ACCURACY FOR EACH CLASS IS THE PROPORTION OF SAMPLES CLASSIFIED CORRECTLY IN THE TMS-DERIVED CLASSIFICATION SHOWN IN FIGURE 3, BASED ON 141 TEST SITES (SEE DAVIS (1987) AND DAVIS ET AL. (1986) FOR DETAILS).

Class	% Oak Cover	Dominant Species	May Accuracy (%)
Coast live oak forest	>60	Quercus agrifolia Toxicodendron diversilobum	79
Coast live oak woodland	20-60	Quercus argrifolia Adenostoma fasciculatum Arctostaphylos spp.	86
Coast live oak chaparral	10-20	Quercus agrifolia Adenostoma fasciculatum Arctostaphylos spp.	89
Chaparral	0-20	Adenostoma fasciculatum Ceanothus ramulosus C. impressus Arctostaphylos rudis A. purisima	88
Coastal Scrub	0-20	Salvia mellifera Baccharis pilularis Ericameria ericoides Artemisia californica	86
Grassland	0-20	Bromus spp. Vulpia spp. Avena barbata Brassica spp.	89
Conifer Forest	0-30	Pinus muricata Quercus agrifolia Heteromeles arbutifolia	92

10 to 20 percent oak cover (Table 1). This distinction was made because our studies in this area have focused on the distribution of this oak species (Davis, 1987; Davis *et al.*, 1988).

Natural vegetation was mapped using Thematic Mapper simulator data (resampled to 30 m resolution) collected at noon on 5 July 1984 (Plate 1). The first four principal components of the imagery (excluding TM band 6) were subjected to unsupervised classification, and image classes were assigned to vegetation types by analysts familiar with the study area (Davis, 1987). Weighted map accuracy was 89 percent, and all natural vegetation classes were mapped at greater than 85 percent accuracy except for oak forest, which was mapped at 79 percent accuracy (Table 1; Davis *et al.*, 1986; Davis, 1987). Most errors in mapping oak forest were confusion with dense oak woodland, and thus not severe. The vegetation map was co-registered in Universal Transverse Mercator projection to the geologic map by Dibblee (1950).

Topographic parameters slope angle, slope azimuth, and local horizons were derived from the U.S. Geological (USGS) 30-m digital elevation model (DEM) for the Lompoc quadrangle using image processing software described by Frew and Dozier (1986).

As a measure of drainage basin position, we calculated each cell's drainage area from the number of cells upstream expected to drain through that cell based on maps of slope and exposure. The procedure for calculating drainage area was a modification of the algorithm described by Marks *et al.* (1984) for delineating drainage basin boundaries.

We measured the mutual information of vegetation pattern and maps of clear-sky insolation integrated over individual months of the growing season (December through June) to compare monthly and seasonal patterns of association. Incident radiation on a slope was calculated using maps of slope angle and azimuth as well as a horizon file which gave, for each cell in the elevation model, the angle to the local horizon in eight azimuthal sectors (Dozier *et al.*, 1981). To calculate diffuse irradiance, a sky view factor, the ratio of diffuse sky radiance at a point to that on an unobstructed horizontal surface, was calculated from slope angle, azimuth, and horizon information under the approximation that diffuse irradiance was isotropic. For each point, reflected radiation from surrounding terrain was also estimated based on the difference between the sky view factor on an infinitely long slope (i.e., no facing terrain) and the calculated sky view factor at the point. Average surface albedo for the region was estimated as 0.14, a representative number for chaparral vegetation that covered a majority of slopes in the area (Miller *et al.*, 1981)

The range in elevations was small enough that the atmosphere was treated as the same at all locations. Monthly values for atmospheric transmission were estimated based on visibility data collected at the Vandenburg Air Force Base. Because we could only roughly estimate atmospheric properties that prevail in a particular month, the calculated insolation values were treated as relative rather than absolute and scaled to one-byte integers between 0 and 255 (e.g., Figure 3).

Goetz (1987) conducted third-order field surveys on two dense grids to study errors in the DEM model for the Lompoc quadrangle. Digital and survey elevations agreed fairly well ($r^2=0.93$), but errors were amplified during the differencing operations needed to calculate slope and exposure (slope $r^2 = 0.44$, exposure $r^2 = 0.38$). Errors were concentrated in areas of rapid change in slope and exposure such as ridges and ravines, and included both resolution errors (i.e., undersampling in areas of rapid change) and stereo-model errors (e.g., overestimating surface elevation in riparian corridors filled with continuous tree canopies). Accordingly, prior to sampling and classification we masked ridges and ravines, where we expected digital

topographic parameters to have lowest reliability (and registration errors between scanner data and topographic data to be most problematic). This was done by computing the gradient (first derivative) of the March insolation image and eliminating locations where the rate of change exceeded a subjectively chosen threshold.

