G I S
93

GIS'93
Symposium

Vancouver
British Columbia
February 1993

From pixels to polygons: the rule-based aggregation of satellite image classification data using ecological principles

Kenneth A. Stumpf

Geographic Resource Solutions(GRS)
1125 Sixteenth Street, Suite 213
Arcata, CA 95521
(707) 822-8005

Abstract

Raster pixel data developed using satellite image classification techniques are frequently difficult to convert to a polygon (vector) format due to the extreme heterogeneity of the pixel classification data. Many groups of pixels are too small to map as polygons without yielding an unusable database. The small areas that are less than a user defined minimum size mapping unit must be aggregated with neighboring groups (stands) prior to developing a usable vector database. Conventional spatial operators based on either grid or polygon analysis (neighborhood and/or sliver filters) often cause degradation of stand boundaries and descriptive attributes, and decrease the reliability of the final map. A rule-based pixel filtering methodology that is based on user-defined concepts of stand similarity is presented in this paper. This technique considers stand characteristics such as the major vegetation type, species composition, density of canopy closure, average tree size, and canopy structure during the evaluation of stand similarity. Aggregation rules representing ecological relationships, minimum size constraints, and the relative importance of the different vegetation characteristics are also used to guide the aggregation process. The rules are flexible and may be defined relative to project objectives and the desired use of the resulting database.

Introduction

Image processing and classification techniques have being used to develop thematic information describing vegetation characteristics for extensive areas of rugged forested terrain (Brown and Fox, 1992). During the past two years, GRS has been mapping the vegetation characteristics of two and one-half million hectares (six-million acres) of both public and private lands in northern California. An additional five-million hectares (twelve-million acres) is to be mapped during 1993. One of the major challenges of the image classification and database development processes used in such a project involves the large datasets that must be evaluated and processed. The basic unit of analysis is a raster pixel that is 25 meters square or 0.0625 hectares (0.15 acres) in size.

Individually, each classified pixel represents a data sample of a specific location. Together, these pixels may comprise a "pretty picture" that helps the user visualize and understand the status of the resources being mapped. However, the information these pixels represent is difficult to manage as (millions of) pixels as compared to when they are grouped or aggregated to form (hundreds or thousands of) vegetation types or stands. To effec-

tively evaluate and understand these data we must reduce the size and heterogeneity of the database. Therefore, the development of a thematic vegetation database in a vector format is often a desirable goal of a mapping effort of this nature.

Data management problems: Excessive information and detail

The basic problem concerning the generalization of the raster data to a vector format is that there is too much detail. The results of the image classification processes are raster grids that normally represent between 100 to 200 types and sometimes as many as 400 types. Large groups of homogeneous pixels are uncommon. Instead the pixel grids usually consist of a very heterogeneous mix of pixels that are frequently isolated or found in very small groups, smaller in size than the minimum size allowed in the database. One cannot simply vectorize the resulting raster data without creating many, many, tiny polygons the size of an individual pixel.

A commonly used method of resolving problems of excessive detail as represented by isolated or small groups of homogeneous pixels that differ from their neighboring pixels is to filter the raster data and remove these undesirable pixels. The value of the undesirable pixel(s) is altered to smooth the data and produce "cleaner" data. Modal filters, or other mathematically based filters are often used to alter the value of the undesirable pixel(s). A modal filter changes the value to reflect the pixel type that occurs most frequently in the immediate area (window) evaluated around the undesirable pixel. This approach may be appropriate for pixels completely surrounded by another type of pixel. However, modal filters may be inappropriately used to smooth or clean pixel data representing small interspersed homogeneous groups of data or along the edges of different types where mixed pixels are commonly found. In these situations we have found that mathematically based filtering has two negative impacts: the edges of stand boundaries are moved as multiple passes are made over the raster grid, and the stand characteristics or attributes of the resulting types are sometimes incorrectly changed. The mathematical filtering adversely effects the delineation of stand type boundaries as the edges "creep" when subjected to multiple passes of the filter. Filtering can also alter the type characteristics to reflect a different and incorrect type than that represented by the previous stand type. This situation occurs in particular along the edges of different types, such as a conifer type and a grassland type, where mixed pixels are present.

