

# Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A.

Janet L. Ohmann and Matthew J. Gregory

**Abstract:** Spatially explicit information on the species composition and structure of forest vegetation is needed at broad spatial scales for natural resource policy analysis and ecological research. We present a method for predictive vegetation mapping that applies direct gradient analysis and nearest-neighbor imputation to ascribe detailed ground attributes of vegetation to each pixel in a digital landscape map. The gradient nearest neighbor method integrates vegetation measurements from regional grids of field plots, mapped environmental data, and Landsat Thematic Mapper (TM) imagery. In the Oregon coastal province, species gradients were most strongly associated with regional climate and geographic location, whereas variation in forest structure was best explained by Landsat TM variables. At the regional level, mapped predictions represented the range of variability in the sample data, and predicted area by vegetation type closely matched sample-based estimates. At the site level, mapped predictions maintained the covariance structure among multiple response variables. Prediction accuracy for tree species occurrence and several measures of vegetation structure and composition was good to moderate. Vegetation maps produced with the gradient nearest neighbor method are appropriately used for regional-level planning, policy analysis, and research, not to guide local management decisions.

**Résumé :** Afin d'effectuer l'analyse des politiques touchant les ressources naturelles et appuyer la recherche écologique, il est nécessaire d'obtenir une information spatiale précise sur la structure de la végétation forestière et sur la composition des espèces et ce, à une vaste échelle spatiale. Nous présentons une méthode de cartographie prévisionnelle de la végétation qui intègre l'analyse de gradient directe et l'application au plus proche voisin pour attribuer des caractéristiques détaillées de la végétation à chaque pixel sur une carte numérique du paysage. L'analyse de gradient du plus proche voisin intègre des mesures de la végétation provenant de réseaux régionaux de parcelles sur le terrain, des données environnementales cartographiées et l'imagerie Landsat capteur TM. Dans la province côtière de l'Oregon, les gradients des espèces sont plus fortement corrélés au climat régional et à la localisation géographique, tandis que les variations dans la structure de la forêt sont mieux expliquées par des variables provenant de Landsat TM. À l'échelle régionale, les prédictions cartographiées représentent bien l'intervalle de variabilité qui caractérise les données échantillonnées et la prédiction des zones par type de végétation correspond bien aux estimations basées sur les échantillons. À l'échelle du site, les prédictions cartographiées maintiennent la structure de covariance parmi les variables à réponses multiples. La précision est bonne à modérée pour les prédictions sur la présence des espèces et ainsi que sur plusieurs mesures de la structure et de la composition de la végétation. L'utilisation des cartes de végétation produites avec la méthode de gradient du plus proche voisin est appropriée pour une planification à l'échelle régionale, pour l'analyse des politiques et pour la recherche environnementale. Elle se prête cependant moins bien aux décisions locales d'aménagement.

[Traduit par la Rédaction]

## Introduction

Issues in forest management grow increasingly complex, involving an array of ecological and commodity values and their interactions. Issues such as biodiversity conservation, long-term productivity and sustainability, and global climate

change require consideration of broad geographic scales (landscapes to regions) and long time frames (decades to centuries). Policy analysis often must address the distribution of forest resources and uses across multiple-ownership regions, as well as changes in landscape patterns and forest conditions over time. Recently, forest assessments have applied simulation models to forest stands in a geographic information system (GIS) to examine regional landscape change (He et al. 1998; Spies et al. 2002). These analyses demand regional-scale information about forest vegetation that is spatially explicit, spans all ownerships and land uses, and describes multiple attributes of composition and structure. Because regional assessments consider multiple components of forest ecosystems and their interactions, it is important that the covariance among vegetation components be realistically portrayed at the local level and that the full

Received 10 February 2001. Accepted 8 January 2002.  
Published on the NRC Research Press Web site at  
<http://cjfr.nrc.ca> on 12 April 2002.

**J.L. Ohmann**<sup>1</sup> USDA Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331, U.S.A.

**M.J. Gregory**, Department of Forest Science, Oregon State University, Corvallis, OR 97331 U.S.A.

<sup>1</sup>Corresponding author (e-mail: [johmann@fs.fed.us](mailto:johmann@fs.fed.us)).

range of variability in each component be represented across the region. Vegetation maps with these characteristics also are needed for basic ecological research on distributions of plant species and communities and on stand and landscape processes.

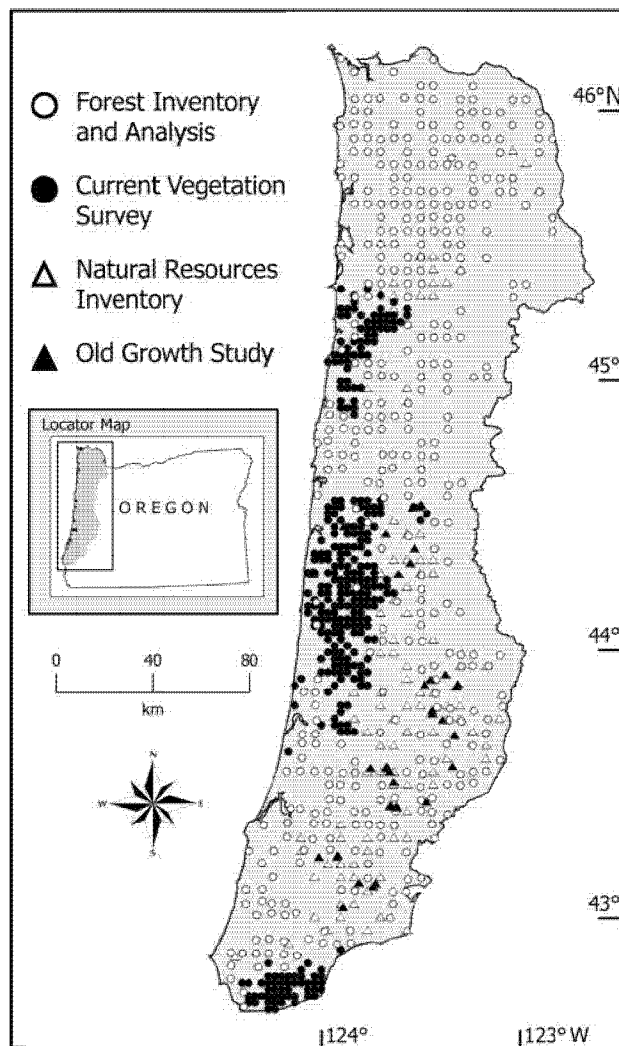
In this paper we present a method for predictive vegetation mapping (*sensu* Franklin 1995) that combines direct gradient analysis (Gauch 1982) with nearest-neighbor imputation to produce digital maps that are rich in floristic and physiognomic information, spatially explicit, and regional in scope. Direct gradient analysis is used to quantify relations between vegetation and environment for a sample of field plot locations, and imputation is used to make spatial predictions that can be displayed in a GIS. Imputation is a statistical analysis tool for incomplete data, whereby measured values are assigned to observations that lack such data (see Van Deusen 1997). We demonstrate the mapping method, which we call gradient nearest neighbor (GNN), for the coastal province of Oregon, U.S.A. (Fig. 1). We developed this approach concurrently with, and independently from, Gottfried et al. (1998), who mapped alpine vegetation in Austria.