Each layer of the database was comprised of 81,000 900 m^2 cells. However, the variables were all highly spatially autocorrelated (Cliff and Ord, 1981). Analysis of the full set of observations would have weakened the use of the chi-square stopping rule, because the spatial dependence in mapped variables violated the assumption of sample independence. To avoid this problem, we sub-sampled the maps at a sampling density low enough so that sample values were expected to be independent at the average inter-sample distance. We analyzed a 3.5 percent random sample of the image ($n = 2853$), which was sufficiently low to remove most sample interdependence, based on semi-variogram analyses of the topographic variables (Oliver and Webster, 1986).

Mutual information analysis was conducted using software developed at UCSB. We converted continuous topographic variables to binary variables using cutpoints where the mutual information between vegetation and each topographic variable was maximized. We retained all classes that were significant at χ^2 $p<0.05$.

We tested the predicive value of the classification using 300 samples identified by interpreting 1983 1:24,000-scale air photos (extensive ground reconnaissance in the study area confirmed the reliability of identifying vegetation from the photographs). The study region was stratified into six subregions within which 40 to 60 vegetation stands were sampled which were at least 60 by 60 metres in area on uniform geology (far from mapped boundaries) and topography. Sample neighborhood was randomly selected, but sample locations were sometimes adjusted by one or two pixels to meet the criteria of uniform vegetation and site conditions. The vegetation class was identified by photointerpretation, and the geologic and topographic parameters were taken from the corresponding addresses in the database (values were checked against air photos, and topographic and geologic maps to confirm their reasonableness). It was not possible to collect ground observations of radiation and drainage basin position. However, it is reasonable to assume that this test sample of 300 sites was much more accurate than 300 samples taken at random from the database.

RESULTS

Nineteen land classes accounting for 18.5 percent of the information in the vegetation data were identified using the stopping criterion of χ-square probability less than 0.05 (Figure 4 and Plate 2, Table 2). Figure 5 shows relative vegetation proportions in seven of the 19 land classes that had relatively low similarity among them and compared to overall vegetation composition of the study area. Some other classes did not differ very much in vegetation composition, for two different reasons. For some divisions, the very large sample size meant that even slight differences in vegetation were significant (e.g., Table 2, classes 2, 3, and 4). This is because the magnitude of the information statistic is proportional to sample size (Michaelsen *et al.*, 1986). In other instances, the vegetation in classes on different branches of the classification (e.g., on different rock types) was quite similar (e.g., Table 2, classes 4 and 9, 6 and 8). This could occur when different combinations of substrate and topography provided similar site conditions for plant establishment and growth. Thus, the classification could have been simplified by merging similar classes, for example, by merging classes 15 and 16, classes 6,8, and 10, and classes 2,3,4,9, and 11. We retained all 19 classes for the analyses reported here in order to provide consistency in evaluating redundancy statistics and the test data.

Geology and calculated insolation were the most important variables in the classification. The first stratification by surface geology was highly significant. Most importantly, nearly all bishop pine forest was mapped on diatomaceous shale (Classes 15 through 18 in Table 2). The importance of geology in controlling vegetation pattern can be seen in the plot of r_T against the number of classes (Figure 6). This function shows a large