If enough mixed pixels that represent some level of tree cover are filtered into the grassland type, the tree cover of the grassland type may exceed the minimum percent cover threshold (e.g. 10 percent) required for a tree type definition, and the grassland type is then identified as a low density conifer type rather than a grassland type.

The challenge is to aggregate the very detailed pixel data into more generalized vegetation type information without changing the basic description of the vegetation. Aggregation should be compensating. Major type boundaries should be preserved, and the distribution of hectares by vegetation type of the pixel grid should approximate the distribution of hectares by vegetation type of the generalized database. Significant differences would tend to indicate that the aggregation process is biased. This principle cannot always be demonstrated as there are new types that are derived from the aggregation of pixels. For example, there may be individual pixels that represent different vegetation types, such as either Douglas-fir, mixed conifer/hardwood, or alder. However, there may be polygons that are characterized as mixed conifer/hardwood types that do not contain any mixed conifer/hardwood pixels. The mixed conifer/hardwood characteristic is determined from the combination of pixel characteristics at the polygon level rather than the characteristics of the individual pixels within the polygon; a mixture of Douglas-fir pixels and alder pixels within the stands boundaries combine to make the stand a mixed conifer/hardwood type.

Aggregation methodology

GRS has developed a process to accomplish the aggregation of pixels to polygons. A flow chart of this process is shown in Figure 1. The aggregation of the pixel data to form a generalized thematic database requires the definition of two very important sets of rules. One set of rules describes the definitions of type characteristics that will be recognized and evaluated and the relative significance (weight) of these characteristics. The second set of rules defines the minimum allowable size mapping unit - the smallest mappable area that will be included in the database. These sets of rules are related; each type or class of characteristics can have its own specific minimum area. These rules are significant as the aggregation process reduces between stand variation and increases within stand variation. The rules effect what types of variation will occur within the larger more generalized stand types.

Vegetation type definitions, rules, and weights

Aggregation is based on an evaluation of the similarity of adjacent stands. Similarity is estimated on the basis of the type definitions and rules that have been defined. Only those characteristics present in the database may be considered in this process.

Many of the different vegetation characteristics, such as the vegetation type, crown density, average size diameter, stems per unit area and canopy structure, may be used in this evaluation. Other attributes that may be used include the major vegetation type (i.e. conifer, hardwood, shrub, herbaceous, and so forth), the predominant species cover, and the percent conifer or hardwood composition. Each type characteristic used in