Predictive vegetation mapping rests on the premise that vegetation pattern can be predicted from mapped environmental data (Franklin 1995). Predictive models are based on various hypotheses as to how environmental factors control the distribution of species and communities (Guisan and Zimmerman 2000). In our analysis we chose to use canonical correspondence analysis (CCA) (ter Braak 1986; ter Braak and Prentice 1988), a method of gradient analysis, for several reasons. CCA is used widely by ecologists, is multivariate, and can be used for prediction. CCA directly quantifies relations between two multivariate matrices representing the vegetation and environmental data. The ordering of plots is constrained in a regression step, so that resulting plot scores on CCA axes are linear combinations of the environmental variables, and the canonical coefficients can be used for prediction. CCA has been shown to be robust to multicollinearity among explanatory variables (Palmer 1993). Furthermore, the weighted averaging algorithm of CCA implies unimodal response curves of species to the environment. Species distributions along environmental gradients often are nonlinear (Austin et al. 1994), especially along long gradients typical of regional studies like ours. Regional data matrices also typically are sparse (contain many zeros), and CCA is robust to these data. In contrast, linear methods such as principal components analysis and canonical correlation analysis are considered appropriate to data with monotonic species distributions (Jongman et al. 1987). Lastly, CCA is consistent with a conceptual model of vegetation that varies continuously in space in response to environmental and disturbance gradients.

#### Existing methods for regional vegetation mapping and characterization

Most existing regional maps of forest cover are based on classified satellite imagery. Although these data are spatially complete, information content is limited to general characteristics of the upper forest canopy (Cohen et al. 2001; Wolter et al. 1995; Woodcock et al. 1994). Few examples exist of integrating imagery with field plot and environmen-

Fig. 1. The Oregon coastal province, showing locations of field plots. See Table 1 for descriptions of plot data sets.



tal data for ecological modeling and characterization at the regional scale (but see Tomppo 1990; Moeur and Stage 1995; Nilsson 1997). He et al. (1998) used field plot data to populate digital forest cover maps, but assignment of plot data to mapped polygons was done probabilistically rather than based on empirically derived relationships between ground and mapped variables. Our study differs conceptually from image classification, in which plots may be used for training sites and accuracy assessment, and from traditional forest inventories, where plot and remotely sensed data are used in regression or stratified sampling designed to estimate collective measures of a population or stratum. Sample-based inventories can be used to make estimates of known precision for vegetation attributes such as species abundance, tree size distributions, or dead wood characteristics (e.g., see Ohmann et al. 1994); however, within-stratum variation cannot be accurately mapped, and information about variance may be lost (Moeur and Stage 1995). In addition, if individual ground attributes are predicted independently, the joint distribution of estimated ground attributes is distorted if at least one variable is difficult to predict (Moeur

and Stage 1995). Geostatistical methods such as kriging (Isaacs and Srivastava 1990) preserve the spatial structure and variability inherent in the sample data but usually predict a univariate response (e.g., see Ohmann and Spies 1998; Lister et al. 2000), often do not utilize ancillary data layers to improve results, and may truncate the distributions of predicted attributes (Moeur and Hershey 1999).

Franklin (1995) reviewed methods for predictive vegetation mapping of individual plant species or communities but did not address modeling of multivariate responses. Currently, statistical methods of proven utility in predicting responses of multiple, continuous biotic response variables (usually species) are limited to direct ordination, usually CCA (Guisan and Zimmerman 2000). A handful of recent studies have used CCA in predictive vegetation mapping (Hill 1991; Gottfried et al. 1998; Guisan et al. 1999), but we know of only one (Gottfried et al. 1998) that has combined CCA with nearest-neighbor imputation to map multiple response variables. Two imputation methods have been developed for estimating multiple forest variables simultaneously: *k* Nearest Neighbor (*k*NN) (Tomppo 1990; Nilsson 1997) and most similar neighbor (MSN) (Moeur and Stage 1995). In *k*NN, multiple forest variables are simultaneously calculated for unsampled pixels as weighted averages of *k* nearby samples. The sample weights are proportional to distances in feature space defined by spectral (Landsat Thematic Mapper (TM)) data, and all independent variables are given equal weights. Larger values of *k* improve the predicted response for a given pixel but reduce the resemblance between the predicted and actual covariance structures (Nilsson 1997; Tokola et al. 1996). Values of *k* > 1 result in unrealistic assemblages of species or structures, and such estimates can be biased (Moeur and Stage 1995; Nilsson 1997).

The MSN procedure (Moeur and Stage 1995) populates unsampled stands having only mapped data with the detailed ground attributes of the most similar stand for which ground data are available. The similarity measure, which is derived from canonical correlation analysis, weights mapped elements according to their predictive power for all ground elements simultaneously and incorporates the covariance among ground elements. Although MSN has performed well at predicting measures of stand structure (Moeur and Stage 1995), its efficacy for mapping species composition is unknown. Linear methods such as canonical correlation analysis can perform poorly on species relative abundance data across long gradients (Jongman et al. 1987), since species response to environment often is nonlinear (Austin et al. 1994) and data matrices contain many zeros (species absences).

### Study objectives

The purpose of our study was to characterize, both quantitatively and spatially, the current patterns of forest vegetation in the Oregon coastal province. Specific objectives were to (i) quantify spectral, environmental, and disturbance factors associated with regional gradients of tree species composition and structure; (ii) develop GIS-based analytical tools and models to integrate field plot, remotely sensed, and mapped environmental data to map current vegetation; and (iii) produce vegetation maps that are model predictions. The maps were needed to describe initial landscape condi-

tions as input to a simulation model for the coastal landscape analysis and modeling study (CLAMS) (Spies et al. 2002). The simulation model required mapped data on tree density by species and diameter at breast height (DBH), at 25-m pixel resolution to characterize fine-scale heterogeneity needed for wildlife habitat suitability models. To be ecologically realistic, we sought a multivariate method that would predict the co-occurrence of assemblages of tree species and stand structures and maintain their covariance structure. We also wanted mapped predictions to represent the full range of variability in forest vegetation present in the study area. We were interested in simultaneously mapping multiple vegetation attributes that vary continuously, rather than discrete vegetation classes.

