Plant Ecol (2008) 194:37–49
DOI 10.1007/s11258-007-9273-1

ORIGINAL PAPER

Considerations for ecological reconstruction of historic vegetation: Analysis of the spatial uncertainties in the California Vegetation Type Map dataset

Maggi Kelly · Ken-ichi Ueda · Barbara Allen-Diaz

Received: 3 November 2006 / Accepted: 10 February 2007 / Published online: 8 March 2007

Abstract Historical ecological data are valuable for reconstructing early environmental and vegetation community conditions and examining change to vegetation communities and disturbance regimes over decadal and longer temporal scales, but these data are not free from error. We examine the spatial uncertainties associated with 18,000 vegetation plots in the decades-old California Vegetation Type Mapping (VTM) dataset that has been digitized for use in modern ecological analysis. We examine the relationship between plot location error and basemap year, basemap scale, plot elevation, plot slope, and general plot habitat type. Bivariate plots and classification and regression tree analysis (CART) confirm that basemap scale and age are the strongest explanation of total error. Total error in spatial location for all plots ranged from 126.9 m to 462.3 m; plots drawn on 15-min (1:62,500-scale) basemaps had total error ranging from 126 m to 199.7 m, and plots drawn on coarser-scale basemaps (1:125,000-scale) had total errors ranging from 241 m to 461.2 m. Relocation of individual VTM plots is considerably easier for plots originally marked on 1:62,500-scale maps produced after 1904, and more difficult for plots originally marked on 1:125,000-scale maps produced before 1898. Biogeographical analyses that rely less on relocating individual plots, such as environmental niche modeling or multivariate analyses can alleviate some of these concerns, but all researchers using these kinds of data need to consider errors in spatial location of plots. The paper also discusses ways in which the differing spatial error might be reported and visualized by those using the dataset, and how the data might be used in modern environmental niche models.

Keywords CART · Historical vegetation data · Spatial error · VTM

Introduction

Databases of historical ecological data have proved useful in modern ecological research for reconstructing early environmental and vegetation community conditions, studying disturbance regimes over long temporal scales, and examining change to vegetation communities over decadal and longer temporal scales. In many cases, data originally captured for taxation or land-surveying purposes often included some information on vegetation; this information has become useful to us now for reconstructing historic vegetation. One of the most often used historical databases is the

M. Kelly (✉) · K.-i. Ueda · B. Allen-Diaz
Department of Environmental Science, University of California, 137 Mulford Hall #3114, Berkeley, CA 94720-3114, USA
e-mail: mkelly@nature.berkeley.edu

Public Land Survey System (PLS), which was instituted by the US General Land Office in 1785 and partitioned US land into Townships, Ranges and Sections. At each section corner a surveyor blazed two to four trees, known as "witness trees". The species, diameter, compass bearing, and distance to the corner of each tree were recorded, and these data have become one of the most important sources for reconstructing pre-European settlement vegetation in parts of the US (Bourdo 1956; White and Mladenoff 1994; Manies and Mladenoff 2000; Manies et al. 2001; Schulte and Mladenoff 2001; Mladenoff et al. 2002; Schulte et al. 2002; Wang 2005). Wang (2005) and Mladenoff et al. (2002) provide good reviews of the data. These witness tree point data have been used in interpolations and statistical clustering routines to reconstruct pre-European settlement forest species composition and structure across the Midwest, the Southeast and Eastern US, including Wisconsin (He, Mladenoff et al. 2000; Radeloff et al. 2000; Schulte et al. 2002; Bollinger et al. 2004); in Alabama (Black et al. 2002; Rathbun and Black 2006); Minnesota (Friedman and Reich 2005); Michigan (Leahy and Pregitzer 2003); Pennsylvania (Black and Abrams 2001); West Virginia (Abrams and McCay 1996), and Southern Illinois (Anderson et al. 2006). Other researchers have used the PLS data with archeological sites to examine Native American influence on pre-European settlement vegetation patterns (Black et al. 2002; Foster et al. 2004).

Where PLS data is not available, some researchers have used less systematic historical records to reconstruct past vegetation conditions, or land use patterns (Russell 1981; Jackson et al. 2000; Ryavec 2001; Wilson 2005). For example, Wilson (2005) used 19th century lumber surveys to estimate potential forest in Maine; and Jackson et al. (2000) used land survey notes and forest resource inventories to reconstruct forest abundance in Ontario, Canada.

While valuable, historical ecological data are not free from error. Numerous researchers have discussed caveats associated with use of historical ecological data, and recommended caution in using older collections. Uncertainties can exist in both aspatial or attribute, and spatial content of the data in these collections (Schulte and Mladenoff 2001; Mladenoff et al. 2002; Plewe 2002). For example, reports cite surveyor variability in reporting vegetation characterization and species descriptions (i.e. attribute errors) (Bourdo 1956; Galatowitsch 1990; Manies et al. 2001; Schulte and Mladenoff 2001; Schulte et al. 2002) and in the reporting of the locations of both point samples, linear features, and polygonal features like habitat polygons or land parcels (i.e. positional errors) (Ryavec 2001; Gregory 2002). These uncertainties need to be considered and evaluated before the use of historic ecological data begins, and this paper deals specifically with the positional error found in a historic database.

Case study—The Wieslander vegetation type map dataset

The Wieslander Vegetation Type Map (VTM) dataset, collected originally in the 1920s and 1930s in California is a geographically broad and floristically detailed dataset, valuable for biogeographical reconstruction and vegetation change analysis. Albert E. Wieslander led the effort, funded through the Forest Service and other federal, state, and county agencies to map 16 million ha (nearly 41 million acres) of California's wildlands (Wieslander 1935; Wieslander 1961; Colwell 1977). The VTM collection consists of five components: plot data gathered at nearly 18,000 plots around the state and including floristic and environmental detail; plot maps depicting the locations of the plots sampled; vegetation maps, showing hand drawn polygons of forest type, and their associated species; landscape photographs and associated information about location and content of the photographs; and herbarium specimens for every species recorded on the vegetation maps or in the sample plots (Ertter 2000). The collection is a data-rich resource for foresters, ecologists, land managers, and others interested in the environment of early 20th century California, and land use changes since then. The entire VTM collection (with the exception of the herbarium specimens) has been digitized and made available via the Internet for use in modern ecological and geospatial analysis (Kelly et al. 2005).