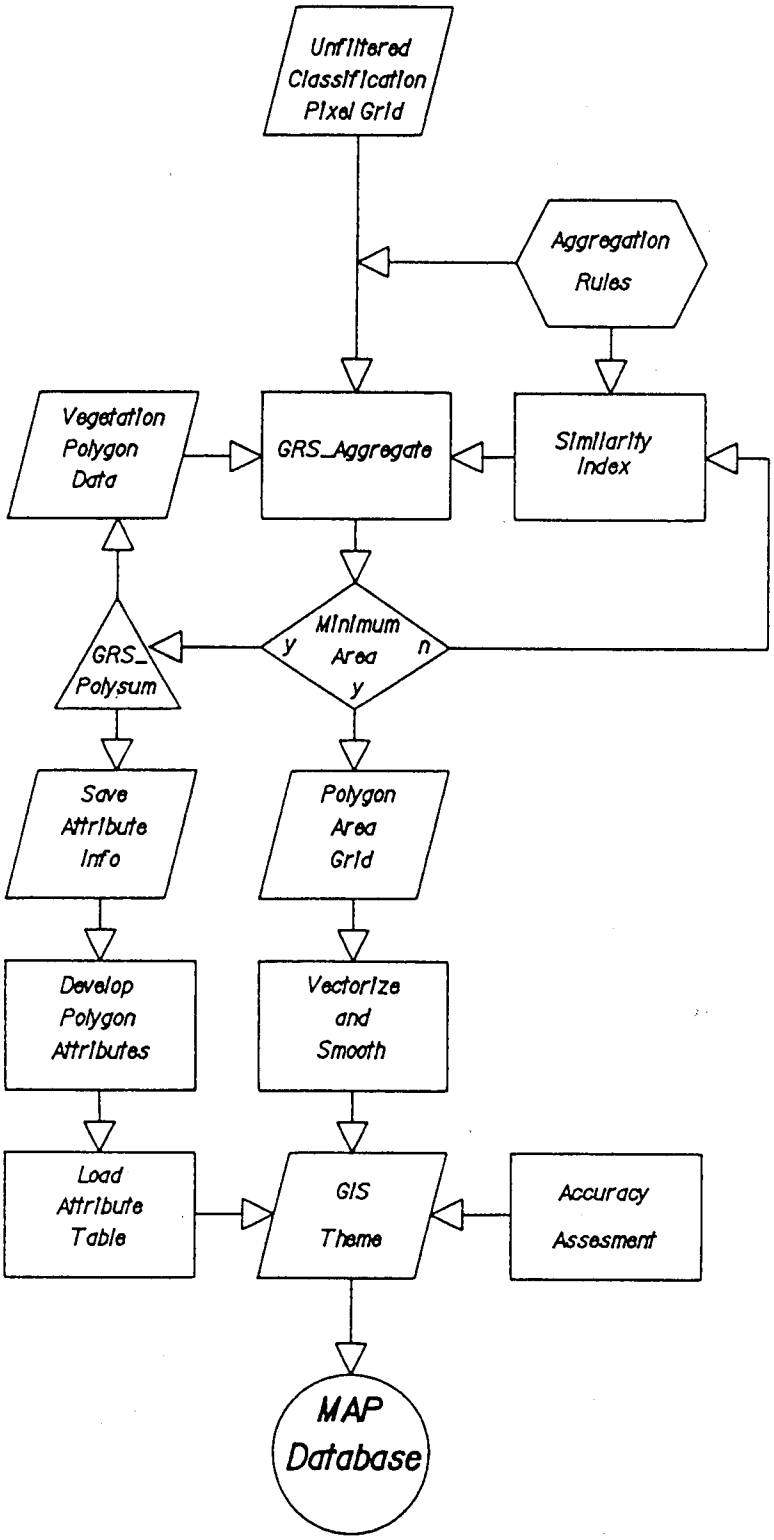

Figure 1: *Schematic workflow for aggregating pixel data into thematic polygon.*

the evaluation must have a quantitative definition as either a value or a range of values. For example, the measure of tree size may be the estimate of quadratic mean tree diameter, whereas the definition of a non-tree type may be an area with less than 10 percent tree canopy closure.

Aggregation is based on the premise that if all the characteristics of adjoining groups are the same except for one, then the most similar adjoining stand is the stand with the most similar or least different characteristic. Differences between the subject stands characteristics and the adjacent stands' characteristics are estimated to enable a quantitative estimate of similarity. Figure 2 shows three examples of single theme characteristics of type, size, and density. The similarity based on size and density can be estimated by evaluating the difference in the values of these characteristics. The small stand A would be merged into the most similar stand B in each of these cases. Similarity based on the type evaluation would yield a different answer, as stand A is most similar to the other redwood stand C. An evaluation that considers a characteristic such as the average size or density is fairly straightforward. An evaluation that considers a more subjective characteristic such as the vegetation type is more difficult. Rules and weights must be developed to estimate the relative magnitude of the contributions of the different vegetation types. The most similar adjacent type is obviously the same type. If the subject area is a Douglas-fir (conifer) type and adjacent areas are all different types, then the next most similar type would probably be another conifer type, as opposed to a hardwood type. An adjacent tree type would be more similar than an adjacent non-tree type, and of the adjacent non-tree types the brush and herbaceous types would probably be more similar than the water, river bar, and bare ground types. These kinds of relationships cannot always be uniformly applied. The estimation of these rules is heavily reliant on ecological relationships and concepts. For example, hardwoods are not always more similar to conifers than they are to grassland types. The associations of live oak with grassland types and tanoak with conifer types are known and may be included in the estimation of similarity. Hardwood pixels should be aggregated with the type that the specific hardwood type is commonly associated with in its natural range instead of generalizing and always merging a hardwood type with a conifer type rather than with a grassland type. The aggregation process must be flexible to accommodate generalization of specific types according to specific rules.

