Predicting map error by modeling the Sacramento River floodplain

Joshua H. Viers[1], Alexander K. Fremier[2], and Rachel A. Hutchinson[1]
[1]*Department of Environmental Science and Policy, University of California, Davis*
[2] *College of Natural Resources, University of Idaho*
jhviers@ucdavis.edu

Abstract
We quantified the map accuracy for the Sacramento River Monitoring and Assessment Project to help land and water manager's better plan for restoration efforts. While map errors are quantifiable and even predictable, linking the causes of error to complex environmental and geographic variables would improve decision making. We evaluated patterns of GIS-induced map error on over 32,000 acres based on environmental and GIS variables like floodplain age and edge complexity. We conducted extensive field validation and used spatial statistics to compare environmental variables with vegetation map inaccuracies. We then constructed a multivariate model to predict errors in certain vegetation types. We field validated 15% of map polygons (n=8,067) which were 85% correct (K=0.83). Using validated polygons, we found errors occurred most frequently on older floodplains but rates varied by vegetation type. By incorporating error in attribution and spatial assignment, restoration planners have a more realistic assessment of current conditions.

The Sacramento River

Recommended Citation: Viers, JH, AK Fremier, and RA Hutchinson. 2010. Predicting map error by modeling the Sacramento River floodplain. Proceedings from the 2010 ESRI International User Conference, San Diego, California. 21 ppd.

Introduction

Large floodplain rivers are heavily impaired in many parts of the world and will require large-scale restoration to maintain habitat diversity that supports naturally functioning riparian and aquatic ecosystems. Broad-scale maps of vegetation are valuable as they provide data that quantifies ecosystem state and land cover change for efforts to create management plans. Vegetation maps are generally constructed by spatially locating and delineating the dominant species of distinct plant communities found within a specific map extent. This mapping process is often informed by on-the-ground vegetation surveys and airborne and /or spaceborne imagery (McDermid et al. 2005; Ustin et al. 2004; Lucas et al. 2008; He et al. 2006; Greenberg et al. 2006). However, vegetation maps of large river floodplains and riparian areas are often lacking critical information about the environmental condition at the time of depiction, which can be useful not only to managers and researchers trying to develop management plans, but also to understand change in condition through time as riverscapes are inherently dynamic both spatially and temporally. Understanding dynamic riverscape processes through time is difficult, and understanding vegetation dynamics is further challenged if the underlying spatial data are incorrect, either in location or codification.

Riparian ecosystems are considered some of the most diverse and complex ecosystems in the world and, while this makes them appealing to conservationists and resource managers, it is a challenge to inventory existing resources and direct conservation efforts in such a dynamic system. Vegetation maps in particular are difficult to produce in riparian areas due to habitat heterogeneity, poor understanding of vegetation communities, and low resolution imagery, among other factors (Congalton et al. 2002; Gergel et al. 2007). As they are relatively difficult to produce, these maps possess a variety of spatial and attribution errors that are often linked to difficulties associated with

interpreting habitat complexity. While often poorly understood, or worse ignored, these errors have a wide range of consequences including improper management decisions and misdirected restoration objectives (Gergel et al. 2007; Langford et al. 2006). Some of these errors could be avoided by the use of advanced imagery and by relating errors to specific causes or environmental variables (Ustin et al. 2004; Greenberg et al. 2006; Oldeland et al. 2010; Hestir et al. 2008). Error rates in mapping attribution can be easily quantified and even predictable, but providing in depth information that links complex environmental and geographic variables to error rates might be more useful to map makers and land managers. Further, important elements of a map created specifically for management and restoration are often overlooked and create missing or misleading attributes, such as structural components like canopy height. Vegetation maps of riverscapes are increasingly indispensible for management and restoration planning, as they capture the status of not only vegetation for a specific time in history, but also the dynamic and structural character of fluvial ecosystems, such as channel position and form (Figure 1).

Large floodplain riverscapes typically have meandering channels, which drive habitat formation and heterogeneity and create successional sequences that modify floodplain surfaces. Early succession occurs on newly created surfaces that mature as the bank stabilizes and secondary succession begins. Floodplain age, or the time elapsed since sediment deposition, becomes central in the understanding of successional sequences or why specific vegetation communities colonize in specific locations. Where datasets that quantify these processes are available, such as floodplain age and the relative elevation of a surface (see Fremier 2003; Greco et al. 2007 for examples), vegetation mapping efforts can integrate information about the successional history of a stretch of river and use that data to determine how floodplain age might influence particular stand traits and patterns while providing

a basis to determine where difficulties might exist when mapping highly heterogeneous riparian communities.

Figure 1. *Abandoned channel adjacent to the Sacramento River*

On the Sacramento River, multi-year attempts to map riparian vegetation now offer a nearly decadal view of existing riparian corridors and vegetation (Greco and Plant 2003; Nelson et al. 2008). Each map has its own assortment of errors ranging from minor (e.g., height misclassifications) to major (e.g., vegetation misclassifications). This study investigates the most recent vegetation map of the Sacramento River and evaluates patterns of error based on environmental variables like floodplain age and relative elevation. In this paper we ask three questions: How predictable are

misclassification errors? How useful are riverscape parameters in predicting misclassification errors? How useful is this approach for evaluating error in future mapping projects?