The geographic extent (the maps cover nearly 16 million ha (nearly 41 million acres) of the state) and floristic detail found in the VTM plot data have been compelling reasons for ecologists to use the data in numerous applications. For example, the VTM plot data have been used to create classification schemes for vegetation communities in the state (Allen 1989; Allen et al. 1991; Allen-Diaz and Holzman 1991; Griffin and Critchfield 1972; Jensen 1947) and to validate modern models of vegetation composition (Franklin 2002; Vayssières et al. 2000). In addition, researchers have used the data to reconstruct California vegetation community conditions in the early part of the 20th century in order to examine changes as a result of disturbances such as development, fire, and disease (Bradbury 1974; Minnich et al. 1995; Walker 2000; Franklin 2002).

Despite the broad geographic coverage, large sample size and floristic detail found in the VTM data, there are methodological challenges associated with the original collection that researchers should consider, including the lack of permanent ground marking, the large symbol used to mark plot location on basemaps, the lack of precise surveying equipment available to the samplers, the non-random location of the plots, and the lack of direct measurement of individual tree diameters or of within-stand variance (Garrison, In Review). These combine to make absolute relocation of all original VTM plots unlikely, and approximate relocation challenging.

Experiences vary in published work describing VTM plot relocation. Allen-Diaz and Holzman (1991) estimated <50 m relocation accuracy in blue oak woodlands using VTM map comparisons, photographs, and other geographic cues (Allen-Diaz and Holzman 1991). Minnich et al. (1995) reported relocation accuracy at under 100 m in conifer forests by reference to fixed features such as roads, and the location of prominent trees include in the original dataset, and described a protocol with many samples in the area of the plot to compensate for uncertainty. In contrast, Keeley (2004) found that the spatial variability in coastal sage scrub and chaparral communities in Southern California was too great to perform plot-by-plot analysis of community change. Others have commented on the relocation issue. Walker (2000) found that while the overall fidelity of species composition mapping was good, the spatial accuracy of the plots and vegetation map polygons varied with topography, and spatial error increased in areas of high terrain. Franklin (2002) found discrepancies between the VTM data and the more modern USDA Forest Service Forest Inventory and Assessment (FIA) plots over a large area, but proposes that this might be a result of differing sampling schemes between the VTM and FIA plots, not as a result of spatial accuracy.

These inherent errors are to be expected due to the mapping technologies available to early 20th century surveyors, and our digitization process (involving scanning and georeferencing maps and plot locations), which introduces additional spatial uncertainty to the dataset. We consider these experiences, and in this paper present a summary of the error found in the digital plot database that results from all sources: error inherent in the database and that error introduced by the digitization process. We examine error trends by basemap year, basemap scale, plot elevation, slope, and general habitat type in order to provide information for scientists and managers using the database to make judgments about the likelihood of finding individual VTM plots for historical ecological analysis. We also discuss ways in which the differing spatial error might be reported and visualized by those using the VTM webGIS system designed to allow access to the dataset. And finally, we discuss uses for the VTM data that rely less on spatial accuracy of individual plot location.

Methods

Dataset

There are nearly 18,000 VTM sample plots in California, primarily concentrated along the central and southern coastal ranges, and along the Sierra Nevada Mountains (Fig. 1), covering nearly two-fifths of the total area of the state, and much of its wild area exclusive of the deserts. The sample plots were located by the VTM crews

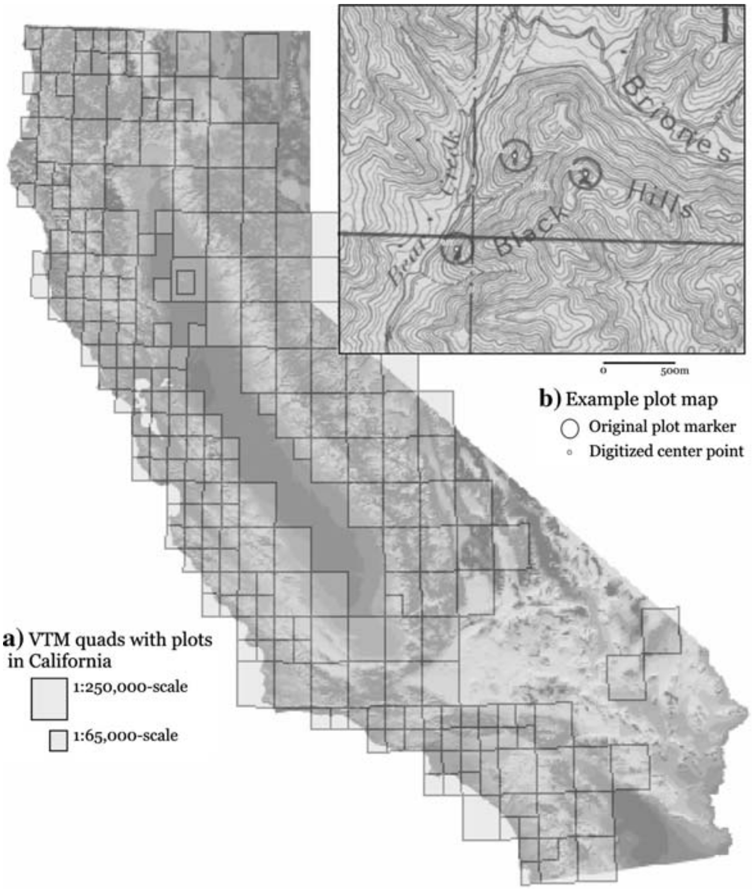


Fig. 1 (**a**) VTM quadrangles in California with plot data, (**b**) an example of a basemap with plots marked: in this field the outer, larger circle is the original plot marker and the smaller mid-point is our digitization of the center of the plot