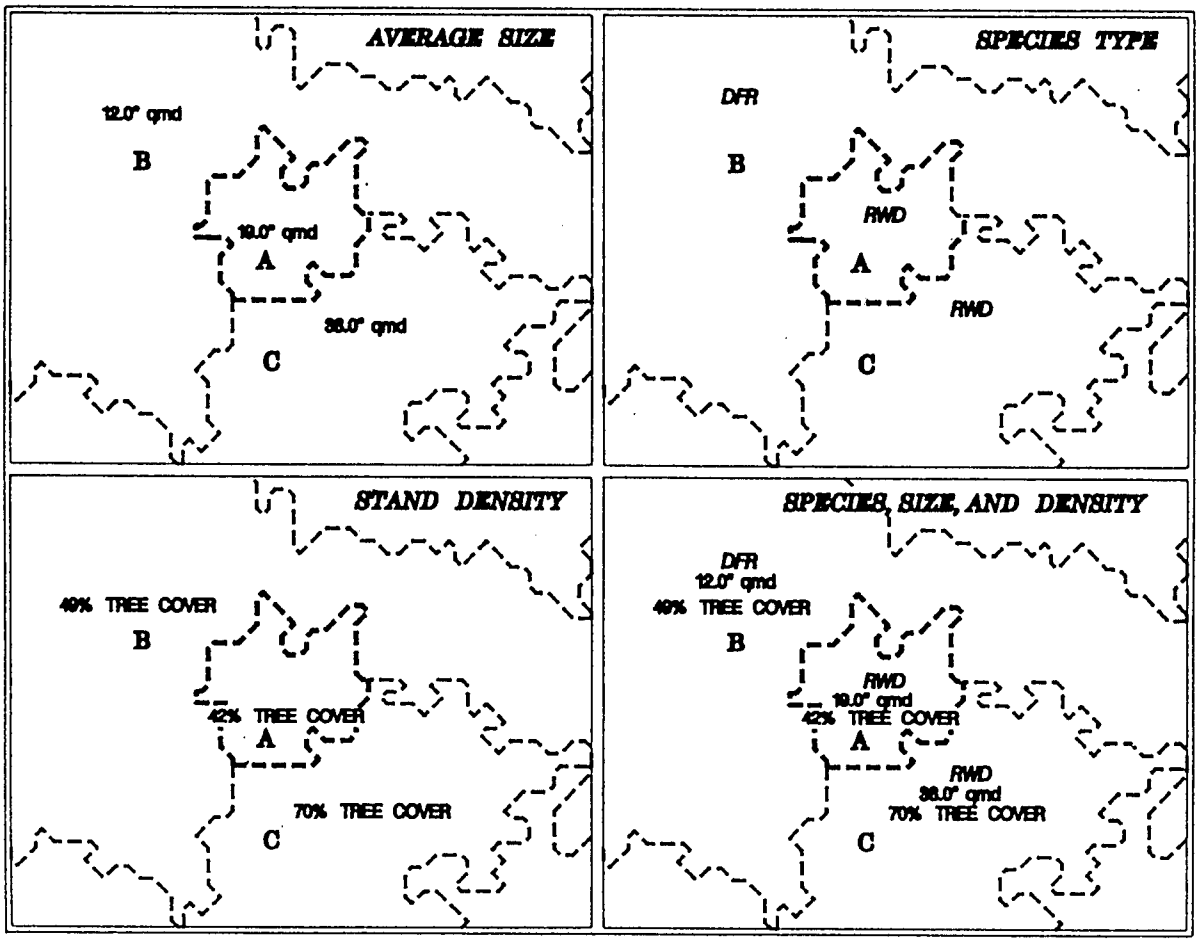

Figure 2: *Stand aggregation by characteristic.*

Aggregation characteristics that involve only one characteristic are relatively easy to make compared to choices involving multiple characteristics. Most often, differences of multiple characteristics, such as tree crown cover, average diameter, and vegetation type, are observed as shown in the lower right cell of Figure 2. The question of similarity most often requires a more complex evaluation than demonstrated in Figure 2. There are usually more than two adjacent stands and the stand characteristics are frequently dissimilar, as shown in Figure 3. Some of the kinds of differences (vegetation type versus average diameter versus crown density) may be more significant than others. Levels of significance can be estimated and represented by assigning weights or factors to the kind of difference being estimated. Diameter differences may be twice as important as density differences. Species type differences may be small between conifer types but large between major vegetation types such as hardwood, brush, and herbaceous. The relative weights of different characteristics must be evaluated and estimated before aggregating pixels to produce a final map.

The factors and relationships used to develop similarity indices are attempts to recognize differences between vegetation types and characteristics, and they reflect the classification rules selected for the thematic database. These weights and factors are not fixed and they can be modified to reflect other interpretations of ecological associations and the relative significance of the different vegetation characteristics. Project objectives can play a significant role in the determination of the significance of the different characteristics and the role any one characteristic plays in the aggregation of the pixel data. For example, from a wildlife managers viewpoint, if wildlife respond to tree size more than canopy cover, then the aggregation process should preserve groups by size or seral stage representing a variety of densities rather than groups by similardensity representing a wide range of sizes. Similarly, a botanist may be more interested in species' purity or diversity and may develop a different set of rules that accentuates species similarity or diversity. A forester may be interested in mapping areas with characteristics of volume per hectare by species type and develop rules and weights that would tend to generate strata required for a first stage sample for a timber inventory of commercial species. Different maps and information may be developed depending on the goals and objectives of a project as defined in the rules and weights used to guide the aggregation process.