Methods

The 2007 Sacramento River riparian vegetation map extends from Red Bluff in the north to Colusa in the south (River Mile 144 to River Mile 245) in California's northern Sacramento Valley (Figure 2). The mapped area covers almost 13,278 hectares, delineates 8,067 individual polygons and includes 15 unique vegetation types (Nelson et al. 2008). There was no vegetation classification conducted specifically for this vegetation map; the 15 types included were compiled from surveys conducted by Vaghti (2003) and riparian specific vegetation communities documented in the Manual for California Vegetation (Sawyer et al. 2009; Sawyer and Keeler-Wolf 1995).

All digitization for the riparian map creation was conducted using a heads-up display in the ArcGIS 9.3 environment (ESRI, 2009) by California State University, Chico, Geographic Information Center (Nelson et al. 2008),with orthorectified aerial photographs (1:15,840 scale). Distinct stands were identified based on field reconnaissance points and implementing the California Native Plant Society and Department of Fish and Game Rapid Assessment protocol for quick identification of vegetation stand composition in the field. Stands were distinguished by their dominant overstory species and stands that had less than 10% overstory cover were considered herbaceous communities. Minimum mapping area was set to 0.5 acre, but smaller polygons (0.01 acre) were allowed for vegetation types of interest, such as invasive weeds (e.g., *Arundo donax*).

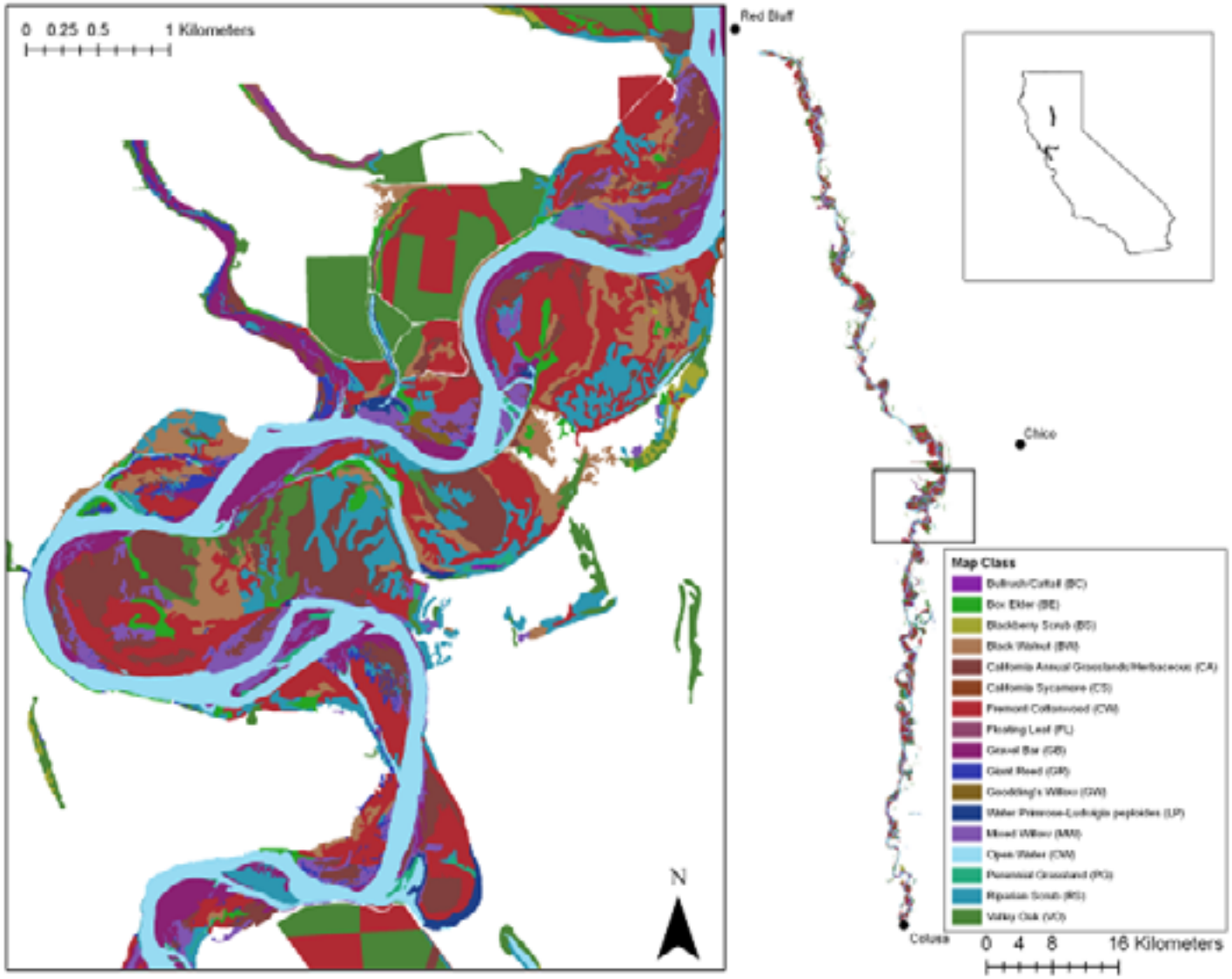
We used a combination of field visits and visual accuracy checks to determine attribute validation, described below. Once classification accuracy was known for polygons visited in the field, a

multivariate recursive partitioning analysis was performed to determine which if any riverscape variables could be used to improve mapping efforts in the future, and to predict additional polygons – not yet checked in the field for classification accuracy – that are likely incorrect in classification. These model variables include some elements of patch configuration, vis-à-vis FRAGSTATS (McGarigal and Marks. 1995), as well as fundamental factors of riverscape ecology, such as floodplain age and relative elevation