across a gradient of vegetation types, and the historic records contain data regarding tree stand structure (number of trees per diameter class), percent cover of dominant overstory and understory vegetation by species, soil type, parent material, leaf litter, elevation, slope, aspect, parent material, and other environmental variables (Kelly et al. 2005). Each plot was 1/5th acre (0.08 ha) in size (800 m^2) in forests, and 1/10th acre (0.04 ha) in scrub and chaparral communities (Wieslander 1935). All the plot data were stored on paper data sheets, and individual plots were numbered according to USGS quad name, quad section number, and plot number. The plot locations were stamped with hollow circles of approximately 3.5 mm in diameter (1.75 mm radius) in red ink on nearly 150 15-min (1:62,500-scale) and 30-min (1:125,000-scale) United States Geological Survey (USGS) quadrangles (Fig. 1). These maps were produced by the USGS from 1890 through 1946 (Varanka 2006). The map sheets had been previously cut into sections, mounted on canvas, and folded, to facilitate use in the field. This allowed for repeated folding along the seams, without loss of mapped information.

Digitization and georeferencing of plot maps

We digitized (scanned) and georeferenced the plot data and associated plot maps to protect the original collection from further field use and to allow more researchers to utilize the collection. Georeferencing is the process of registering one map or image to real-world coordinates using a series of tie-points common to the image with known coordinate system (the ‘‘Master’’ image) and the image that needs registering (the ‘‘Slave’’ image). These tie-points can either be feature-based (mapped features common to both maps), or marked map coordinates (map corners or latitude/longitude grid marks), and they are used to calculate a mathematical transformation that warps the slave image. Our digital georeferencing process utilized image-to-image registration using modern DRGs (Digital Raster Graphics) as reference maps. Others have tried different approaches, for example, Walker (2000) used modern satellite imagery as the reference map in the georeferencing process. We decided to use modern DRGs because of their statewide coverage and the abundance of feature-based tie points. We strove to use feature-based tie points for image-to-image registration, because marked map coordinate points can be troublesome as projections and datums are not always reliably reported on early 20th century maps. In addition, some features, such as roads and road intersections represent some of the most consistent surveyed features on the map that persist from the date of the original maps to the present. When feature-based tie-points were not available or insufficient, we used map coordinate points.

We performed digital image-to-image georeferencing using Erdas Imagine 8.7 software (Leica, 2004), and ArcGIS software (ESRI, 2004). We first scanned each plot map section at 600 dpi. Next, digital scanned and uncut versions of each USGS topographic map of the same edition and reprint were acquired from map libraries in California, and modern 1:24,000-scale USGS Digital Raster Graphics (DRG) of quads corresponding to each VTM quad were also collected. The uncut historic topographic maps were registered to the modern USGS DRGs using stable feature-based tie points. The VTM plot map sections were then registered to the uncut topographic maps using each map's unique Polyconic projection (the projection was different for each quad, as each USGS map from the early 20th century was projected with an origin at the center of the map (USGS 1928)). In this second step common tie points from both maps were used. We used first order polynomial transformations for each step.

Total error for each step was calculated for each quad using a series of independent checkpoints that did not form a part of the rectification model. Each point on the plot map segments was then hand-digitized with a high zoom factor (1:10,000 or greater) in ArcMap by creating a point at the center of the red circle indicating each plot point. The resulting quad's plots points were re-projected to Albers California projection system and stored as a shapefile.

Plot elevation, slope, and general habitat type were derived by overlaying the final plot location shapefile with a California Digital Elevation Model and derived slope (90 m resolution), and a previously digitized 1:100,000-scale habitat map produced by the VTM project in 1945. Habitat types included barren, forest, chaparral and sagebrush, cultivated, urban and industrial, desert, and grasslands. Map age was calculated by subtracting the basemap's edition date from 2006.

Since the size of the plot stamp varied slightly across the collection, we averaged the radii from the center to the northern edge of the plot circle from 20 randomly selected plots on 15-minute quads and 20 randomly selected plots on 30-min quads.

Quantification of total error

Error accumulates through the digitization process: total propagated error includes the error associated with the basemaps used combined with the error introduced in the process. We account for six sources of error in this process; the first three result from the initial maps and collection process and are constants based on map scale, and the last three result from the digitization process and are different for each map quad:

e_1 = Error associated with the historic basemaps (constant for each map scale, assumed to adhere to NMAS 1947);

e_2 = Error associated with the Surveyor's plot marker (the scale-dependent radius (1.75 mm) of the red marker (for 1:62,500-scale maps, this is 112.5 m, and 215.7 m for 1:125,000-scale basemaps);

e_3 = Error associated with the modern DRG (derived from NMAS standards),

e_4 = Error associated with the registration of the basemap (RMSE per quad, reported from the georeferencing process);

e_5 = Error associated when the section map is registered (RMSE per section, reported from the georeferencing process); and

e_6 = Error associated with the analyst locating the plot in the center of the marked circle (derived from sample tests);

We used the formula for calculating total error propagation provided by Thapa and Bossler (1992) and cited elsewhere (Wieczorek et al. 2004). We assume no functional relationship exists between individual errors, and we assume that a linear relationship exists between the total error and the individual errors. By adopting these assumptions we contend there is no reason to assume that all the errors necessarily compound each other, or that they are all necessarily in the same direction. This seems reasonable because these errors are independent of each other, and from numerous sources. We then calculated total error by using the law of propagation of errors:

$$\text{Total error} = (e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 + e_6^2)^{1/2}.$$

The total error associated with the digitization process was then provided for the user as a fields labeled "error" in the point shapefile, and provided as an "error buffer" in our on-line webGIS application (vtm.berkeley.edu).