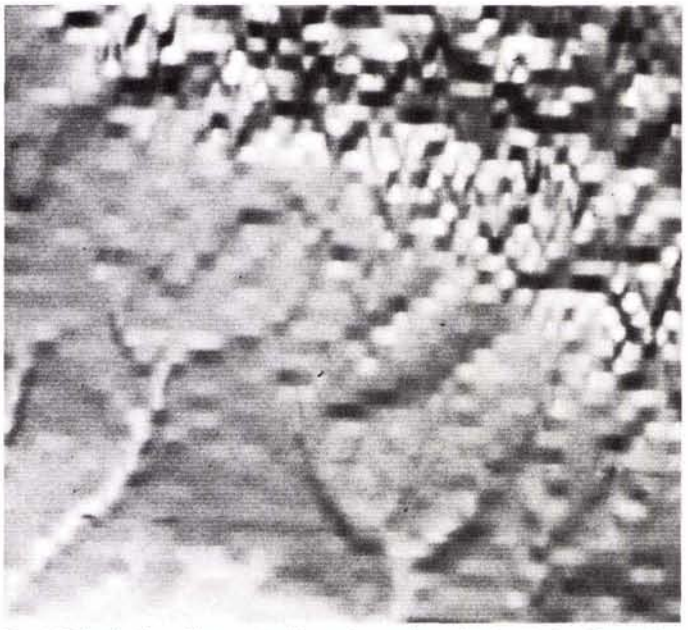

FIG. 3. Distribution of integrated incoming solar radiation during December calculated from digital elevation data. Scale and orientation as in Figure 2. Image brightness is proportional to total radiation.

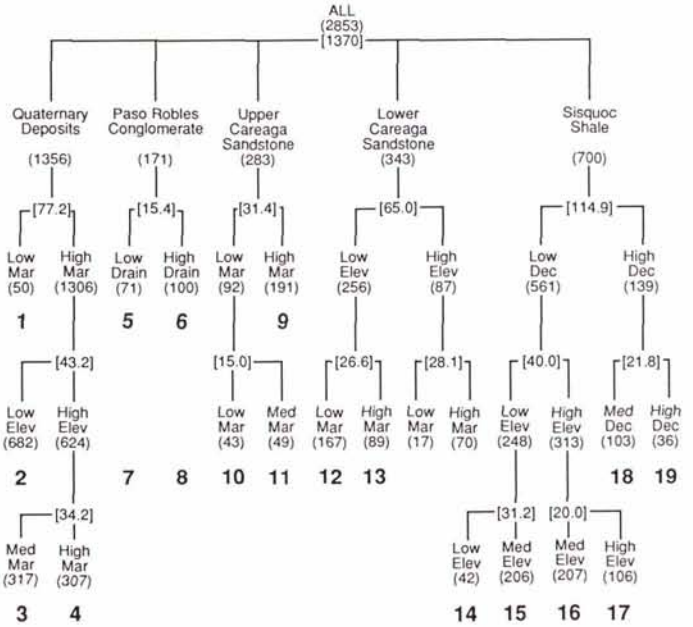

FIG. 4. Ecological land classification based on mutual information analysis of 2853 samples from the regional database. Bold numbers identify land classes (*cf.* Table 2), numbers in parentheses are the sample size for each class, numbers in brackets are two times the mutual information statistic (I) (Equation 3) between vegetation and the stratifying variable.

TABLE 2. VEGETATION COMPOSITION (CLASS PROPORTIONS AND MARGIN TOTALS) OF THE 19 LAND CLASSES IDENTIFIED BY MUTUAL INFORMATION ANALYSIS. LAND CLASS NUMBERS CORRESPOND TO THOSE IN FIGURE 2.

	Vegetation Class							
Land Class	Oak Forest	Oak Woodland	Oak Chaparral	Chaparral	Coastal Scrub	Grassland	Conifer Forest	Total
1	0.600	0.040	0.160	0.140	0.060	0.0	0.0	50
2	0.110	0.139	0.180	0.229	0.180	0.161	0.0	682
3	0.136	0.300	0.215	0.132	0.110	0.107	0.0	317
4	0.088	0.238	0.166	0.235	0.199	0.075	0.0	307
5	0.141	0.155	0.085	0.141	0.197	0.282	0.0	71
6	0.310	0.200	0.140	0.020	0.150	0.180	0.0	100
7	0.372	0.256	0.0	0.116	0.186	0.070	0.0	43
8	0.224	0.204	0.224	0.041	0.224	0.082	0.0	49
9	0.068	0.209	0.126	0.236	0.246	0.115	0.0	191
10	0.365	0.168	0.132	0.066	0.144	0.126	0.0	167
11	0.191	0.146	0.124	0.236	0.247	0.056	0.0	89
12	0.412	0.118	0.059	0.0	0.059	0.294	0.059	17
13	0.0	0.0	0.114	0.057	0.329	0.429	0.071	70
14	0.476	0.167	0.190	0.071	0.0	0.048	0.048	42
15	0.141	0.097	0.078	0.107	0.005	0.0	0.573	206
16	0.082	0.048	0.092	0.043	0.024	0.005	0.705	207
17	0.0	0.038	0.160	0.028	0.038	0.047	0.689	106
18	0.049	0.107	0.223	0.223	0.078	0.0	0.320	103
19	0.028	0.111	0.056	0.472	0.306	0.028	0.0	36
Total	413	456	432	454	416	304	378	2853