Attribute Validation

Field validation of the 2007 vegetation map was completed one year after the aerial photography was collected. The vegetation types included in the map were often denoted by a dominant species (cottonwood, valley oak, etc.) or by a collection of similar lifeforms such as annual grassland (CA; grasses and herbaceous species), riparian scrub (RS; shrubs and vines) and as such field validation was conducted by determining the dominant species and/or lifeform of the polygon. Of these vegetation types, we sought to validate >10% of each map class and 15% of the total map product. The 2007 vegetation map was loaded into ArcPad on GeoXM or GeoXT Trimble GPS units, and polygons were individually updated based on overall polygon accuracy and homogeneity within the polygon. If the original polygon was deemed inaccurate for either of those reasons, a suggested vegetation type was recorded and applied to the final version of the vegetation map. We used this method to reduce the error associated with data collection by providing GPS level accuracy, and by eliminating the human error incurred when transposing data from printed maps or datasheets.

Figure 2. Sacramento River study area and the distribution of the vegetation map.



Data Analysis

To provide error estimates, error matrices or contingency analysis of field validated polygons was completed on both a polygon count and polygon area (hectares) basis. In addition to individual class accuracy, a Kappa statistic, or measure of correlation between the map data and validation data, was calculated as an indication of overall map accuracy.

Recursive Partitioning

To incorporate environmental variables into the vegetation map (Table 1), we used all digitized vegetation polygons to ascertain via zonal statistics in Spatial Analyst: ranges in floodplain age surface (1903 to 2007) (Fremier 2003) and relative elevation surface (i.e., cm above water line) (Greco et al. 2008); we then calculated the distance (meters) from polygon centroids to main river channel and levees; and measured edge complexity by calculating a normalized perimeter to area ratio by dividing the perimeter to area ratio of the polygon by the perimeter to area ratio for a circle with the same area(Figure 3 and Figure 4). Some polygons were discarded because the relative elevation surface (Greco et al. 2008) did not fully extend into our study reach.

Table 1. *Riverscape variables included in the recursive partition model.*

Variable	***Description***
LnArea	*Ln* transformed polygon area (m^2)
Normal PA Ratio	Normalized perimeter to area ratio where ((P/A(Polygon)/(P/A(Circle of same area))
Channel Distance (m)	Distance from polygon centroid to main channel (meters)
Levee Distance (m)	Distance from polygon centroid to levee (meters)
Relative Elevation (cm) Range	Range of relative elevation values (cm)
Relative Elevation (cm) Median	Median relative elevation value (cm)

Relative Elevation (cm) Minimum	Minimum relative elevation value (cm)
Floodplain Age (FPA) Range	Range of floodplain age values (years)
Floodplain Age (FPA) Median	Median floodplain age (years)
Floodplain Age (FPA) Maximum	Maximum floodplain age (years)
Height	Polygon Height Class: 1:<2m, 2:2-6m, 3:6-10m, 4:10-20m, 5:>20m.

To evaluate map errors, recursive partitioning was run for each vegetation class for 11 riverscape variables (Table 1) on the field validated dataset in JMP (SAS, Cary, NC). We maximized the split based on variable significance to the model, and independently tested model fits with coefficients of determination and cross-validation on 10% of the sample set. We also evaluated receiver operating characteristic (ROC) score to identify poor model specificity.

Figure 3. Sacramento River study area and the distribution of the floodplain age map.