We analyzed the total error from 16,686 plot points, covering about two-fifths of California (calculated by the amount of quad coverage) on nearly 150 basemaps. The year of basemap edition ranged from 1892 to 1936: all were USGS topographic quadrangles of either 15-min (1:62,500-scale) or 30-min (1:125,000-scale) quads (Fig. 1). Data for each plot (total error, basemap age, basemap scale, plot elevation, plot slope, and plot habitat type) was compiled in an Excel file for use in *R* statistics package. Data summaries by scale are provided in Table 1.

Table 1 Summary information of plot data by basemap scale

	1:62,500-scale (15-min) n = 5,564	1:125,000 (30-min) n = 11,122	Both scales combined n = 16,686
Elevation range (m)	2–2,890	7–3,477	2–3,477
Slope range (%)	0.1–71	0.1–79.7	0.1–79.7
Basemap age (year)	70–109	71–114	70–114
Basemap edition date	1897–1936	1892–1935	1892–1936

Analysis of error

In addition to calculating standard descriptive statistics of total error and creating bivariate regression plots of total error and all explanatory factors (e.g. map age, elevation, slope, and habitat of the plot) we used a non-parametric multivariate technique, Classification and Regression Trees (CART), to analyze how these factors influenced total error. CART is increasing in popularity among researchers analyzing multivariate data, as it requires no advance variable selection, its results are invariant to transformations such as log transforms, it can use any combination of categorical and continuous predictor variables. It can handle missing data (Feldesman 2002), and it has the ability to capture hierarchical and non-linear relationships and expose interactions among predictor variables (Clark and Pregibon 1993; Michaelsen et al. 1994; De'ath and Fabricius 2000; Kelly and Meentemeyer 2002). The tree models are developed by recursively partitioning the response variable (here total error) into increasingly homogeneous binary subsets based on critical thresholds in predictor variables. The split chosen is the one that most reduces the average impurity in the resulting bins (Breiman et al. 1984; De'ath and Fabricius 2000; Venables and Ripley 2002). The resulting "trees" are often displayed graphically, and are easy to understand as a series of if/then conditions, but they can be complex to render cartographically (Muñoz and Felicísimo 2004). All statistical analysis was performed using *R* stats package (R_Statistics 2006). We created trees separately for the plot locations found on 1:62,500-scale basemaps (n = 5,577); and using data associated with 1:125,000-scale basemaps (n = 11,134).

Results

Average total error for all maps was 232.4 m, and ranged from 126.9 to 462.3 m (Table 2). Not surprisingly, for the 15-min (1:62,500-scale) basemaps the errors were smaller (ranging from 126 m to 199.7 m) and larger for the coarser-scale 1:125,000-scale basemaps (ranging from 241 m to 461.2 m). There is a clear, although not statistically significant, trend describing smaller error on newer finer-scale basemaps and larger on older course-scale basemaps (Fig. 2) (R-squared values are 0.3571 for the course-scale maps (n = 11,134) and 0.1342 for the fine-scale basemaps (n = 5,577)).

The largest error was found in the Placerville quad, which is one of the older basemaps in the collection (1893 edition), covering mountainous terrain; the smallest error was found in the Petaluma quad, one of the youngest basemaps (1914 edition), and covering less rugged terrain. The relative contributions from each of the six error sources are listed in Table 3: in the Petaluma quad case the largest contribution is the plot marker size, in the Placerville quad, the largest contribution comes from the registration of the old base map.

Since basemap scale had the largest influence on error, we performed CART analysis on the data separated into two classes by scale: plot locations on 1:62,500-scale basemaps (n = 5,577) and on 1:125,000-scale basemaps (n = 11,134). The tree produced from data associated with the finer-scale maps was a weaker predictor, with a 57% classification rate; while the tree produced from coarser-scale basemaps was a stronger predictor (78% classification rate). The CART

Table 2 Total error (calculated according to the law of propagation) statistics for each scale class, and both scales combined

	1:62,500-scale (15-min) n = 5,564	1:125,000 (30-min) n = 11,122	Both scales combined n = 16,686
Minimum error (m)	126.0	241.2	126.0
Maximum error (m)	199.7	461.2	461.2
Average error (m)	138.3	279.6	232.4

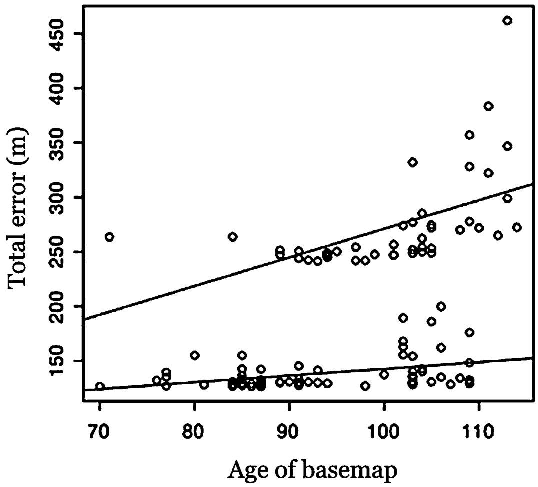


Fig. 2 Relationship between total error and age of basemaps; regression lines fitted through 1:62,500-scale data (lower line) and 1:125,000-scale data (upper line)

Table 3 Example of map error from the quad with the smallest and largest total error