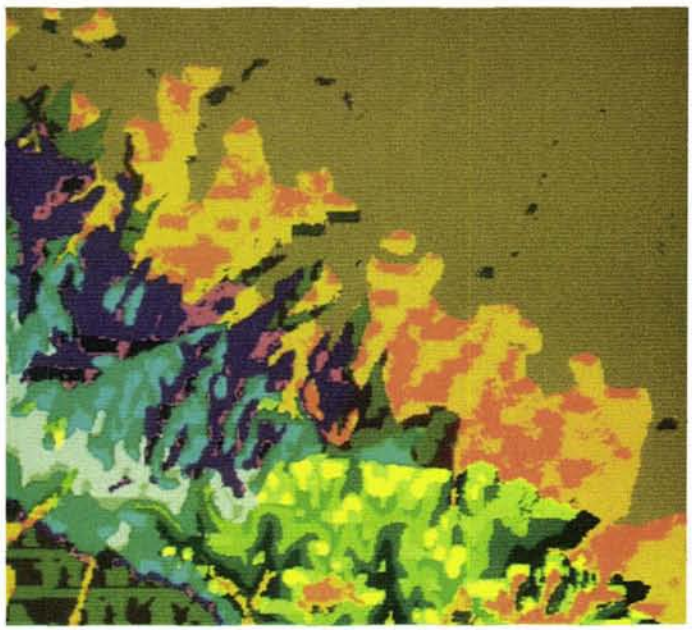

PLATE 2. Distribution of ecological land classes identified by mutual information analysis of digital satellite and terrain data. Color/classes legend: 1,tan; 2,light orange; 3,dark orange; 4,brown; 5,gray; 6,reddish brown; 7,dark aqua; 8,aqua; 9,pale aqua; 10,dark purple; 11, purple; 12,magenta; 13,light gray; 14,dark green; 15,green; 16,yellow green; 17,gray green; 18,pale gray green; 19,ivory.

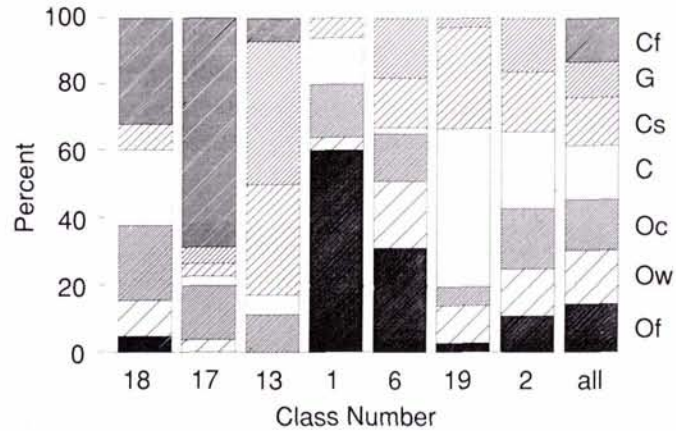

FIG. 5. Histograms of the relative frequency of different natural vegetation classes in 7 of 19 land classes and for the study area as a whole. Land class numbers are as in Table 2. Vegetation class abbreviations are OF (oak forest), OW (oak woodland), OC (oak chaparral), C (chaparral), CS (coastal scrub), CF (conifer forest) and G (grassland).

gain when the data are stratified into five geologic classes, and then smaller gains with further subdivision of the data based on topographic parameters.

After stratifying by geology, many subsequent splits were based on site differences in calculated insolation for the months of March or December. Oak forest and conifer forest were associated with sites receiving relatively low insolation (e.g., classes 1,7,10,12, and 14 through 17); chaparral, coastal scrub, and grassland predominated in classes receiving high insolation (notably classes 9,13, and 19). Vegetation pattern was associated most strongly with March or spring (integrated March through May) radiation on all rock types except for Sisquoc diatomite, where vegetation pattern was more strongly associated with December or total winter (integrated December through February) insolation. Association between vegetation and insolation was weaker for the months of April through June (Figure 7), when higher sun angles reduce spatial variation in insolation created by slope, exposure, and horizon effects. For example, the coefficients of variation for insolation maps from December, March, and June declined from 8.3 to 5.5 to 3.1, respectively.