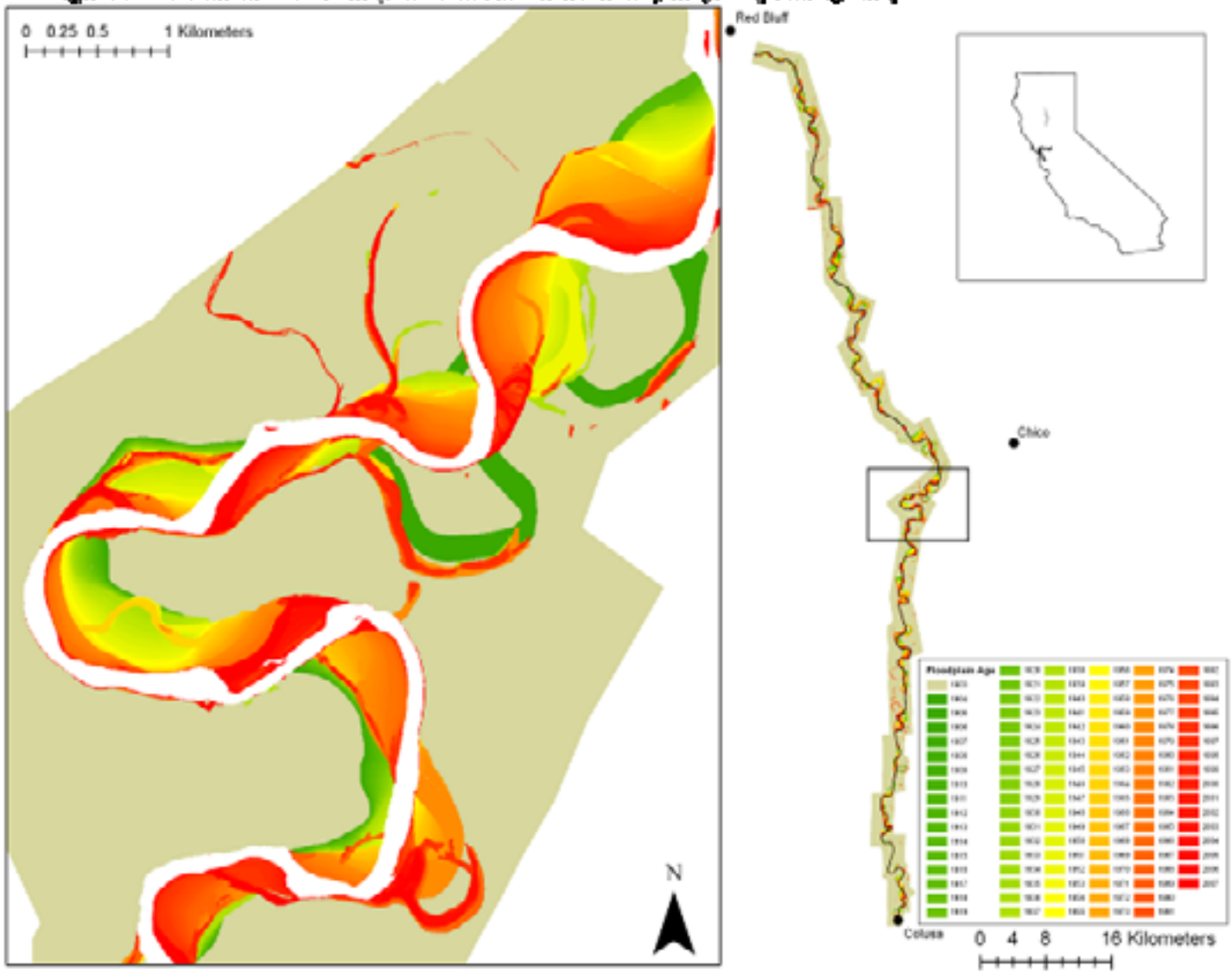
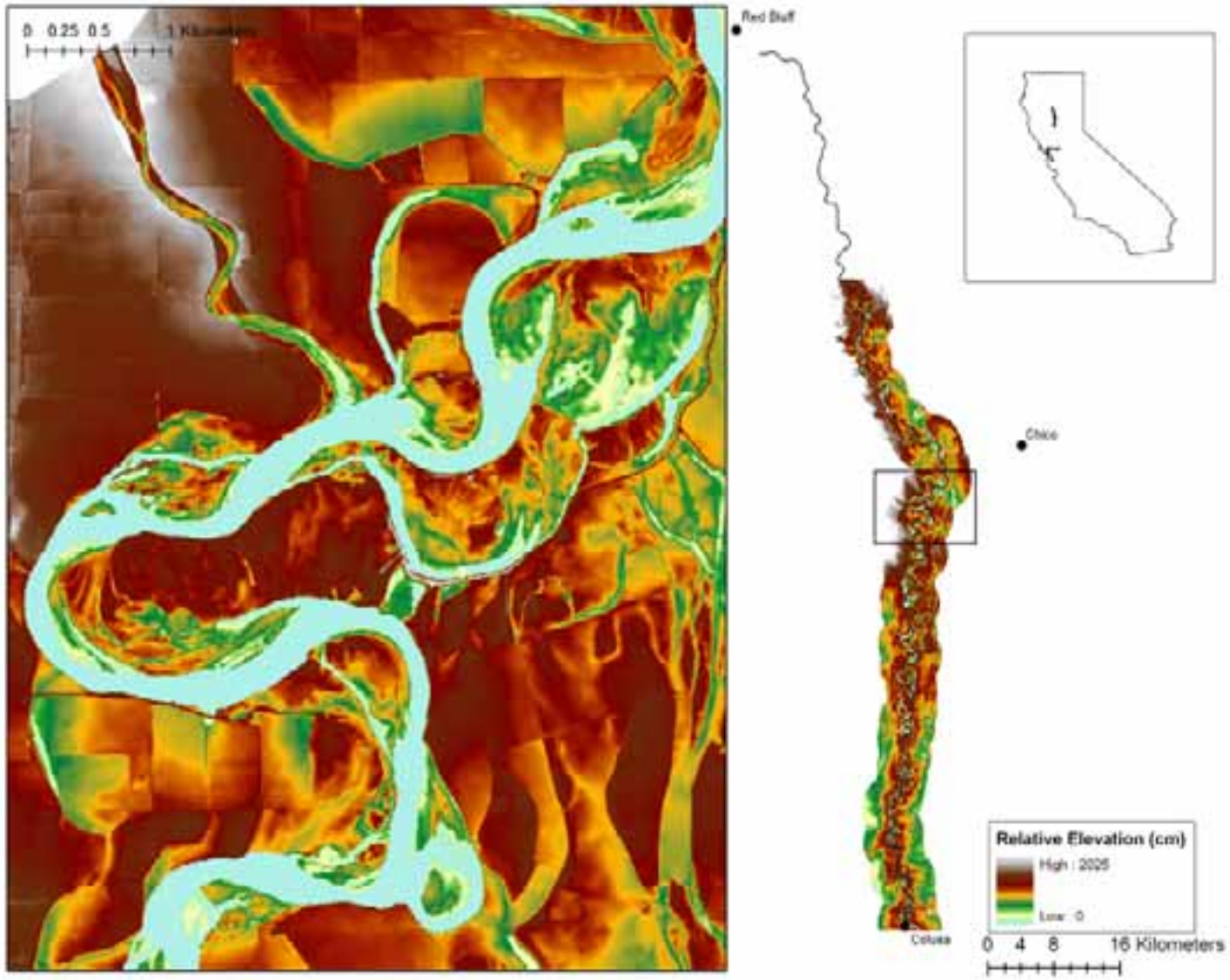