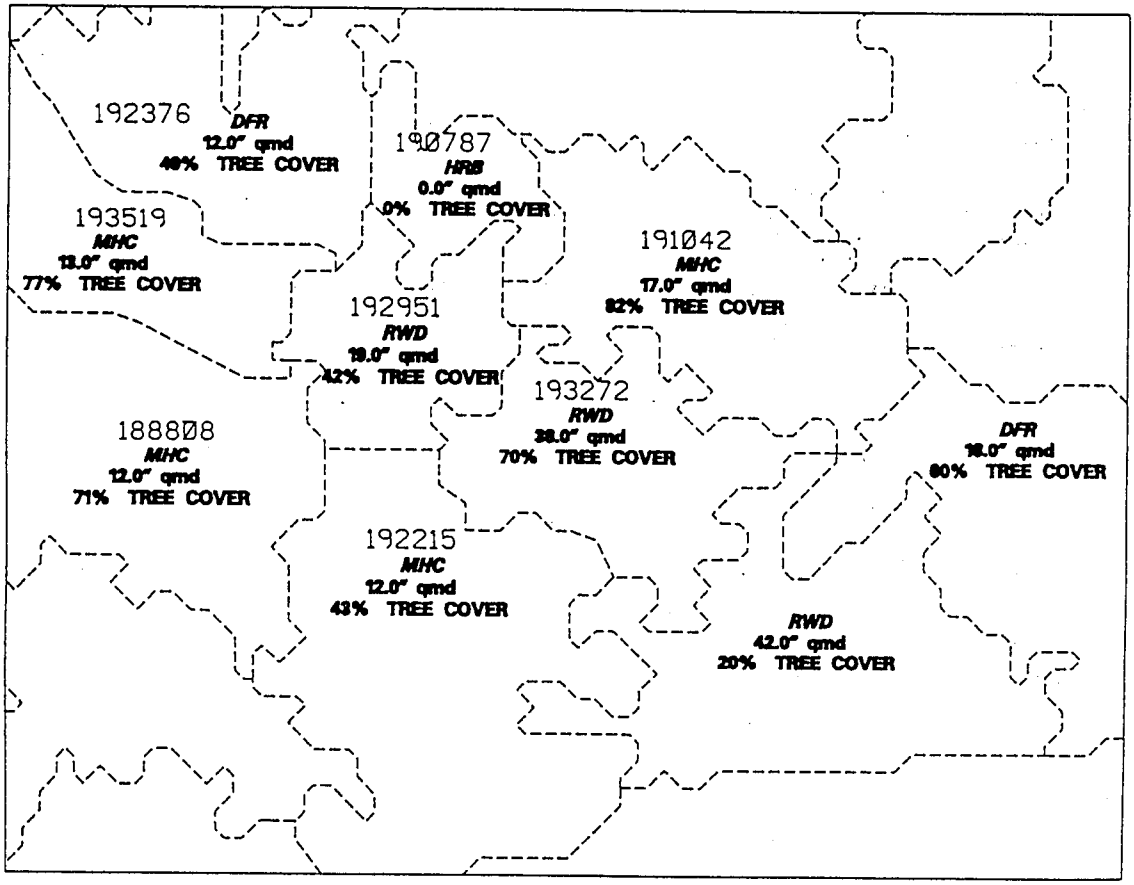

Figure 3: *Typical stand characteristics evaluated for similarity.*

Minimum size mapping unit

The minimum size mapping unit is the smallest size area that will be represented in the final map. Some significantly different stands should not be aggregated with each other, if possible. For example, non-tree types such as small brushfields, prairies, barren areas, and bodies of water should not, if possible, be merged into surrounding vegetation types such as coniferous forest, mixed conifer/hardwood, and hardwood types. Preservation of distinctly different stands is necessary to maintain the accuracy of the mapping effort since fewer stands are generated that represent a mixture of significantly different characteristics solely for the purpose of satisfying a minimum size constraint. The minimum size limit obviously effects the capability of any map to accurately represent what is present on the ground. The larger the minimum size mapping unit, the greater the probability that a stand represents a diverse grouping of heterogeneous types that could have been represented by smaller, more homogeneous stands if the minimum size limit were smaller. The characteristics of the vegetation being mapped and the projects information needs and objectives are integral to the determination of the appropriate minimum size limits developed and used for each vegetation characteristic. This aggregation process allows the definition of different