Elevation was also a significant variable on some rock types in spite of the low range of elevations in the study area. In some cases, elevation served as a surrogate for geologic members of a formation. For example, the mudstone member of the Sisquoc shale occupied lower elevations than the diatomaceous member, and conifer forest occurred mostly on the latter (classes

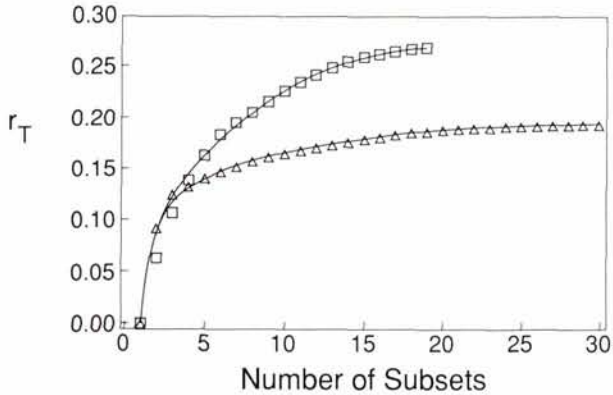

FIG. 6. Information redundancy (r_T, Equation 6) between vegetation and the ecological land classification as a function of the number of classes for the database (triangles) and test data (squares). Points are measured values and lines are fitted curves described by Equation 7 in the text.

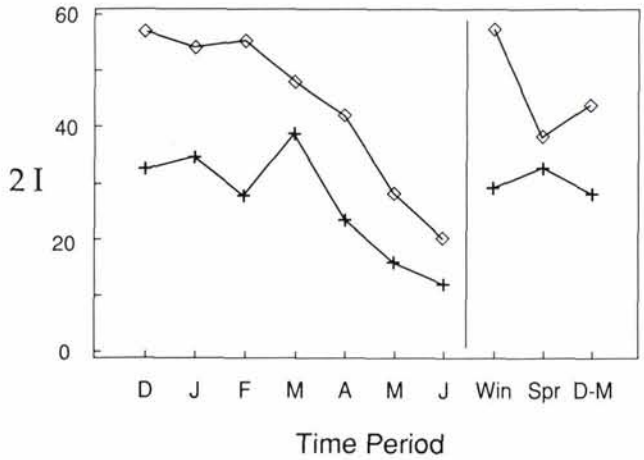

FIG. 7. Mutual information ($2I$) of vegetation and monthly or seasonal insolation, where insolation is cut into two classes at the point where $2I$ is maximized, for samples on Siquoc diatomite (diamonds) and Quaternary deposits (crosses). Winter (Win) months are December through February, spring (Spr) months are March through May, and D-M is integrated December through May.

14 versus classes 15,16, and 17). Also, lower elevations of the same rock type were usually lower hillslope or riparian environments that were probably more mesic. Thus, south facing slopes at higher elevations on Lower Careaga sandstone (class 13) were dominated by coastal scrub and grassland, whereas oak forest, woodland, and chaparral were relatively frequent at lower elevations (class 11).

The ecological land classes were significantly associated with vegetation classes for the 300 test sites analyzed by aerial photointerpretation ($2I=216$, $p<0.001$), and the 19 terrain classes accounted for 26.67 percent of the vegetation information in the 300 sites. This is a substantial improvement over the 18.7 percent information captured in the original training data, and we attribute the improvement to improved quality of the topographic and vegetation data for the test sites.

The relationship between r_T and the number of subsets S can be described by the equation

$$r_T(S) = \frac{a\,(\tanh(bS) - \tanh(b))}{\tanh(cS)} \tag{6}$$

where the coefficients a,b, and c are found by a least-squares fit to the data (Figure 6). The asymptote is defined by

$$\text{limit}_{S\to\infty}\, r_T(S) = a\left(1 - \tanh(b)\right). \tag{7}$$

As discussed by Phipps (1981), the asymptote is of theoretical interest as an estimate of the full association of vegetation and terrain based on the established hierarchical scheme. The relationship between r_T and T for the sample from the database has an estimated asymptote of 0.196, or 19.6 percent ($a=0.2136$, $b=0.0827$, $c=0.103$), while for the test data the asymptote is approximately 0.275 ($a=0.3108$, $b=0.1156$, $c=0.210$). The difference is due largely to the higher mutual information content of topographic variables and vegetation for the test data. This is shown by the greater gains in redundancy with further subdivision of the geologic strata for the test data versus the samples from the database (Figure 6).