Figure 4. Sacramento River study area and the distribution of the relative elevation (cm) data.



Results

Attribute Validation

Field validation was completed for 1,198 polygons which accounted for 15% percent of all vegetation polygons and 25.4% of the total mapped area. Aquatic and herbaceous types were not targeted as aggressively as tree and shrub vegetation types, resulting in lower visitation rates for these categories (Table 2). Of field visited polygons, 85% were deemed correct (85.6% of the validated area). The Kappa statistic for the validated region was 0.83 and indicated strong agreement or correlation between the digitized and field validated datasets (Table 2). Based on these findings, we can expect that 15% of the total mapped area requires some modification and that those modifications will vary for each vegetation type (Figure 5). Vegetation types with the lowest overall accuracy included Gooding's willow, valley oak, and introduced perennials.

Table 2*. Vegetation map classes, associated map codes used in the 2007 Sacramento River Vegetation Map, and percent of polygons that were field validated.*

Code	***Vegetation Class***	***% Polygons Checked***	***% Accuracy***
BC	Bulrush/Cattail	0	n/a
BE	Box Elder	16.4	80.9
BS	Blackberry Scrub	3.9	77.7
BW	CA Walnut	11.9	76.8
CA	CA Annuals	16.4	76.0
CS	CA Sycamore	35.9	100

CW	Fremont Cottonwood	23.2	91.2
FL	Floating Leaf	10.4	25.0
GB	Gravel Bar	24.5	94.5
GW	Goodding's Willow	63.2	66.6
LP	Ludwigia peploides	5.0	100
MW	Mixed Willow	25.5	98.0
PG	Introduced Perennials	12.9	73.5
RS	Riparian Scrub	12.6	75
VO	Valley Oak	13.5	68.8

While accuracy values are variable between each map class, they rarely vary structurally (e.g. one forest vegetation type may have changed to another forest type but not necessarily to an herbaceous vegetation type). These more specific measures of accuracy can be used for in-depth study of the vegetation map, including the ability to assess land cover change over time within and between vegetation categories at large spatial scales. Valley oak (*Quercus lobata*) forest was commonly misclassified as Fremont cottonwood (*Populus fremontii ssp. fremontii*) or black walnut (*Juglans californica var. hindsii)* which may be due to the fact that both are very similar in size, leaf color, and structure, with an average difference of only 2 meters between their heights and crown radii (unpublished data).

Cottonwood forest was most commonly misclassified as valley oak forest (VO). We believe image quality issues with the digital aerial photographs may have further caused these two species, which are already very alike in both height and crown radius (unpublished field data), to become

indistinguishable from each other. California sycamore (*Platanus racemosa*) was correctly classified 100% of the time; however, it was commonly confused with valley oak (*Quercus lobata,* VO) forest and constituted almost 12% percent of the error found in the valley oak (Table 2). Valley oak and California sycamore can easily be confused in the early summer as they are a similar color at that time and have very similar crown radii. Stands of California sycamore, cottonwood and restoration areas were commonly incorrectly classified as valley oak forest. Box elder (*Acer negundo var. californicum*) vegetation types were most commonly misclassified as riparian scrub (RS) which shares a similar structure and dense canopy cover to box elder species which have an average height of 10 meters and a canopy cover of 3-55% (unpublished data).

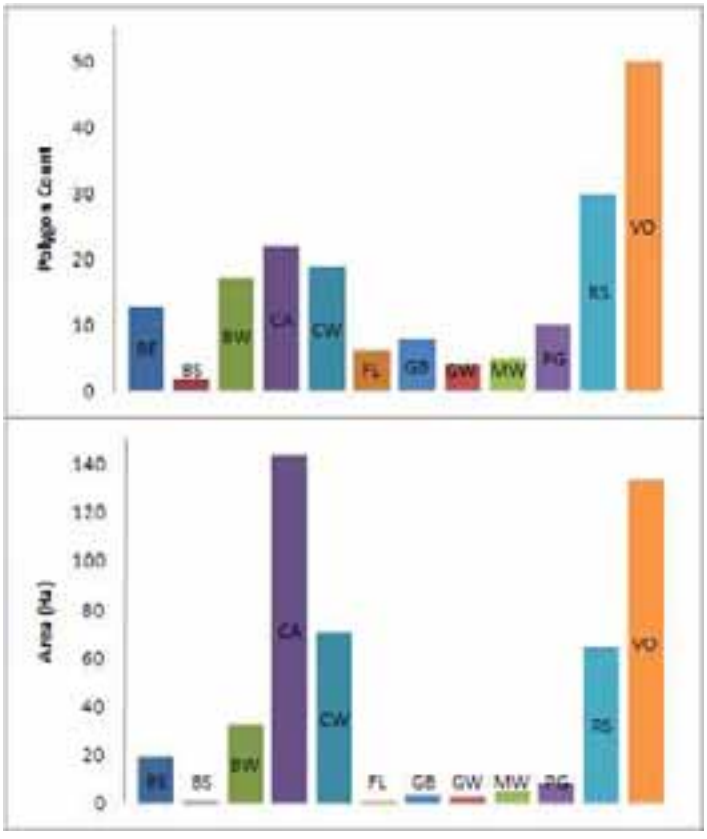


Figure 5*. Polygon count and area (ha) of inaccurate vegetation polygons based on field validation.*