Error source	Error (m)	
	Petaluma Quad 1:62,500-scale (15-min) 1914 Edition	Placerville Quad 1:125,000 (30-min) 1893 Edition
E_1: Historic basemap errors	52.91	105.83
E_2: Plot marker size	112.46	215.73
E_3: Modern DRG	20.32	20.32
E_4: Registration of basemap	14.94	393.70
E_5: Registration of map section	4.30	23.42
E_6: Location of plot in center of marked circle	1.70	3.94
Total	126.90	462.29

results help to describe possible influences on total error (Fig. 3). The trees should be read from top down, each split subdivides the remaining data into homogeneous groups, and can be thought of as a series of greater than/less than splits. The number above each split is the value of the split, the numbers at the end of the tree branch (called a ‘‘node’’) are the mean error for the group (upper value), and the number of observations in that group (lower value). Plot habitat and slope had no consistent influence on error (because these variables either do not appear on the tree, or in inconsistent ways), but there are trends with respect to map scale, age, and elevation revealed by the analysis. The most important explanatory factor is map scale: the error on 1:62,500-scale basemaps is always less than that found on 1:125,000-scale basemaps. Once scale is removed as a factor, basemap age has the strongest influence on total error. For all plots, maps over 100-years-old provide the most error: the threshold for the finer-scale maps is 1905 (101-years-old), and for the coarser-scale basemaps is 1898 (108-years-old). Plots drawn on maps older than these have significant error in plot location. For plots drawn on 1:62500-scale basemaps with editions of 1905 and later all have similar total error (132 m average), and no other factors (elevation, slope, or habitat) help us understand variations in error. For the older (pre-1905) 1:62,500-scale basemaps, plots with the worst error are on slopes less than 48%, and on

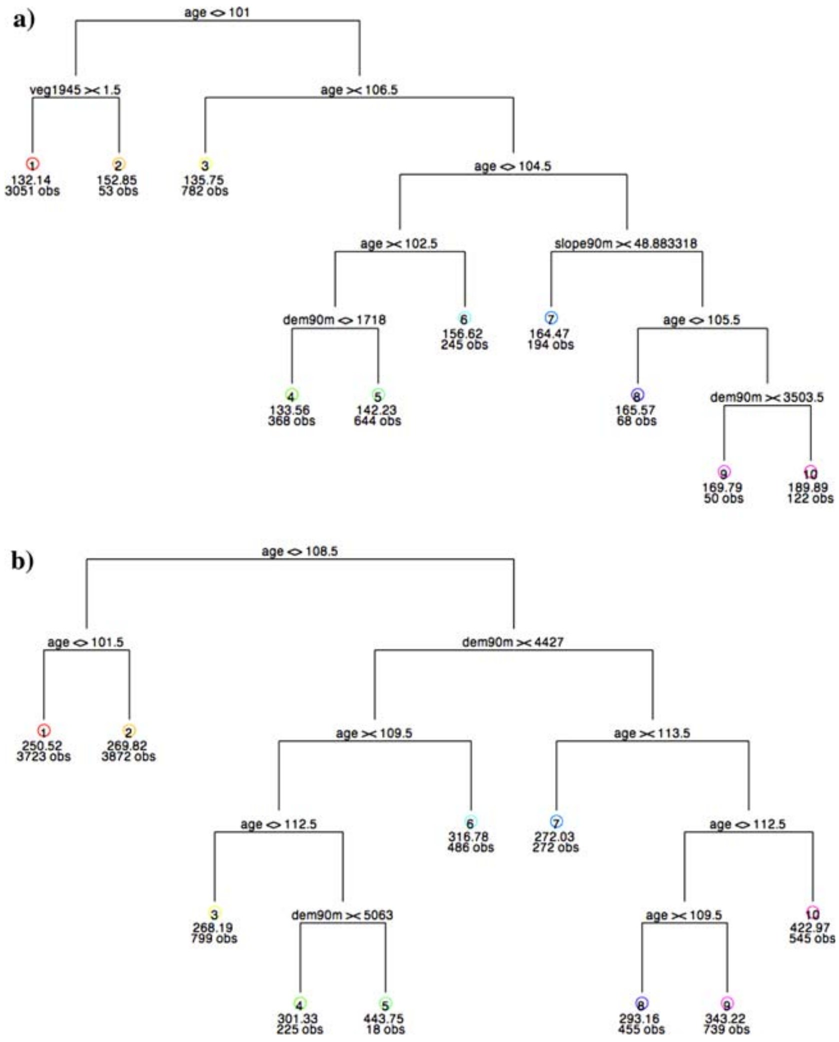


Fig. 3 Classification tree results: (**a**) using data from 1:62,500-scale basemaps (n = 5,577); and (**b**) using data from 1:125,000-scale basemaps (n = 11,134). The trees should be read from the top down, the text above each split indicates the value that separates the data into less than (<) and greater than (>) homogeneous groups. At the end of the lines (or ‘‘branches’’) are ‘‘nodes’’: these are homogeneous groups of data. These nodes are numbered with an integer in a circle; below the node number is the mean error of the group, and the number of observations in the group. Variables include: age = age of basemap; dem90 = elevation derived from 90 m DEM; slope90 = slope derived from 90 m DEM

elevation less than 3,503 ft (1067 m), but these are neither strong nor consistent relationships.

Of the plots located on coarser-scale basemaps, editions from 1905 and later produced the best total error numbers (250.5 m average); plots on basemaps produced between 1898 and 1904 had slightly more error (269 m). The worst errors on coarse-scale basemaps were on pre-1898 edition basemaps, and of this group, errors increased on elevations between 4,427 and 5,063 ft (1349 and 1543 m), but the relationship between elevation and total error is again not consistent. Plot slope and general habitat are not factors in explaining error (Fig. 3).