## Methods

### The study area

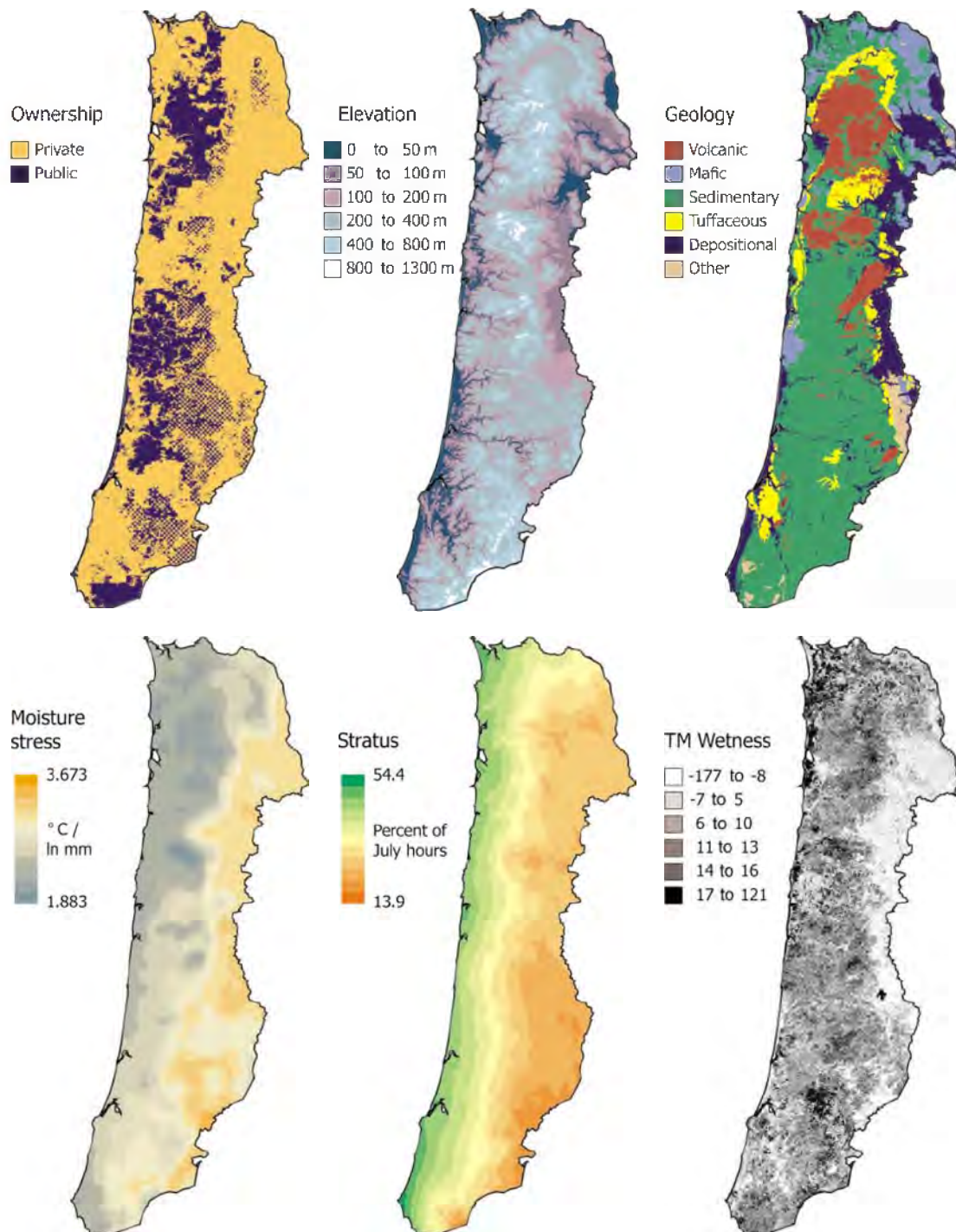
The Oregon coastal province spans  $3 \times 10^6$  ha between 42.6 and 46.3°N and 122.6 and 124.5°W and is bounded on the west by the Pacific Ocean (Fig. 1). The rugged terrain ranges from sea level to 1249 m in elevation (Fig. 2). Geologic formations are primarily marine sandstones and shales, basaltic volcanic rocks, and related intrusives (Fig. 2). Most soils are well drained and have poorly developed horizons, dark surface horizons high in organic matter, and high capacity to hold exchangeable cations. Soils on steep slopes tend to be shallow and stony loam-textured, whereas soils on uneven, benchy, and unstable slopes are deeper and derived from colluvium. The overall climate is maritime, with mild wet winters and cool dry summers, but climate varies geographically with proximity to the ocean, latitude, and orographic effects (Fig. 2).

Regional gradients in species composition in the Pacific Northwest are associated primarily with climate (Ohmann and Spies 1998), whereas patterns of forest structure vary with history of wildfire (Wimberly and Spies 2001a) and timber management and, thus, land ownership (Cohen et al. 2002). National Forests retain landscape patterns created by decades of staggering small harvest units in space. Lands managed by the Bureau of Land Management occur in a "checkerboard" pattern interspersed with private lands, and contain a mix of old and young forest. Forest industry lands typically occur in large blocks that are intensively managed for timber production. Virtually all private forest lands have been harvested at least once and are less than 80 years old. Forests are dominated by coniferous trees, but disturbed sites can be occupied by pioneer broad-leaved trees or shrubs. Broad-leaved trees also occur in riparian areas and in woodlands at the Willamette Valley margin. See Franklin and Dyrness (1973) and Ohmann and Spies (1998) for more detailed descriptions of vegetation and environment.

### Vegetation data from field plots

We obtained vegetation data collected on field plots established in regional forest inventories and research studies: the Natural Resource Inventory (NRI) of the Bureau of Land Management; the Current Vegetation Survey (CVS) of the USDA Forest Service, Pacific Northwest Region (Max et al. 1996); the Forest Inventory and Analysis (FIA) of the USDA Forest Service, Pacific Northwest Research Station; and the

**Fig. 2.** Geographic patterns of selected explanatory variables used in the gradient nearest neighbor method. See Table 3 for variable descriptions.



Old Growth Study (OGS) of the USDA Forest Service, Pacific Northwest Research Station (Spies and Franklin 1991) (Table 1, Fig. 1). The field plots sampled all forest lands in the province; inventory plots on nonforest land were not measured in the field. All plots were installed on systematic grids except the OGS plots, which were selected subjectively to sample older forests.