minimum sizes for different vegetation attributes or characteristics. For instance, areas mapped as "large size" might have a minimum of five hectares while areas mapped as "small size" might have a minimum size of twenty hectares. Two minimum sizes are used for each characteristic considered during the check of minimum

Table 1: *Aggregation possibilities for stand 192951.*

Aggregating Stand 192951 - 7 adjacent stands to check ...

Veg Stand ID#	Type	Pr SP	Pct Cover	Pct Conif	Avg DBH	Veg Form	Canopy Structure	
192951	RWD	RWD	42	78	19	10	EVEN	
188808	MHC	HWC	71	42	12	17	EVEN	
simindex =	8.0	5.5	7.3	7.2	4.7	3.5	0.0 =	36.1
192951	RWD	RWD	42	78	19	10	EVEN	
190787	HRB	ARC	0	0	0	35	UNDF	
simindex =	26.0	25.0	10.5	25.0	25.0	12.5	0.0 =	124.0
192951	RWD	RWD	42	78	19	10	EVEN	
191042	MHC	DFR	82	64	50	17	EVEN	
simindex =	8.0	0.5	10.0	2.8	41.3	3.5	0.0 =	66.1
192951	RWD	RWD	42	78	19	10	EVEN	
192215	MHC	HWC	43	55	12	17	EVEN	
simindex =	8.0	5.5	0.1	4.6	4.7	3.5	0.0 =	26.4
192951	RWD	RWD	42	78	19	10	EVEN	
192376	DFR	DFR	49	73	12	10	EVEN	
simindex =	1.0	0.5	0.9	1.0	4.7	0.0	0.0 =	8.0
192951	RWD	RWD	42	78	19	10	EVEN	
193272	RWD	RWD	70	79	38	10	UNEVEN	
simindex =	0.0	0.0	7.0	0.2	12.7	0.0	15.0 =	34.9
192951	RWD	RWD	42	78	19	10	EVEN	
193519	MHC	HWC	77	47	13	17	EVEN	
simindex =	8.0	5.5	8.8	6.2	4.0	3.5	0.0 =	36.0

minimum simindex = 8.0 for stans id# = 192376

RWD = redwood
DFR = Douglas-fir
MHC = Mixed conifer/hardwood
HWC = Hardwood associated with conifers
SHR = shrub
ARC = Manzanita sp.