Discussion

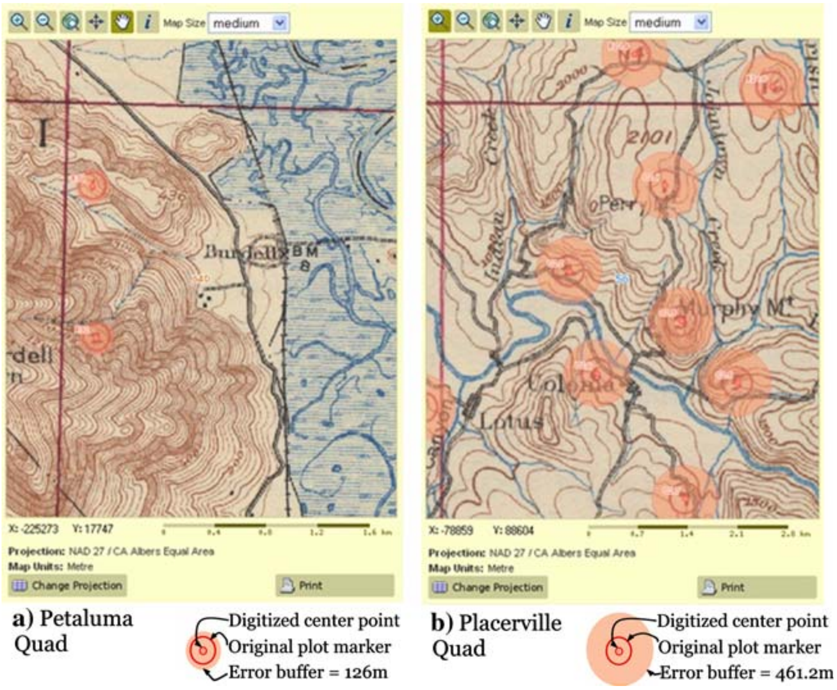
Plot relocation considerations

The use of the VTM plot data in modern ecological analysis within a digital framework has several advantages. First, our webGIS system (vtm.berkeley.edu) allows researchers to search, query and, download species and location data from plots across California knowing a consistent digitization and georeferencing process has taken place. Second, we have tools to assist researchers in plot relocation, including a protocol-describing plot data download from our database and plot location upload to a Global Positioning System (GPS) for field relocation. We also provide the total error as a field in the point file, so that users can create their own error for each plot point; the digitized location of the plot location and the error buffer are available on our webGIS application (vtm.berkeley.edu) for researcher query and error visualization (Kelly et al. 2005). Finally, the digital database saves the collection from repeated field use and damage: the original maps and plot data cards in the VTM collection are now stored in the UC Berkeley Koshland Biosciences and Natural Resources Library.

Despite these advantages, the larger spatial location errors found in the database make it challenging for researchers to quickly and absolutely relocate individual plot sites given a latitude/longitude point and a search radius alone. For example, researchers relocating plots in the Petaluma quad (which had the smallest overall error) would have to cover about 5 ha (circle with radius = 126.9 m), an area which is only slightly larger than that covered by the marked circle on the original map (about 4 ha). In contrast, researchers interested in the plots on the Placerville quad will have a tremendous challenge: the search area for each plot is 67 ha (radius = 462 m). These contrasting cases are shown in Fig. 4; in the Placerville quad, the search area is considerably greater than original marked circle, largely due to the problems encountered registering pre-1900 maps. The graphics in Fig. 4 are from our webGIS site, demonstrating how researchers can quickly visualize the error associated with their plots while searching the database prior to download.

This work clearly suggests that relocation of individual plots will be easier for plots marked on 1:62,500-scale basemaps produced later than 1904. In these cases, average error is about 132 m, and can be as low as 127 m. The pre-1900 maps have less spatial fidelity than later versions because often the cartographers relied on planetable surveying, with transportation by mule pack train. Regardless of basemap age, plots marked on 1:62,500-scale maps always have less error than those on their coarser-scale counterparts. When plots are marked on 1:125,000-scale basemaps, researchers will encounter less error with plots marked on basemaps produced from 1904 onwards. While Walker (2000) suggested a relationship at the quad scale between plot location and topography, we found no consistent relationship between plot elevation, slope, or general habitat type and location error at the regional scale.

There are additional methods that can be employed to assist researchers to focus the search for plots within the larger search area. Geographical inferences can be made about plot location from the original plot maps; patterns of drainages and slope on the original map can give insight into where the VTM crews placed the plot marker on that map. For example, Fig. 5 shows a plot in the Concord quad (plot number 82 AC26): this plot has a total error of 145 m. The original placement of the plot on the 1915-edition map is clearly in the first drainage southeast of the confluence of



Map Size: medium

X: -225273 Y: 17747

Projection: NAD 27 / CA Albers Equal Area

Map Units: Metre

Map Size: medium

X: -78859 Y: 88604

Projection: NAD 27 / CA Albers Equal Area

Map Units: Metre

Fig. 4 Examples of error buffers from our webGIS site for researcher search, query, and download of VTM data: (**a**) Petaluma quad (lowest total error = 126.0 m); (**b**) Placerville quad (highest total error = 461.2 m)

Redwood creek (the northerly waterway) and the other waterway (not named). On the digitized version, the plot is near the ridge between the first and second drainages southeast of the same confluence. The actual plot location is likely located more to the northwest on the modern map, toward the edge of the error buffer, but likely within it. With print-outs of the original map and the modern DRG, with the error buffer superimposed on both (available from the website), and slope and aspect, a researcher will be more able to relocate the original plot.

Other uses for VTM data

Alternative uses of the VTM data lessen the reliance on spatial accuracies of individual sites. For example, some researchers have conducted sub-sampling around individual sites to capture local spatial variability in community structure (Minnich et al. 1995). Alternatively, the VTM data have been used in analyses that do not require the spatial location of individual plots. For example, Allen-Diaz and Holzman (1991) used the dataset to examine vegetation community structure through multidimensional statistical analysis (Allen-Diaz and Holzman 1991), and Keeley (2004) conducted meta-scale analyses to examine vegetation change at regional scales (Keeley 2004). Moreover, the data can be used as samples of presence data in predictive modeling of species distributions. While not distributed randomly, the geographic coverage, and total number of VTM plots make these data useful in newer non-parametric predictive habitat distribution modeling techniques such as Support Vector Machines (Guo et al. 2005). Predictive habitat distribution models, sometimes called environmental niche models (ENM), are increasingly recognized as important tools that can support our understanding of historical habitats and climate change impacts (Iverson and Prasad 1998; Clark et al. 2001) and biodiversity patterns (Rushton et al. 2004; Graham et al. 2004). The SVM algorithm seeks to find an optimal series of ‘‘hyper-planes’’ around clusters of presence training points in multidimensional space (Cristianini and Scholkopf 2002; Huang et al. 2002). The multidimensional class description can then be used to map a niche across a landscape. The SVM algorithm is non-parametric, it is able to handle non-linear and categorical data, and makes no assumption about the probability density of the data, and techniques such as this will be useful for