The inventory plots averaged about 1 ha in area. The OGS plots sampled irregularly shaped stands of 7–60 ha, but sub-

plots were clustered within a smaller portion of the stand. Within each plot, trees  $\geq 2.54$  cm diameter at breast height (DBH) were sampled on a series of 1–10 nested fixed- and variable-radius plots, and the species and DBH of each tree were recorded. We combined data from the four data sets into a consistent format and computed the basal area and number of trees per hectare represented by each tree. For each plot we summarized basal area by species (Table 2) and size class. Size classes were based on tree DBH: 2.5–

**Table 1.** Sources of field plot data on forest vegetation.

| Data set                      | Ownerships sampled                    | Years measured | Sample design  | No. of plots | No. of pixels per plot |
|-------------------------------|---------------------------------------|----------------|--|--------------|------------------------|
| Natural Resources Inventory   | Bureau of Land Management             | 1997           | Systematic grid: 5.5 km  | 99           | 13                     |
| Current Vegetation Survey     | Siskiyou and Siuslaw National Forests | 1993–1996      | Systematic grid: 2.7 km outside wilderness, 5.5 km in wilderness | 304          | 13                     |
| Forest Inventory and Analysis | Nonfederal lands                      | 1984–1986      | Systematic grid: 5.5 km  | 381          | 9 or 22                |
| Old Growth Study              | Federal lands                         | 1984           | Located subjectively in forest >80 years old                     | 39           | 112–963                |

**Table 2.** Tree species in this study.

| Scientific name  | Code  | Frequency<br>( <i>n</i> = 823) |
|--|-------|--------------------------------|
| <i>Abies amabilis</i> (Dougl.) Forbes  | ABAM  | 2                              |
| <i>Abies grandis</i> (Dougl.) Forbes and <i>Abies concolor</i> (Gord. & Glend.) Lindl. | ABGR  | 62                             |
| <i>Abies procera</i> Rehder  | ABPR  | 11                             |
| <i>Acer macrophyllum</i> Pursh   | ACMA  | 258                            |
| <i>Alnus rubra</i> Bong.   | ALRU  | 470                            |
| <i>Arbutus menziesii</i> Pursh   | ARME  | 67                             |
| <i>Calocedrus decurrens</i> (Torr.) Florin.  | CADE  | 23                             |
| <i>Castanopsis chrysophylla</i> (Dougl.) DC.   | CHCH  | 75                             |
| <i>Chamaecyparis lawsoniana</i> A. Murray  | CHLA  | 53                             |
| <i>Cornus nuttallii</i> Aud.   | CONU  | 48                             |
| <i>Fraxinus latifolia</i> Benth.   | FRLA  | 12                             |
| <i>Lithocarpus densiflorus</i> (Hook. & Arn.) Rehder                                   | LIDE  | 54                             |
| <i>Picea sitchensis</i> S. Watson  | PISI  | 127                            |
| <i>Pinus attenuata</i> Lemmon  | PIAT  | 2                              |
| <i>Pinus contorta</i> var. <i>latifolia</i> Engelm.                                    | PICO  | 3                              |
| <i>Pinus lambertiana</i> Dougl.  | PILA  | 6                              |
| <i>Pinus monticola</i> Dougl.  | PIMO  | 1                              |
| <i>Pinus ponderosa</i> Dougl.  | PIPO  | 4                              |
| <i>Prunus emarginata</i> (Dougl.) Walp.  | PREM  | 52                             |
| <i>Prunus virginiana</i> L.  | PRVI  | 2                              |
| <i>Pseudotsuga menziesii</i> (Mirb.) Franco  | PSME  | 722                            |
| <i>Quercus garryana</i> Dougl.   | QUGA  | 38                             |
| <i>Quercus chrysolepis</i> Liebm.  | QUCH  | 12                             |
| <i>Quercus kelloggii</i> Newberry  | QUKE  | 12                             |
| <i>Rhamnus purshiana</i> DC.   | RHPU  | 15                             |
| <i>Salix</i> spp. L.   | SALIX | 22                             |
| <i>Taxus brevifolia</i> Nutt.  | TABR  | 21                             |
| <i>Thuja plicata</i> Donn  | THPL  | 176                            |
| <i>Tsuga heterophylla</i> (Raf.) Sarg.   | TSHE  | 408                            |
| <i>Umbellularia californica</i> (Hook. & Arn.) Nutt.                                   | UMCA  | 44                             |

**Note:** Nomenclature is from Little (1979).

25.4 cm, 25.5–50.4 cm, 50.5–75.4 cm, 75.5–100.4 cm, and ≥100.5 cm. We computed a variety of plot-level measures of vegetation structure and composition from the basic tree data for use in vegetation mapping and accuracy assessment.

#### Landsat 5 TM imagery

We developed 10 data layers from bands 1–5 and 7 of Landsat 5 TM imagery (Table 3). Because the plots were measured across a wide range of dates (1984–1997), we developed TM data for 2 years, 1988 and 1996. Portions of

five TM scenes were needed to cover the study area. We normalized values for the TM bands among adjacent and overlapping scenes within each year, then between the 2 years, using a histogram equalization function (Lillesand and Kiefer 1994) in Erdas Imagine. Before normalizing the images, we excluded pixels that changed significantly between dates, primarily clear-cut timber harvests. We transformed each mosaiced image into tasseled cap brightness, greenness, and wetness indices (Kauth and Thomas 1976), which have demonstrated utility for mapping forest cover in

**Table 3.** Mapped explanatory variables used in the gradient nearest neighbor method.

| Variable class and code | Definition   |
|-------------------------|--|
| <b>Ownership</b>        |  |
| PUB                     | Public land ownership (federal, state, or local government)  |
| <b>Topography</b>       |  |
| ELEV                    | Elevation (m), from 30-m digital elevation model (DEM)   |
| ASPECT                  | Cosine transformation of aspect (degrees) (Beers et al. 1966), 0.0 (southwest) to 2.0 (northeast), from 30-m DEM                 |
| SLOPE                   | Slope (percent), from 30-m DEM   |
| SLPOS                   | Slope position, from 0 (bottom of drainage) to 100 (ridgetop), from SLOPEPOSITION macro in ArcInfo on 30-m DEM                   |
| SOLAR                   | Solar radiation (cal/cm <sup>2</sup> ) from program SolarImg (Harmon and Marks 1995) and 100-m DEM                               |
| <b>Geology</b>          |  |
| VOLC                    | Volcanic and intrusive rocks   |
| MAFO                    | Mafic rocks (basalt, basaltic andesite, andesite, gabbro); Miocene and older   |
| SEDR                    | Siltstones, sandstones, mudstones, conglomerates (sedimentary)   |
| TUFO                    | Tuffaceous rocks and tuffs, pumicites, silicic flows; Miocene and older  |
| DEPO                    | Depositional (dune sand, alluvial, glacial, glaciofluvial, loess, landslide and debris flow, playa, lacustrine, fluvial)         |
| <b>Climate</b>          |  |
| ANNPRE                  | Mean annual precipitation (natural logarithm, mm)  |
| SMRPRE                  | Mean precipitation from May to September (natural logarithm, mm)   |
| CVPRE                   | Coefficient of variation of mean monthly precipitation of December and July (wettest and driest months)                          |
| SMRTMP                  | Moisture stress during the growing season, computed as SMRTMP/SMRPRE, where SMRTMP is the mean temperature (°C) in May–September |
| ANNTMP                  | Mean annual temperature (°C)   |
| AUGMAXT                 | Mean maximum temperature in August (°C) (hottest month)  |
| DIFTMP                  | Difference between AUGMAXT and DECMINT (°C), where DECMINT is the mean minimum temperature in December (coldest month)           |
| STRATUS                 | Percentage of the hours in July with cloud ceiling of marine stratus <1524 m and visibility <8 km                                |
| <b>Landsat TM</b>       |  |
| B1                      | Band 1 (blue)  |
| B2                      | Band 2 (green)   |
| B3                      | Band 3 (red)   |
| B4                      | Band 4 (near-infrared)   |
| B5                      | Band 5 (mid-infrared)  |
| B7                      | Band 7 (mid-infrared)  |
| R43                     | Ratio of B4 to B3  |
| R54                     | Ratio of B5 to B4  |
| R57                     | Ratio of B5 to B7  |
| BRT                     | Brightness axis from tasseled cap transformation   |
| GRN                     | Greenness axis from tasseled cap transformation  |
| WET                     | Wetness axis from tasseled cap transformation  |
| DISTURB                 | No. of years since clear-cut harvest, from multitemporal Landsat TM analysis (Cohen et al. 2002)                                 |
| <b>Location</b>         |  |
| X                       | Longitude (decimal degrees)  |
| Y                       | Latitude (decimal degrees)   |