Recursive partitioning

Recursive partitioning is a statistical technique for parsimonious data disaggregation that minimizes within group variance and maximizes across group separation based on the underlying independent variable matrix. The recursive partition model (RPM) identified multiple variables as significant for predicting whether a polygon had a high or low probability of being correctly classified based on our riverscape variable matrix. The use of RPM was largely successful across most of map classes, with the exception of Goodding's willow and perennial grasslands (Table 3), with each of these two classes having low coefficients of determination ($R^2 < 0.40$). Relative elevation range and floodplain age range were common predictors of the polygon accuracy, indicating that habitat complexity or a range of elevations and floodplain ages might influence the accuracy of correctly identifying a map class.

Table 3*. Likelihood of vegetation class correctness as determined by the probability value produced by the recursive partition model. Likely misclassified polygons were identified if they had a >0.5 probability of being incorrectly classified based on the model variables.*

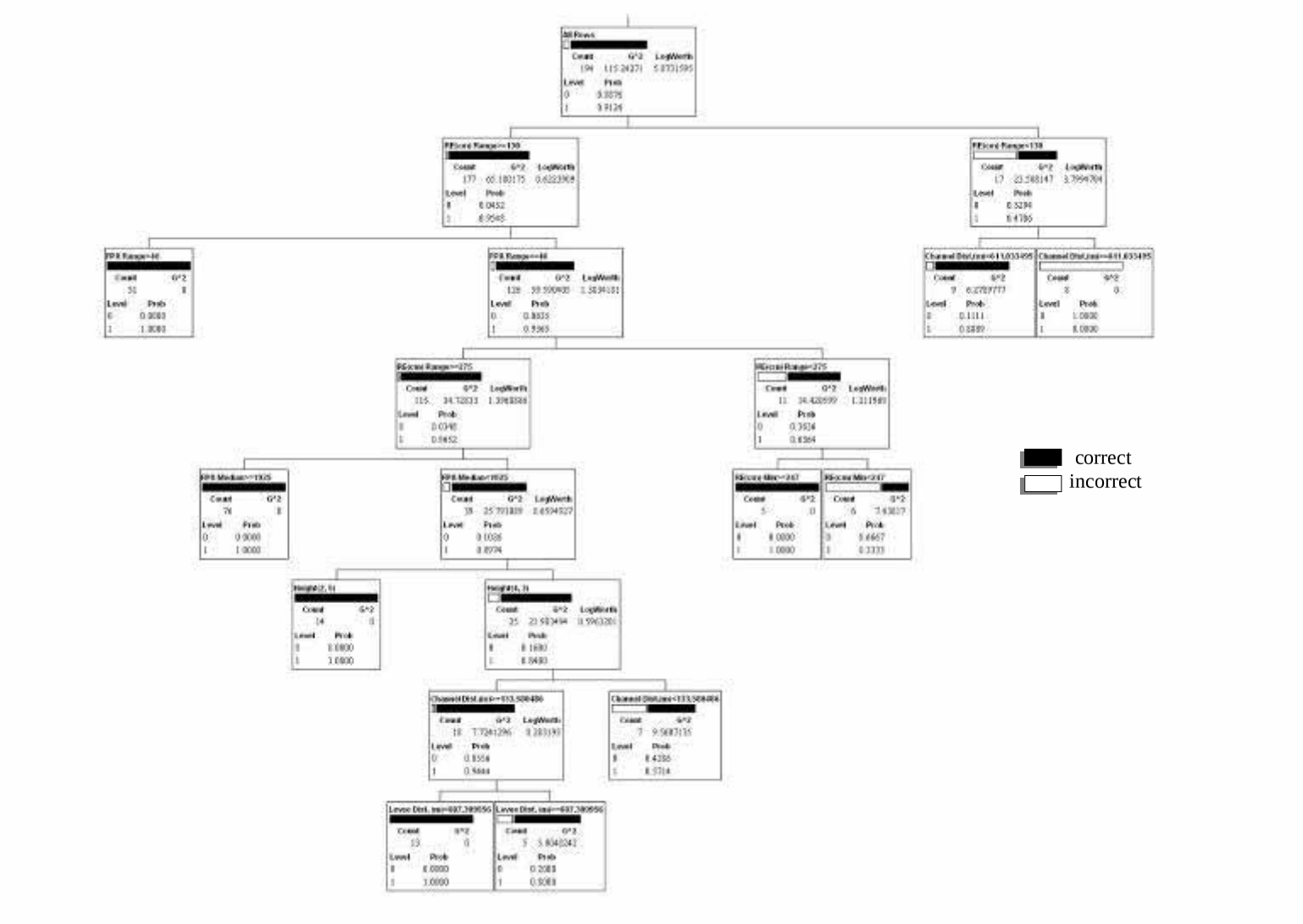
Class Code	***n (# polys)***	***Likely Correct n (%)***	***Likely Misclass n (%)***	***Likely Correct ha (%)***	***Likely Misclass ha (%)***	***R^2***	***# of Splits***	***k-fold***
BE	63	306 (79.89)	77 (20.10)	277.7 (81.96)	61.10 (18.03)	0.56	5	0.50
BW	69	478 (82.55)	101 (17.44)	733.8 (82.20)	158.8 (17.79)	0.46	7	0.27
CA	92	444 (79.42)	115 (20.57)	1195. (76.33)	370.6 (23.66)	0.65	8	0.54
CW	194	723 (86.37)	114 (13.62)	2707. (89.22)	327.0 (10.77)	0.75	9	0.69
GB	91	301 (82.01)	66 (17.98)	536.1 (86.81)	81.44 (13.18)	0.63	4	0.49
GW	12	6 (31.57)	13 (68.42)	21.08 (66.81)	10.47 (33.18)	0.37	1	0.33
MW	151	593 (100)	--	626.8 (100)	--	0.6	5	0.4