DISCUSSION

Ecological land classifications have been developed for many regions and at many scales. When suitable classifications already exist, the main uses of digital satellite and terrain data will be in mapping known land classes over areas of interest. There are many regions and applications for which suitable land classification schemes may not exist. In these instances, cartographic modeling such as demonstrated here, that is, GIS-based sampling, classification, and mapping, may actually contribute to defining regional ecological land classes. We are not suggesting that cartographic modeling can substitute for field sampling in developing ecological land classification systems. However, the types of cartographic analyses conducted here complement traditional field survey methods by measuring associations or testing field results with many more random samples and at larger spatial scales than can practically be collected in the field.

The usefulness of the land classification scheme that we developed for the study area depended on both cartographic and ecological considerations as well as complex interactions between the two. The cartographic considerations included the resolution, accuracy, and bias of terrain maps, and these in turn depended on the classification systems used to map categorical variables and on the precision at which continuous variables were measured. For example, here we used seven vegetation classes defined by vegetation structure and dominant overstory species. Different results may have been obtained with a different classification scheme or had dominant understory species been analyzed (e.g., Carleton *et al.*, 1985; Davis *et al.*, 1988).

Much of the predictive skill of the classification was due to the strong influence exerted by surface geology. From a cartographic perspective, this macro-scale terrain variable could be accurately resolved and co-registered with other variables at 30-m resolution. On the other hand, 30-m resolution elevation data were too coarse to reliably reconstruct microtopography in the study region. Masking of regions in which the DEM was most sensitive to sampling resolution increased the measured association between vegetation and topographic variables, but had the undesirable effect of systematically removing distinctive environments such as ridges and riparian areas from the analysis. To improve the analysis, these need to be digitized directly from air photos or topographic sheets and stored as a separate variable in the database.

An important test of the land classification is whether it is consistent with existing ecological studies for the region. Support for the classification is provided by existing observations of the study area and similar Mediterranean ecosystems. The strong association of vegetation and geology, notably the as-

sociation of bishop pine forest with north-facing slopes on Sisquoc diatomite, was described by Cole (1980). The distribution of vegetation types with respect to solar radiation are what one would expect based on physiognomy and documented drought relations. For example, coast live oak forest, which was associated with more mesic classes, usually occurs on mesic sites in chaparral landscapes (e.g., Griffin, 1973; Campbell, 1980; Cole, 1980). Coastal scrub, which was associated with high insolation classes, is characteristic of more xeric substrates and topographic positions (e.g., Harrison *et al.*, 1971; Cole, 1980; Westman, 1983). Our analysis also suggests that oak woodland and oak chaparral are very similarly distributed with respect to topography. This is consistent with observations by Wells (1962) and Davis *et al.* (1988) that coast live oak cover increases during fire-free intervals on many chaparral-covered sites in the study area, and on these sites chaparral is probably seral to woodland (Wells, 1962; Davis *et al.*, 1988).

The use of digital topographic data to model monthly and seasonal radiation patterns has produced interesting results, although the generality of these results needs further testing using elevation data of higher resolution and quality and better atmospheric data. Most importantly, vegetation pattern in this example was more strongly associated with maps of winter or spring radiation than with maps of slope orientation that neglected horizon shading effects or with maps of integrated annual radiation. Differences in correlation between monthly radiation and plant species' distributions have also been documented for forest species by Kirkpatrick and Nunez (1980). These differences possibly reflect differences in the sensitivity of establishment or growth of the canopy species to seasonal patterns in radiation-related parameters such as soil temperature, evapotranspiration, or total photosynthetically active radiation (Kirkpatrick and Nunez, 1980).

The success of a land classification ultimately depends on its onsite explanatory power, utility, or predictive skill (Goodall, 1966). The data used here to test the classification are less than ideal because the topographic variables were not measured directly. Nevertheless, the fact that the ecological land classification had high predictive skill when tested with high quality map data and air photos is encouraging. The fact that the classification accounted for only one-fourth of the vegetation information in the test data is neither surprising nor discouraging. The vegetation pattern in the test area is extremely complex, reflecting land use and fire histories as well as variations in geology, soil, and topgraphy. Also, the maps used in the analysis supplied an imperfect model of the terrain variables. In spite of these limitations, the classification was sensible based on previous ecological surveys in the region and showed reasonable predictive skill.