our region (Cohen and Spies 1992; Cohen et al. 1995, 2001). We filtered each of these TM grids twice in succession, using a 3 × 3 pixel window and assigning the median value to the center pixel. This filtering reduced fine-scale heterogeneity, retained vegetation boundaries, and improved prediction accuracy. We also obtained maps of clear-cut harvests from

1972 to 1995 developed by Cohen et al. (2002) from multi-temporal TM data, which we converted to maps of number of years since harvest.

To assign values from the TM-based grids to plots, we represented each plot as a template of pixels with a configuration that approximated the plot's layout on the ground, an-

chored by its  $X$  and  $Y$  coordinates. We used an ArcInfo macro to overlay the plot templates on each TM grid and retrieve the mean values associated with each plot. For the disturbance grid we used the majority value. We assigned TM data from both 1988 and 1996 to each plot, but in the analyses we used TM data from the year most closely matching the date of ground measurement. We eliminated the following kinds of plots from all analyses: plots in shadow, water, or cloud in the imagery; plots with obvious mismatches between ground and spectral data (due to location errors or to harvesting between date of imagery and date of field measurement); and plots on obvious edges such as harvest units, roads, or streams.

### Mapped data on climate, topography, geology, and location

We obtained map layers for climatic, topographic, and geologic variables (Table 3) that are available in digital format and that have been shown to be associated with patterns of forest vegetation in the Pacific Northwest (Ohmann and Spies 1998). We converted the layers to grids as needed, resampled them to  $25 \times 25$  m (the resolution of our predictions), and assigned mean or majority values for each grid to the plots using the procedure described above for the TM data.

We derived climate data from mean annual and mean monthly precipitation and temperature surfaces generated by the precipitation–elevation regressions on independent slopes model (PRISM) (Daly et al. 1994). PRISM uses DEMs to account for topographic effects in interpolating weather measurements from an irregular network of weather stations to a uniform grid. The PRISM surfaces were generated at 4.7-km resolution from 1961–1990 weather data. We log transformed all precipitation surfaces, because vegetation does not respond linearly to amount of precipitation. From the mean monthly PRISM grids we computed several climatic indices that approximate growing season conditions, seasonal variability, and continentality (Table 3). We also acquired a map of July frequency of low stratus clouds (C. Daly, Spatial Climate Analysis Service, Oregon State University, Corvallis, OR 97331, unpublished data) (Table 3). Summer fog, common along the Pacific coast, is thought to influence plant species distributions by reducing moisture stress during the growing season.

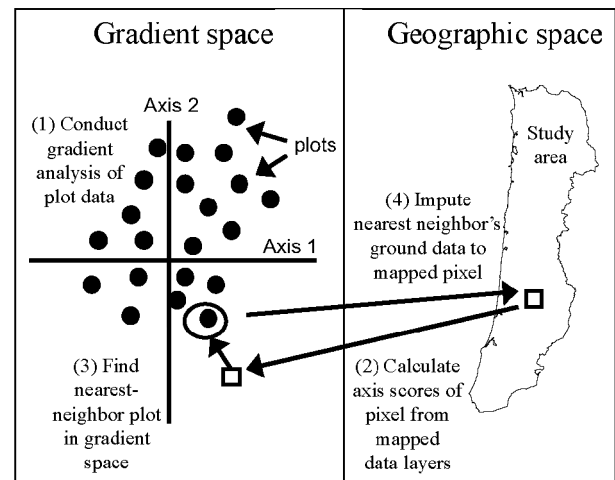
We derived several topographic measures from a 30-m DEM (Table 3). We derived 14 generalized geologic types from a digital version of the geologic map of Oregon (Walker and MacLeod 1991), five of which occurred in the study area (Table 3). Lastly, we used the latitude ( $Y$ ) and longitude ( $X$ ) coordinates for each plot as explanatory variables.

### The gradient nearest neighbor method of predictive vegetation mapping

Our method for predictive vegetation mapping involves the following steps (Fig. 3), which we refer to collectively as the gradient nearest neighbor (GNN) method:

- (1) Conduct direct gradient analysis using stepwise CCA (ter Braak 1986; ter Braak and Prentice 1988) to develop a model that quantifies relations between ground (response) data and mapped (explanatory) data.

Fig. 3. Steps in the gradient nearest neighbor method (pixel size not to scale).



- (2) For each mapped  $25 \times 25$  m pixel (the spatial resolution of our mapped predictions), predict scores for the first eight CCA axes by applying coefficients from the model developed in step 1 to the mapped values for the explanatory variables.
- (3) For each mapped pixel, identify the single plot that is nearest in eight-dimensional gradient space, where distance is Euclidean and axis scores are weighted by their eigenvalues. Also identify the second-nearest plot for accuracy-assessment purposes (see below).
- (4) Impute the ground attributes of the nearest-neighbor plot to the mapped pixel. Following imputation, maps can be constructed for any vegetation attribute measured on the field plots.