						0		9
PG	34	211 (80.22)	52 (19.77)	77.69 (74.81)	26.15 (25.18)	0.32	3	0.13
RS	96	546 (71.93)	213 (28.06)	687.1 (70.40)	288.8 (29.59)	0.46	10	0.27
VO	109	559 (69.61)	244 (30.38)	831.0 (53.86)	711.6 (46.13)	0.59	12	0.45

Mapping errors for the Box Elder (BE) vegetation class was partitioned by minimum relative elevation (> 3.0 m) and median relative elevation (> 4.8 m) having the highest error rates. In other words, polygons labeled BE on high elevations relative to the water line were unlikely to contain box elder. The Black Walnut (BW) class was first partitioned by distance to channel, with polygons close to channel edge (< 80 m) having a high error rate in classification (> 80%), and those further away further portioned by median relative elevation (< 4.0 m) and range in floodplain ages (> 100 yr) having the highest misclassification. Thus, black walnut was less likely to mapped correctly if near channel edge, or in low lying areas with high dynamism (i.e., range in floodplain age). The California annual grasses (CA) were always incorrect if the normalized perimeter to area ratio of the polygon was < 1.4, meaning that digitized polygons that were not indicative of increased edge complexity were more likely to be incorrect. If edge complexity was high in CA polygons, topographic homogeneity (RE range < 1.7 m) and recent scour (FPA median > 1993) were also predictors of high misclassification likelihood. The Cottonwood class (CW) was perhaps the best modeled RPM (R^2 = 0.75), with topographic homogeneity (RE range < 1.3 m) and distance from channel (> 611 m) having 100% mapping error. For CW polygons with greater topographic heterogeneity (RE range >= 1.3 m), a wide range of floodplain age (FPA range >= 46 yr) and low lying position (RE minimum < 2.5 m) were predictors of misclassification. One vegetation map

class with a high overall misclassification rate was Riparian Scrub (RS), with 1 in 4 polygons mislabeled by the interpreter. The RPM solution to these misclassifications pointed to young floodplains (FPA median >= 1966) and simplified edge complexity (Normal P/A Ratio < 1.86) as the most likely incorrect classification (> 85%), but complex polygons were also erroneous if near channel edge (< 202 m) and recent in age (FPA max >= 2003). Older floodplains (FPA median < 1966) were often misclassified if far from levees (distance >= 608 m) and on high terraces (RE median > 5.4 m). The final vegetation map class of interest was Valley Oak (VO), in part because of its iconic presence in the Sacramento River Valley. The RPM for VO was the most complex solution with 12 splits in its regression tree to render a 30% misclassification rate. With VO height classes not equal to "4", or moderately tall (5 is the tallest), there was a high likelihood of misclassification (12%), of which distance from channel (> 800 m) was the strongest predictor of misclassification. If the polygon height class was equal to "4", then misclassification rates increased for polygons with homogenous topographic surfaces (RE range < 3.5 m) and on younger floodplains (FPA median >= 1911). For VO polygons on heterogeneous surfaces and older floodplains, they were likely misclassified if adjacent to levees (< 153 m). Overall RPM statistics indicated that for the riparian forest classes (i.e., woody species), the average misclassification rate was 30%, equating to about 1560 hectares of the mapped area which should be revisited or reclassified in some way.

Figure 4. *Example of the Fremont cottonwood (Populus fremontii ssp. fremontii) decision tree produced by the recursive partition model. In this example, correct polygons are coded with a black bar and incorrect with a white bar.*



Discussion and Conclusion

Mapping the aerial extent of vegetation in river corridors provides valuable insight into system condition and change through time and over space. Analysis of aerial photography allows for broad scale interpretation of a landscape; however, vegetation mapping can be error prone due to the complexity of vegetation composition and structure, as well as the resolution of the imagery. Many methodological improvements are being made to reduce interpretation errors yet relatively few studies have applied associated secondarily derived data of vegetation polygons (edge metrics) or separate environmental data to quantify classification errors.