In summary, the classification technique described by Phipps (1981) and here called mutual information analysis has proven to be useful for an application of geographical information system software to ecological land classification. The technique is one of a variety of algorithms that have been developed for predictive classification (see Legendre and Legendre, 1983; Breiman *et al.*, 1984) and its main advantages in the context of GIS applications are the specification of a categorical dependent variable, the use of categorical independent variables, and the relationship of the resulting association measure to a well-known statistical distribution (χ-squared). In this application, the method has been usefully applied to digital remotely sensed and terrain data to identify terrain classes that were associated with vegetation pattern.

ACKNOWLEDGMENTS

This study was supported in part by the National Aeronautics and Space Administration under Grant No. NAG 5-917 and by the California Space Institute.

REFERENCES

Bailey, R. G., 1983. Delineation of ecosystem regions. *Environmental Management*, Vol. 7, pp. 365–373.

Band, L. E., E. F. Wood, 1988. Strategies for large-scale distributed hydrologic simulation. *Applied Mathematics and Computation*, Vol. 27, pp. 23–37.

Berry, J. K., 1987. Fundamental operations in comuputer-assisted map analysis. *International Journal of Geographic Information Systems*, Vol. 1, pp. 119–136.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984. *Classification and Regression Trees*. Wadsworth, Belmont, California.

Burrough, P. A., 1986. *Principles of Geographic Information Systems for Land Resource Assessment*. Clarendon Press, Oxford.

Campbell, B., 1980. Some mixed hardwood forest communities of the coastal ranges of southern California. *Phytocoenologia*, Vol. 8, pp. 297–320.

Carleton, T. J., R. K. Jones, and G. Pierpont, 1985. The prediction of understory vegetation by environmental factors for the purpose of site classification in forestry: an example from northern Ontario using residual ordination analysis. *Canadian Journal of Forest Research*, Vol. 15, pp. 1099–1108.

Cliff, A. D., and J. K. Ord, 1981. *Spatial Processes*. Poin Ltd, London.

Cole, K., 1980. Geologic control of vegetation in the Purisima Hills, California. *Madroño*, Vol. 27, pp. 79–89.

Davis, F. W., 1987. Thematic Mapper analysis of Coast live oak in Santa Barbara County. *Proceedings of the Symposium on Multiple-Use Management of California's Hardwood Resources*, U.S. Forest Service Gen. Tech. Report PSW100, San Luis Obispo, pp. 317–324.

Davis, F. W., S. Goetz, and J. Franklin, 1986. The use of digital satellite and elevation data in chaparral ecosystems research. *Proceedings of the Conference on Chaparral Watersheds*, Water Resources Center, Davis, pp. 19–27.

Davis, F. W., D. E. Hickson, and D. C. Odion, 1988. Composition of maritime chaparral related to fire history and soil, Burton Mesa, California *Madroño*, Vol. 35, pp. 169–195.

Dibblee, T. W., 1950. *Geology of Southwestern Santa Barbara County*. Bulletin 150, California Division of Mines, Sacramento, 127 p.

Dozier, J., J. Bruno, and P. Downey, 1981. A faster solution to the horizon problem. *Computers and Geosciences*, Vol. 7, pp. 145–151.

Ferren, W. R., H. C. Forbes, D. A. Roberts, and D. M. Smith, 1984. *The Botanical Resources of the La Purisima Mission State Historic Park*. Herbarium Publication No. 3, University of California, Santa Barabara.

Forman, R. T. T., and M. Godron, 1986. *Landscape Ecology*. John Wiley and Sons, New York.

Franklin, S. E., 1987. Terrain analysis from digital patterns in geomorphometry and Landsat MSS spectral response. *Photogrammetric Engineering and Remote Sensing*, Vol. 53, pp. 59–65.

Frew., J and J. Dozier, 1986. The Image Processing Workbench - portable software for remote sensing instruction and research. *Proceedings of the 1986 International Geoscience and Remote Sensing Symposium*, European Space Agency, Paris, pp. 271–276.

Gauch, H. G., Jr., and R. H. Whittaker, 1981. Hierarchical classification of community data. *Journal of Ecology*, Vol. 69, pp. 537–557.

Goetz, S., 1987. *Predictive Mapping of Coast Live Oak in California Using Digital Terrain Data*. M. A. Thesis, Department of Geography, University of California, Santa Barbara, 99 p.

Goodall, D. W., 1966. Classification, probability and utility. *Nature*, Vol. 211, pp. 53–54.