We ran CCA in the program CANOCO, version 4 (ter Braak and Smilauer 1998), using 823 plots. We used information listed by CANOCO to identify explanatory variables that were highly correlated (variance inflation factors  $>20$ ). We used the forward stepwise procedure in CANOCO to identify and retain those among the collinear variables that explained the most variation in the species data. In this way we identified a subset of variables that avoided collinearity between variables but retained as much environmental information as possible. Response variables were basal area ( $\text{m}^2/\text{ha}$ ) by species (Table 2) and size classes described previously. Within species, we combined size classes that had very low frequencies of occurrence. We square-root transformed basal area values to dampen the influence of dominant species and because square root transformed values were most strongly correlated with the explanatory variables. In CANOCO, we downweighted rare species and selected species scores as weighted mean sample scores. We added explanatory variables to the stepwise CCA models in the order of greatest additional contribution to explained variation. Variables were added only if they were significant ( $P < 0.01$ ), where significance was determined by a Monte Carlo permutation test using 99 permutations ( $H_0$ : additional influence of variable on vegetation is not significantly different from random) and only if adding the variable did not cause any variance inflation factors to exceed 20, which indicates strong multicollinearity (ter Braak and Smilauer



1998). We excluded  $X$  and  $Y$  from the stepwise procedure, because they are strongly correlated with several of the explanatory variables and do not directly measure environmental factors that influence plants. However, we added  $X$  and  $Y$  to the final model so that geographic location would be considered in the selection of nearest-neighbor plots. This constrained the nearest-neighbor distances and slightly improved prediction accuracy.

Nillson (1997) compared several distance measures in the  $k$ NN method and determined that Euclidean distance was appropriate for applications similar to ours. We used eight CCA axes, because they accounted for almost all (94%) of the total variation explained, and because prediction accuracy was better than with fewer axes. By weighting the axes by their eigenvalues in the distance calculations, we gave more weight to axes with greater explanatory power. In addition, use of unweighted axes resulted in overfitting of the model and reduced prediction accuracy for independent observations.

We produced two versions of the GNN predictions, for 1988 and 1996, the 2 years for which we had Landsat TM imagery. Only one CCA model was developed from analysis of the plot data, using TM variables for the year closest to each plot's measurement date. The CCA coefficients were then applied to both years to make the GNN predictions. Because our model was applicable only to forested areas where plot data were available, we masked out nonforest areas (water, urban, agriculture, sand dunes, etc.) from our GNN predictions and accuracy assessments using a locally developed land-use map.

### Model evaluation and accuracy assessment

We evaluated performance of the GNN method in several ways. At the aggregate, regional level, we compared relative proportions of mapped vegetation classes predicted by GNN with those estimated from the systematic grids of field plots. We also compared overall means and ranges of variability of the mapped GNN predictions to those of the plot data for several vegetation attributes, to evaluate how well GNN retained the variability present in the observed data.

We assessed the site-level accuracy of GNN by comparing predicted to observed (ground-measured) values for the 823 plot locations. These comparisons also indicated how well GNN maintained the known variability in the plot data across the site-specific locations. The plot data were regarded as truth and assumed to be measured and geo-referenced without error. For each of several vegetation attributes, means of the GNN-predicted, pixel-level values corresponding to each plot location were calculated. For each plot location we used the predicted value associated with the second-nearest neighbor, rather than the nearest neighbor (which would be the plot itself). We expected this method to be effectively the same as a data-splitting analysis where 823 versions of the model are run, each time leaving out one plot, but it was computationally much more efficient. Although the substitution of the second-nearest neighbor is not mathematically equivalent to a run that omits the  $i$ th observation, since each CCA model is influenced by the  $i$ th observation, it is extremely unlikely that omitting one plot would cause a large change in the CCA model. A 10-fold crossvalidation analysis supported our assumption that

**Table 4.** Variation explained by subsets of variables in canonical correspondence analysis.

| Subset of explanatory variable | Percentage of total inertia |
|--------------------------------|-----------------------------|
| Ownership                      | 2.2                         |
| Topography                     | 4.5                         |
| Geology                        | 1.8                         |
| Climate                        | 8.0                         |
| Landsat TM                     | 15.2                        |
| Location                       | 5.2                         |

**Note:** See Table 3 for variable membership in subsets.

the CCA model is robust to changes in plot input data. We divided the 823 plots into 10 random subsets and developed 10 CCA models, each time leaving out a different 10% of the plots and developing the model using the other 90%. Comparisons of predicted to observed for the 10 models were nearly identical to results from the second-nearest-neighbor approach. We also assessed accuracy by reserving 25% of the plots and developing the GNN model with the remaining 75% of the plots. Again, results were nearly the same as those from the second-nearest-neighbor analysis. For these reasons, we present accuracy results only from the second-nearest-neighbor analysis in this paper.

We used the kappa coefficient of agreement (Cohen 1960), a measure of classification accuracy that discounts chance agreement, to compare predicted to observed values for vegetation classes and species occurrence. The formula for kappa ( $\kappa$ ) is  $\kappa = (p_o - p_c) / (1 - p_c)$ , where  $p_o$  is the overall classification accuracy (probability, over all classes, that the predicted and observed values agree) and  $p_c$  is the chance agreement between predicted and observed values. Errors of omission and commission are treated equally.

To reduce bias in our accuracy assessment caused by temporal differences between the TM imagery and plot measurement, we compared observed and predicted values for a given plot for either 1988 or 1996, whichever year was closer to the year of plot measurement. Thus, our summaries of accuracy actually reflect a composite of the 1988 and 1996 predictions.

Finally, we mapped the nearest-neighbor distances from GNN as a measure of the geographic distribution of confidence in the GNN predictions. Shorter distances indicate areas of greater confidence in the results, and greater distances represent potential areas of poorer accuracy, as well as environmental conditions that may be undersampled by field plots (Moeur and Stage 1995).