Based on the model fit statistical approach used here, we showed that this method worked well to quantify the accuracy of most of the riparian forest classes. Using environmental data improved our understanding of map error. The most consistent parameters in all recursive partitioning models were the range in relative elevation and floodplain age, proxies for topographic surface heterogeneity, river dynamism and time since disturbance. Based on these conclusions, we believe that, although there are more sophisticated methods to identify misclassification errors in categorical maps, our approach is useful for a number of important reasons: (1) to ascertain the nature of errors for potential correction (e.g., training sets to fine tune interpretation), (2) to guide map users in interpretation and utility (e.g., removing erroneous polygons from analysis), and (3) to place bounds of confidence around any change detection analyses that are computed from such maps.

In this study we explored this error analysis approach to quantify errors across the floodplains and found that although there was 15% misclassification error rate over the entire map extent, much of this error could be partitioned into certain classified groups. That is, we were able to

quantify not only the error rate by class but also the error rate by class across the selected environmental gradients. These results will improve our ability to separate land cover change from classification errors. Being able to isolate the characteristics of these errors will advance future interpretation and validation methods, as well as reduce the uncertainty about where errors occur in the current map.

Acknowledgments

Portions of this project were funded by the CALFED / Bay-Delta Science program, including the Sacramento River Mapping and Assessment Program (Award # P0620021). We would like to thank our field crew: L. Calbert, S. Lewis, J. Loyko, and C. Stouthamer, and T. Le and M. Jensen for help in the lab.

References

Congalton R.G., Birch K., Jones R. and Schriever J. 2002. Evaluating remotely sensed techniques for mapping riparian vegetation. Computers and Electronics in Agriculture 37: 113-126.

Fremier A.K. 2003. Floodplain Age Modeling Techniques to Analyze Channel Migration and Vegetation Patch Dynamics on the Sacramento River, California. Masters thesis in Geography, UC Davis.

Gergel S.E., Stange Y., Coops N.C., Johansen K. and Kirby K.R. 2007. What is the value of a good map? An example using high spatial resolution imagery to aid riparian restoration. Ecosystems 10: 688-702.

Greco S.E. and Plant R.E. 2003. Temporal mapping of riparian landscape change on the Sacramento River, miles 196-218, California, USA. Landscape Research 28: 405-426.

Greco S.E., Fremier A.K., Larsen E.W. and Plant R.E. 2007. A tool for tracking floodplain age land surface patterns on a large meandering river with applications for ecological planning and restoration design. Landscape and Urban Planning 81: 354-373.

Greco S.E., Girvetz E.H., Larsen E.W., Mann J.P., Tuil J.L. and Lowney C. 2008. Relative elevation topographic surface modelling of a large alluvial river floodplain and applications for the study and management of Riparian landscapes. Landscape Research 33: 461-486.

Greenberg J.A., Dobrowski S.Z., Ramirez C.M., Tuil J.L. and Ustin S.L. 2006. A bottom-up approach to vegetation mapping of the Lake Tahoe Basin using hyperspatial image analysis. Photogrammetric Engineering and Remote Sensing 72: 581-589.

He Y., Guo X.L. and Wilmshurst J. 2006. Studying mixed grassland ecosystems I: suitable hyperspectral vegetation indices. Canadian Journal of Remote Sensing 32: 98-107.

Hestir E.L., Khanna S., Andrew M.E., Santos M.J., Viers J.H., Greenberg J.A., Rajapakse S.S. and Ustin S.L. 2008. Identification of invasive vegetation using hyperspectral remote sensing in the California Delta ecosystem. Remote Sensing of Environment 112: 4034-4047.

Langford W.T., Gergel S.E., Dietterich T.G. and Cohen W. 2006. Map misclassification can cause large errors in landscape pattern indices: Examples from habitat fragmentation. Ecosystems 9: 474-488.

Lucas R., Bunting P., Paterson M. and Chisholm L. 2008. Classification of Australian forest communities using aerial photography, CASI and HyMap data. Remote Sensing of Environment 112: 2088-2103.

McDermid G.J., Franklin S.E. and LeDrew E.F. 2005. Remote sensing for large-area habitat mapping. Progress in Physical Geography 29: 449-474.

McGarigal K. and Marks. B.J. 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA For. Serv. Gen. Tech. Rep. PNW-351.

Nelson C., Carlson M. and Funes R. 2008. Rapid Assessment Mapping in the Sacramento River Ecological Management Zone – Colusa to Red Bluff. Sacramento River Monitoring and Assessment Program. Geographical Information Center, California State University, Chico.

Oldeland J., Dorigo W., Lieckfeld L., Lucieer A. and Jurgens N. 2010. Combining vegetation indices, constrained ordination and fuzzy classification for mapping semi-natural vegetation units from hyperspectral imagery. Remote Sensing of Environment 114: 1155-1166.

Sawyer J.O. and Keeler-Wolf T. 1995. A Manual of California Vegetation. California Native Plant Society.

Sawyer J.O., Keeler-Wolf T. and Evens J.M. 2009. A Manual of California Vegetation, 2nd ed. . California Native Plant Society.
Ustin S.L., Roberts D.A., Gamon J.A., Asner G.P. and Green R.O. 2004. Using imaging spectroscopy to study ecosystem processes and properties. Bioscience 54: 523-534.
Vaghti M. 2003. Riparian Vegetation Classification in Relation to Environmental Gradients, Sacramento River, California. Master thesis in Ecology, UC Davis.