Griffin, J. R., 1973. Xylem sap tension in three woodland oaks of central California. *Ecology*, Vol. 54, pp. 152–159.

Harrison, A. T., E. Small, and H. A. Mooney, 1971. Drought relationships and distribution of two Mediterranean-climate California plant communities. *Ecology*, Vol. 52, pp. 869–875.

Jenny, H., 1980. *The Soil Resource: Origin and Behavior*. Springer-Verlag, New York.

Kirkpatrick, J. B., and M. Nunez, 1980. Vegetation-radiation relationships in mountainous terrain: eucalypt-dominated vegetation in the Risdon Hills, Tasmania. *Journal of Biogeography*, Vol. 7, pp. 197–208.

Kullback, S., 1959. *Information Theory and Statistics.* John Wiley, New York.

Legendre, L., and P. Legendre, 1983. *Numerical Ecology.* Elsevier Scientific, New York.

Mabbutt, J. A., 1968. Review of concepts of land classification. *Land Evaluation: Papers of a CSIRO Symposium* (G. A. Stewart, ed.). MacMillan Press, Melbourne. pp. 11–28.

Marks, D., J. Dozier, and J. Frew, 1984. Automated basin delineation from digital elevation data. *Geo-Processing,* Vol. 2, pp. 279–311.

Michaelsen, J., F. Davis, and M. Borchert, 1986. Non-parametric methods for analyzing hierarchical relationships in ecological data. *Coenoses,* Vol. 1, pp. 97–106.

Miller, P. C., E. Hajek, D. K. Poole, and S. W. Roberts, 1981. Microclimate and energy exchange. *Resource Use by Chaparral and Matorral* (P. C. Miller, ed.). Springer-Verlag, New York. pp. 97–121.

Morissey, L. A., and L. L. Strong, 1986. Mapping permafrost in the boreal forest with Thematic Mapper satellite data. *Photogrammetric Engineering and Remote Sensing,* Vol. 52, pp. 1513–1520.

Naveh, Z., and A. S. Lieberman, 1984. *Landscape Ecology: Theory and Application.* Springer-Verlag, New York. 356 p.

Oliver, M. A., and R. Webster, 1986. Semi-variograms for modelling the spatial pattern of landform and soil properties. *Earth Surface Processes,* Vol. 11, pp. 491–504.

Orlóci, L., 1978. *Multivariate Analysis in Vegetation Research.* Dr. W. Junk, The Hague.

Paysen, T. E., J. A. Derby, H. Black, V. C. Bleich, and J. W. Mincks, 1980. *A Vegetation Classification System Applied to Southern California.* General Technical Report PSW-45, Pacific Southwest Forest and Range Experiment Station, Berkeley, 45 p.

Phipps, M., 1981. Entropy and community pattern analysis. *Journal of Theoretical Biology,* Vol. 93, pp. 253–273.

Rowe, J. S., and J. W. Sheard, 1981. Ecological land classification: a survey approach. *Environmental Management,* Vol. 5, pp. 451–464.

Shipman, G. E., 1972. *Soil Survey of the Northern Santa Barbara Area.* U.S. Department of Agriculture Soil Conservation Service Report, Washington, D.C..

Steward, D., and P. J. Webber, 1981. The plant communities and their environments. *Resource Use by Chaparral and Matorral* (P. C. Miller, ed.). Springer-Verlag, New York. pp. 43–68.

Stocker, M., F. F. Gilbert, and D. W. Smith, 1977. Vegetation and deer habitat relations in southern Ontario: classification of habitat types. *J. of Applied Ecology,* Vol. 14, pp. 419–432.

Strahler, A. H., 1981. Stratification of natural vegetation for forest and rangeland inventory using Landsat digital imagery and collateral data. *International Journal of Remote Sensing,* Vol. 2, pp. 15–41.

Wells, P. V., 1962. Vegetation in relation to geological substratum and fire in the San Luis Obispo Quadrangle, California. *Ecological Monographs,* Vol. 32, pp. 79–103.

Westman, W. E., 1981. Factors influencing the distributions of species of California coastal sage scrub. *Ecology,* Vol. 62, pp. 170–184.

———, 1983. Xeric Mediterranean-type shrubland associations of Alta and Baja California and the community/continuum debate. *Vegetation,* Vol. 52, pp. 3–19.

(Received 25 October 1988; accepted 10 July 1989; revised 21 July 1989)