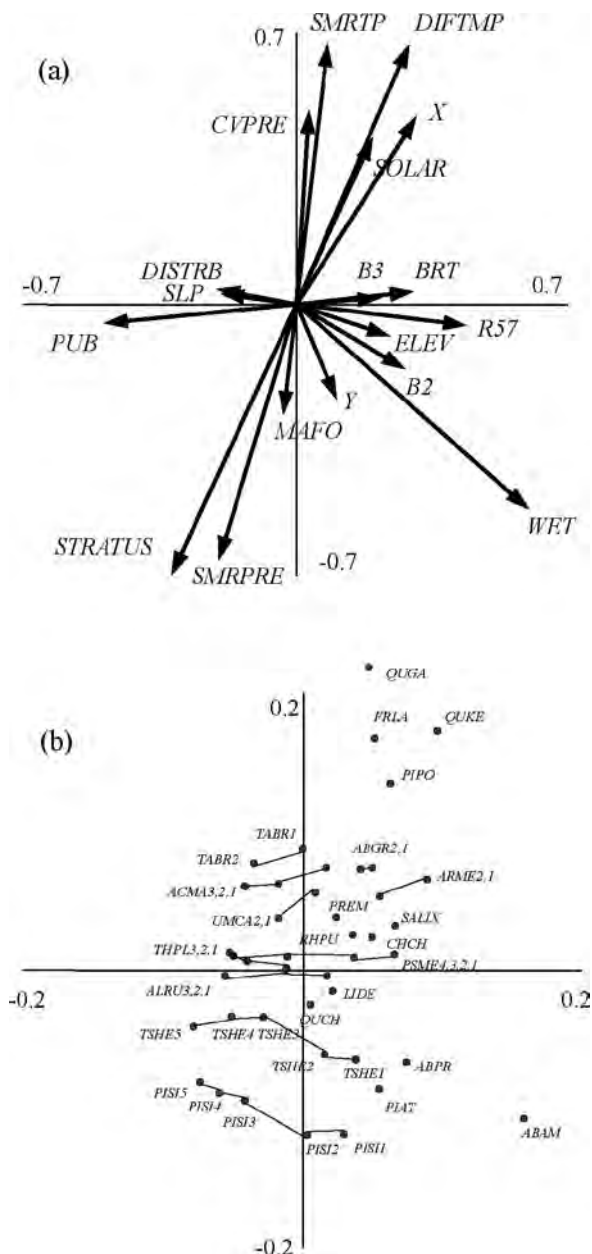
## Results

### Gradients in species composition and structure

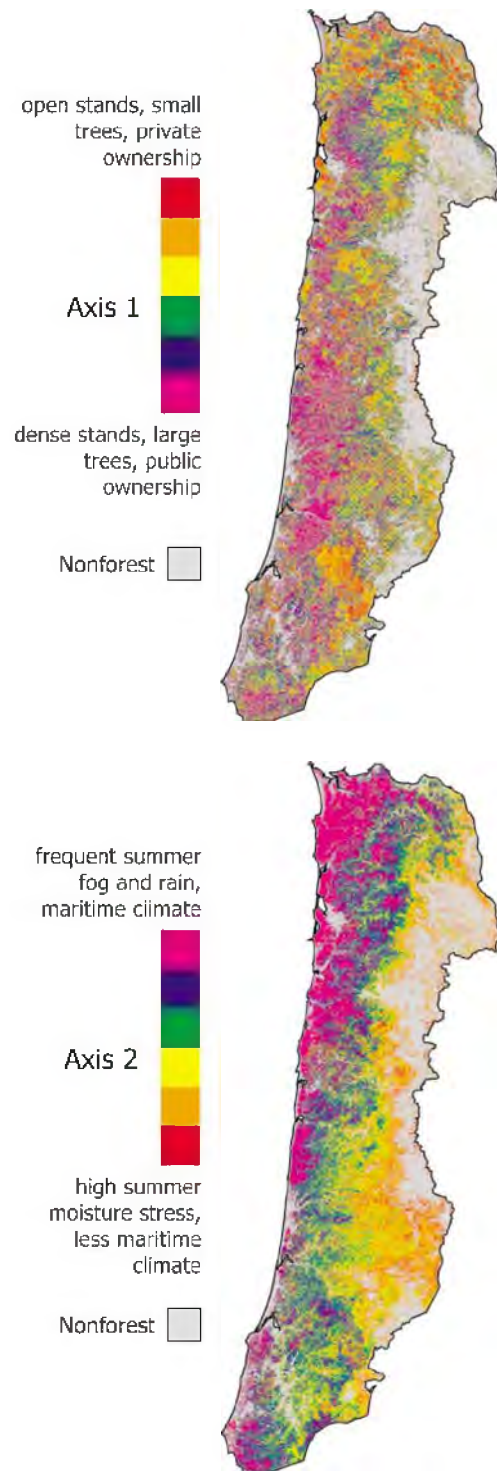
Overall gradients in the species composition and structure of forest vegetation were most strongly associated with Landsat TM variables and climate and, secondarily, with location, topography, ownership, and geology in decreasing order of importance (Table 4). The primary gradient (axis 1) was in forest structure (tree size and density), which varied with TM wetness and ownership (Figs. 4 and 5). Low scores



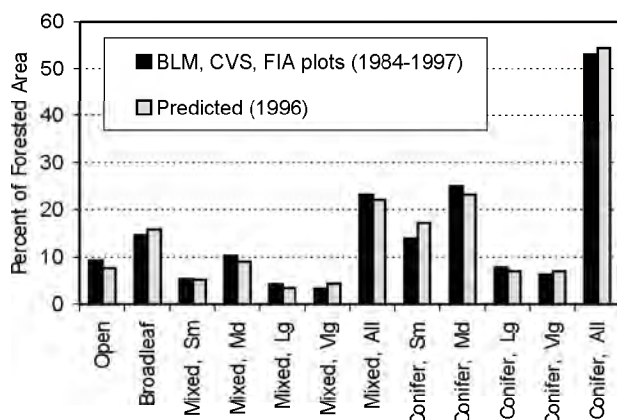
**Fig. 4.** Biplots (see ter Braak and Smilauer 1998) showing associations between vegetation and explanatory variables for the dominant gradients in the Oregon coastal province from canonical correspondence analysis (CCA). (a) Explanatory variables. See Table 3 for variable definitions. Arrow length and position of the arrowhead indicate the correlation between the explanatory variable and the CCA axes, and smaller angles between arrows indicate stronger correlations between variables. (b) Species centroids (circles) in relation to the CCA axes and explanatory variables in Fig. 4a. Lines connect size classes of a given species. See Table 2 for species codes (not shown: CADE, CHLA, CONU, PICO, PILA, PIMO, PRVI). Size-class codes are as follows: (1) small (2.5–25.4 cm diameter at breast height (DBH)); (2) medium (25.5–50.4 cm DBH); (3) large (50.5–75.4 cm DBH); (4) very large (75.5–100.4 cm DBH); and (5)  $\geq 100.5$  cm DBH.



**Fig. 5.** Geographic patterns of dominant gradients in forest vegetation and environment, which are predicted scores on axis 1 and 2 from canonical correspondence analysis.



**Fig. 6.** Comparison of area by vegetation class predicted by the gradient nearest neighbor method (based on  $n = 823$  plots) and estimated from systematic grids of field plots ( $n = 1039$  plots). See Table 6 for definitions of vegetation classes. Sm, small; Md, medium; Lg, large; Vlg, very large.



on axis 1 were stands of large trees on public lands, and high scores were younger stands on private lands. Axis 2 differentiated species along a climatic gradient from coastal areas with frequent summer fog and more summer rainfall to inland areas with greater summer moisture stress and less maritime influence (Figs. 4 and 5). Axis 2 also was associated with variation in TM wetness. Species were arranged on axis 2 from *Picea sitchensis* and *Abies amabilis*, species found along the coast and at higher elevations, to *Quercus garryana* and *Quercus kelloggii*, species found in the driest and least maritime habitats in the eastern and southern parts of the study area (Fig. 4). Axis 3 was correlated with elevation, TM brightness, and TM bands 2 and 3, and separated evergreen species of southwestern Oregon (lowest scores) and broadleaf deciduous species (highest scores). Subsequent axes were difficult to interpret.