sizes. The first value is the critical minimum, a level below which no stand of that value of characteristic may exist in the final database. The second value is a desirable minimum, a level to which stands will be aggregated so long as they are similar enough to each other, as defined in the aggregation rules. In other words, dissimilar stands are aggregated to the critical minimum, whereas similar stands are aggregated to the desirable minimum. The use of two minimum size limits for the different characteristics being mapped enables the preservation of distinctly different stands but also the formation of large stands comprised of fairly similar characteristics. This approach provides flexibility in defining how the various vegetation characteristics will be mapped.

Pixel/polygon aggregation

In order to apply the rule-based aggregation routines, the image processing classification grid is attributed with the vegetation characteristics represented by each pixel. Aggregation must consider a wide range of stand sizes, from those represented by individual pixels, or small groups of pixels to those that represent large groups of pixels that already exceed the minimum size requirements and form valid stands. The sub-minimum size groups or stands must be aggregated with other stands to form a database of valid size stands.

When a sub-minimum size stand is recognized, it is aggregated with the adjacent stand estimated to be the most similar to the subject area. The differences of stand characteristics between the subject stand and the neighboring stands are estimated as absolute values, multiplied by the appropriate weight, and then summed to estimate a similarity index. An example of this process for a sub-minimum size stand (192951 in Figure 3) is shown in Table 1. In this example stand 192951 would be aggregated with stand 192376 since they are the most similar. This example also illustrates that small differences in percent cover, percent conifer composition, and average size diameter are more significant than a species difference between redwood and Douglas-fir, two species that grow in close association in northern California.

Aggregation is performed in several steps, starting with low minimum size limits and progressing to larger levels with each pass through the data. As stands are merged, their attributes are recalculated based on the previous characteristics plus those of the included stand. The aggregation process performed in one step, from the initial pixel groups to the final size limits, is difficult to process and tends to result in larger more generalized types than a step-wise aggregation process. The step-wise aggregation process involves smaller size increases and tends to merge smaller numbers of stands during each step. This approach maintains stands of more similar characteristics rather than merging many small size stands at once into a few large generalized stands. A benefit of the step-wise aggregation approach is that maps reflecting different intermediate minimum size limits may be developed as intermediate products. A five-hectare map may be generated from the intermediate results of aggregation using the five-hectare size limits. A ten- or twenty-hectare map may be developed by continuing the aggregation process using an intermediate ten-hectare and final twenty-hectare size limits (note: these size limits are used as examples whereas the process actually allows variable size limits for different characteristics). These databases are then vectorized using standard vectorization routines. The vectors are then smoothed to remove the stair-step appearance characteristic of vectors derived from pixel (raster) grids to reduce the size of the database. The characteristics of the final polygons are then determined and loaded into the relational database tables.

Estimation of polygon characteristics

The final estimate of each stands vegetation characteristics is based on the summarization of all the different types of image classification pixels found within each of the final stand boundaries. The pixels of all sub-minimum size stands that are merged into the final polygons are included in the polygon summaries. Therefore, aggregated sub-minimum size stands are included in the summaries and contribute to the average characteristics of the final stands.

Summary

Pixel data developed using image processing techniques are characterized by extreme heterogeneity and large size databases. Pixels may be aggregated into polygons based on the vegetation characteristics represented by the pixel data. Rules and weights that estimate the relationships between different characteristics, minimum size areas to be mapped, and the objectives of the mapping effort may be used to guide the aggregation process. Stands are aggregated based on their similarity to adjacent stands as estimated using the rules, weights, and minimum size limits that have been defined by the user.

Literature cited

Brown, G. and L. Fox, 1992. Digital classification of thematic mapper imagery for recognition of wildlife habitat characteristics. Proc: 1992 ASPRS/ACSM Convention, American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, Vol. 4, pp. 251-260.