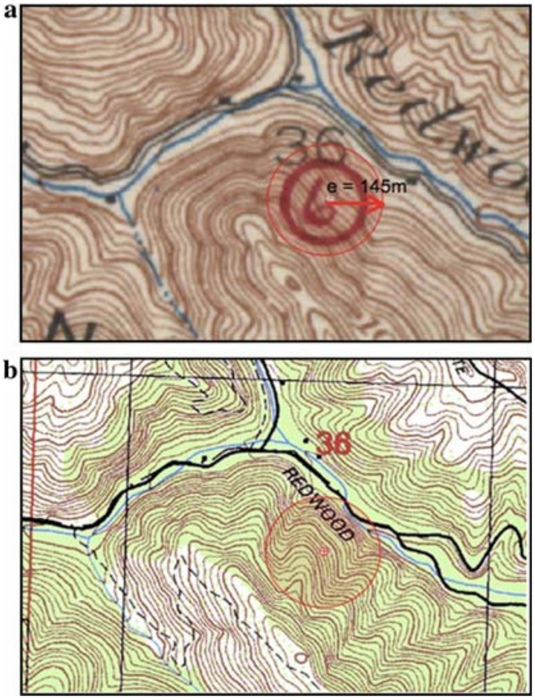


Fig. 5 Plot location and error buffer for a plot in Concord quad: (**a**) plot location and error buffer on original basemap; and (**b**) plot location and error buffer on modern DRG

predicting species distributions in the 1920s and 1930s in order to examine potential impacts of climate change on valuable California habitats.

Conclusions

Data from historical collections describing vegetation are increasingly used in modern ecological analyses to examine vegetation community structure and long-term change. The VTM collection is one such source of data, which has been used recently in biogeographic research. The digitization makes this geographically broad and floristically rich dataset available to researcher via the Internet, but there are considerable spatial uncertainties that must be acknowledged and considered. Relocation of individual VTM plots is considerably easier for plots originally marked on 1:62,500-scale maps produced after 1904, and more difficult for plots originally marked on 1:125,000-scale maps produced before 1898. Plot slope and general habitat type have no clear influence on plot location error, and there are no consistent trends in error across California that can be accounted for plot elevation. In biogeographical analyses that rely less on relocating individual plots, such as environmental niche modeling or multivariate analyses, these spatial errors are not as critical, but all researchers using these kinds of data need to consider error. To assist this, we have developed a webGIS site for researchers to search for, download and visualize the error associated with VTM plots to support evaluation and use of these data.

Acknowledgments We would like to acknowledge the following people for their help in digitization and georeferencing (in alphabetical order): Tim DeChant, Tim Doherty, Ling He, Ann Huber, Desheng Liu, Lauren McGee, Susan McIllroy, Mindy Syfert, Karin Tuxen and Esther Zeledon. This work was funded by USDA Forest Service PNW and PSW, and USDA-CSREES.

References

Abrams MD, McCay DM (1996) Vegetation site relationships of witness trees (1780–1856) in the presettlement forests of eastern West Virginia. Can J Forest Res 26(2):217–224

Allen BH (1989) A classification system for California's hardwood rangelands. University of California Division of Agriculture and Natural Resources, Davis

Allen BH, Holzman BA, Evett RR (1991) A classification system for California's hardwood rangelands. Hilgardia 59(2):1–45

Allen-Diaz BH, Holzman BA (1991) Blue oak communities in California. Madroño 38:80–95

Anderson RC, Jones SL et al (2006) Modifying distance methods to improve estimates of historical tree density from General Land Office survey records. J Torrey Bot Soc 133(3):449–459

Black BA, Abrams MD (2001) Influences of native Americans and surveyor biases on metes and bounds witness-tree distribution. Ecology 82(9):2574–2586

Black BA, Foster HT et al (2002) Combining environmentally dependent and independent analyses of witness tree data in East-Central Alabama. Can J Forest Res 32(11):2060–2075

Bollinger J, Schulte LA et al (2004) Assessing the ecological restoration potentials of Wilsoconsin (USA) using historical landscape reconstructions. Restor Ecol 12(1):124–142

Bourdo EA (1956) A review of the general land office survey and of its use in quantitative studies of former forests. Ecology 37(4):754–768

Bradbury D (1974) Vegetation history of the Ramona quadrangle, San Diego County, California (1931–1972). University of California

Breiman L, Friedman JH, et al (1984) Classification and regression trees. Chapman and Hall, New York

Clark LA, Pregibon D (1993) Tree-based models. Statistical models. Chambers JM and Hastie TJ. Chapman & Hall, Inc., London

Clark ME, Rose KA, Levine DA, Hardgrove WW (2001) Predicting climate change effects on appalachian trout: combining GIS and individual-based modeling. Ecol Appl 11(1):161–178

Colwell WL (1977) The status of vegetation mapping in California today. John Wiley & Sons, Sacramento

Cristianini N, Scholkopf B (2002) Support vector machines and kernel methods–The new generation of learning machines. Ai Mag 23(3):31–41

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192

Ertter B (2000) Our undiscovered heritage: past and future prospects for species-level botanical inventory. Madroño 47(4):237–252

ESRI (2004) Environmental Systems Research Institute, Inc., ArcGIS software. Redlands, CA

Feldesman MR (2002) Classification trees as an alternative to linear discriminant analysis. American J Phys Anthropol 119:257–275

Foster HT, Black B et al (2004) A witness tree analysis of the effects of Native American Indians on the pre-European settlements forests in east-central Alabama. Human Ecol 32(1):27–47

Franklin J (2002) Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. Appl Vegeta Sci 5: 133–146

Friedman SK, Reich PB (2005) Regional legacies of logging: departure from presettlement forest conditions in northern Minnesota. Ecol Appl 15(2):726–744

Galatowitsch SM (1990) Using the original land survey notes to reconstruct presettlement landscapes in the American West. Great Basin Nat 50(2):181–191

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 19(9):497–503

Gregory IN (2002) Time-variant GIS databases of changing historical administrative boundaries: A European comparison. Transactions in GIS 6(2):161–178

Griffin JR, Critchfield WB (1972) The distribution of forest trees in California (No. PSW-82). U.S.D.A. Forest Service, Pacific Southwest Forest and Range Experiment Station, Berkeley, CA

Guo Q, Kelly M et al (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol Model 128(1):75–90

He HS, Mladenoff DJ et al (2000) GIS interpolations of witness tree records (1839–1866) for northern Wisconsin at multiple scales. J Biogeograph 27:1031–1042

Huang C, Davis LS et al (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens 23(4):725–749

Iverson LR, Prasad AM (1998) Predicting abundance of 80 tree species following climate change in the Eastern United States. Ecol Monogr 68(4):465–485

Jackson SM, Pinto F et al (2000) A comparison of pre-European settlement (1857) and current (1981–1995) forest composition in central Ontario. Can J Forest Res 30(4):605–612

Jensen HA (1947) A system for classifying vegetation in California. Calif Fish Game 33:199–266

Keeley JE (2004) VTM plots as evidence of historical change: goldmine or landmine? Madroño 51(4):372–378

Kelly M, Allen-Diaz B et al (2005) Digitization of a historic dataset: the Wieslander California vegetation type mapping project. Madroño 52(3):191–201

Kelly M, Meentemeyer RK (2002) Landscape dynamics of the spread of Sudden Oak Death. Photogram Eng Remote Sens 68(10):1001–1009

Leahy MJ, Pregitzer KS (2003) A comparison of presettlement and present-day forests in northeastern lower Michigan. American Midland Nat 149(1):71–89

Leica (2004) Erdas Imagine 8.7 software. Leica Geosystems, Inc., Heerbrugg, Switzerland

Manies KL, Mladenoff DJ (2000) Testing methods to produce landscape-scale presettlement vegetation maps from the U.S. public land survey records. Landscape Ecol 15:741–754

Manies KL, Mladenoff DJ et al (2001) Assessing large-scale surveyor variability in the historic forest data of the original US public land survey. Can J Forest Res 31:1719–1730

Michaelsen J, Schimel D et al (1994) Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J Vegeta Sci 5:673–686

Minnich RA, Barbour MG et al (1995) Sixty years of change in Californian conifer forests of the San Bernardino Mountains. Conserv Biol 9(4):902–914

Mladenoff DJ, Dahir SE, et al (2002) Narrowing historical uncertainty: Probalistic classification of ambiguously identified tree species in historical forest survey data. Ecosystems 5(6):539–553

Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. J Vegeta Sci 15:285–292

Plewe B (2002) The nature of uncertainty in historic geographic information. Transactions in GIS 6(4):431–456

R Statistics (2006) The R Project for Statistical Computing. Retrieved September 19 2006, from http://www.r-project.org/

Radeloff VC, Mladenoff DJ et al (2000) A historical perspective and future outlook on landscape scale restoration in the northwest Wisconsin pine barrens. Restora Ecol 8(2):119–126

Rathbun SL, Black B (2006) Modeling and spatial prediction of pre-settlement patterns of forest distribution using witness tree data. Environ Ecol Stat 13(4):427–448

Rushton SP, Omerod SJ, Kerby G (2004) New paradigms for modelling species distributions? J Appl Ecol 41:193–2000

Russell EWB (1981) Vegetation of Northern New Jersey before European settlement. Am Midland Nat 105(1):1–12

Ryavec KE (2001) Land use/cover change in central Tibet, c. 1830–1990: devising a GIS methodology to study a historical Tibetan land decree. The Geograph J 167(4):342–357

Schulte LA, Mladenoff DJ (2001) The original US Public Land Survey records: their use and limitations in reconstructing presettlement vegetation. Journal of Forestry October 5–10

Schulte LA, Mladenoff DJ et al (2002) Quantitative classification of a historic northern Wisconsin (USA) landscape: mapping forests at regional scales. Can J Forest Res 32:1616–1638

USGS (1928) Topographic Instructions of the United States Geological Survey. c. t. e. C. H. Birdseye

Varanka D (2006) The 20th-Century Topographic Survey as Source Data for Long-Term Landscape Studies at Local and Regional Scales. U. S. G. S. (USGS)

Vayssières MP, Plant RE, Allen-Diaz BH (2000) Classification trees: An alternative non-parametric approach for predicting species distributions. J Veg Sci 11:679–694

Venables WN, Ripley BD (2002) Modern applied statistics with S. New York, Springer

Walker RE (2000) Investigations in vegetation map rectification, and the remotely sensed detection and measurement of natural vegetation changes. Geography University of California

Wang YC (2005) Persettlement land survey records of vegetation: geographic characteristics, quality and models of analysis. Prog Phys Geograph 29(4):568–598

White MA, Mladenoff DJ (1994) Old-growth forest landscapes transitions from pre-European settlement to present. Nat Resour Res Inst 9(3):191–205

Wieczorek J, Gui Q et al (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. Int J Geograph Informa Sci 18(8):745–767

Wieslander AE (1935) A vegetation type map of California. Madroño 3(3):140–144

Wieslander AE (1961) California's vegetation maps: recent advances in botany. University of Toronto Press, Toronto, 4

Wilson JS (2005) Nineteenth century lumber surveys for Bangor, Maine: implications for pre-European settlement forest characteristics in northern and eastern Maine. J Forestry 103(5):218–223