### Overall performance of GNN

The relative proportions of forest conditions across the province predicted by GNN very closely matched those estimated by systematic grids of inventory plots (Fig. 6). This agreement was not necessarily expected, even though GNN used a subset of the inventory plots (see Discussion). In addition, the mapped GNN predictions reproduced the sampled range of variability in vegetation across the province very closely. The means and standard deviations of several vegetation attributes predicted by GNN nearly exactly matched those observed on the 823 plots (Table 5). The ranges of predicted and observed values matched exactly, because all plots were selected as nearest neighbors at least once. Note that, in these comparisons of predicted and observed, it cannot be determined whether differences are due to errors of prediction or to real change between the dates of plot measurement and GNN prediction. The overall geographic patterns of the GNN predictions appeared reasonable (Figs. 7 and 8) except in some areas along the coast and Willamette Valley margin, which contained the fewest field plots and the longest nearest-neighbor distances (Fig. 9).

**Table 5.** Comparison of descriptive statistics for observed ( $n = 823$  plots) and predicted (mapped) vegetation for selected attributes of forest vegetation, Oregon coastal province.

| Vegetation attribute                          | Mean | Range     | SD   |
|---|------|-----------|------|
| <b>Total basal area (<math>m^2/ha</math>)</b> |      |           |      |
| Observed                                      | 33.9 | 0.0–124.9 | 20.6 |
| Predicted                                     | 31.0 | 0.0–124.9 | 22.3 |
| <b>Broadleaf basal area proportion</b>        |      |           |      |
| Observed                                      | 0.27 | 0.0–1.00  | 0.32 |
| Predicted                                     | 0.26 | 0.0–1.00  | 0.32 |
| <b>Quadratic mean diameter (cm)</b>           |      |           |      |
| Observed                                      | 34.5 | 0.0–166.2 | 22.4 |
| Predicted                                     | 33.2 | 0.0–166.2 | 24.6 |
| <b>No. of trees/ha &gt;100 cm DBH</b>         |      |           |      |
| Observed                                      | 3.0  | 0.0–54.4  | 7.5  |
| Predicted                                     | 3.0  | 0.0–54.4  | 7.7  |
| <b>Stand age (years)</b>                      |      |           |      |
| Observed                                      | 51.1 | 0.0–718.0 | 44.2 |
| Predicted                                     | 52.0 | 0.0–718.0 | 56.3 |
| <b>Tree species richness</b>                  |      |           |      |
| Observed                                      | 3.1  | 0–11      | 1.6  |
| Predicted                                     | 3.0  | 0–11      | 1.7  |

### Accuracy of GNN predictions at the site level

At the site level, overall classification accuracy for 10 classes defined by vegetation density, species composition, and size class was 45% (Table 6). Accuracies were 0–54% better than chance for individual classes, with a mean  $\kappa$  of 0.31 (Table 7). The  $\kappa = -0.03$  for the mixed conifer–broadleaf, very-large class can be attributed to the very small sample size ( $n = 5$ ). Most misclassification errors were minor: the overall classification was 87% correct within one class (Table 6) and 72–98% better than chance for individual classes (mean  $\kappa = 0.83$ ) (Table 7). Among composition classes, classification accuracy was poorest for mixed conifer–broadleaf forests ( $\kappa = 0.30$ ), best for conifer forests ( $\kappa = 0.59$ ), and intermediate for broadleaf forests ( $\kappa = 0.49$ ).

Correlations between predicted and observed values for six measures of vegetation structure and composition ranged from 0.53 for tree species richness to 0.80 for quadratic mean diameter (QMD) (Fig. 10). Correlations generally were greatest for measures associated with successional status of vegetation (0.80 for QMD and 0.71 for stand age). For all continuous vegetation attributes the GNN method overpredicted at low values and underpredicted at high values (Fig. 10).

Prediction accuracy for the occurrence of seven common tree species was 56–89% or 21–53% better than chance (mean  $\kappa = 0.29$ ) (Table 8). For all species, errors of commission were more common than errors of omission. Predictions were most accurate for species whose distributions are geographically limited and strongly associated with climate (e.g., *Picea sitchensis* and *Quercus garryana*). Widely distributed species that occur in locally low abundances (e.g., *Acer macrophyllum* and *Thuja plicata*) or whose local abundances are associated with disturbance history (e.g., *Tsuga*

*heterophylla*; Wimberly and Spies 2001b) were more difficult to predict. Chance-corrected prediction of our most ubiquitous tree species (Table 2), *Pseudotsuga menziesii*, was fairly poor since the probability of predicting its occurrence by chance already was fairly high.

## Discussion

### Accuracy of the GNN vegetation maps

We evaluated the GNN predictions in ways that should be familiar to readers with backgrounds in forest inventory and image classification. However, we caution against directly comparing our accuracies with other published accounts because of differences in methods of both map construction and accuracy assessment. Nevertheless, in a broad sense the GNN predictions appear similar in accuracy to other Landsat TM based studies in western Oregon forests. This is not surprising, given the primary importance of Landsat TM in our predictive model and probably indicates inherent limitations of Landsat TM for mapping forest vegetation. Cohen et al. (2001), the Interagency Vegetation Mapping Project (Weyermann and Fassnacht 2000), and GNN all achieved correlation coefficients ranging from 0.66 to 0.86 for several continuous measures of vegetation structure and composition. Our prediction accuracies for occurrence of individual tree species also were similar to other published studies (e.g., Iverson and Prasad 1998; Guisan et al. 1999).

The tendency for regression methods to yield biased predictions, as we observed with GNN (Fig. 10), is a problem that has long been recognized in remote sensing and other studies (Curran and Hay 1986). Measurement errors in  $X_i$ , which are assumed in regression analysis to be zero, result in an underestimate of the slope of the regression (where slope is positive). Whereas several methods have been suggested for addressing this problem during model calibration (Curran and Hay 1986), research is needed into how such methods might be incorporated into CCA.

The ability to predict a given vegetation attribute with GNN is influenced by the response variables specified in the underlying CCA model. The response variables are summary measures calculated from the basic tree-level data on each plot, which are chosen by the analyst and can be tailored to study objectives. Presumably, the closer the resemblance between a predicted vegetation attribute and the response variables used in model development, the better the expected accuracy. Improving prediction accuracy for some vegetation attributes may come at the cost of reduced accuracy for others, and it may be possible to optimize the model for particular attributes. Similarly, perfect accuracy for multiple vegetation attributes in the GNN predictions is impossible, because two plots never are exactly alike nor are the vegetation and explanatory factors perfectly correlated.

Although the sample-based estimates of vegetation classes are not completely independent of those predicted by GNN (Fig. 6), there are several reasons we might expect them to differ, and why we think this represents a useful comparison. The GNN predictions were based on 79% of the total inventory plots (784 of 1039 plots), plus 39 OGS plots that were not part of the sample-based estimates. The subset of plots used in GNN was not selected randomly or systematically

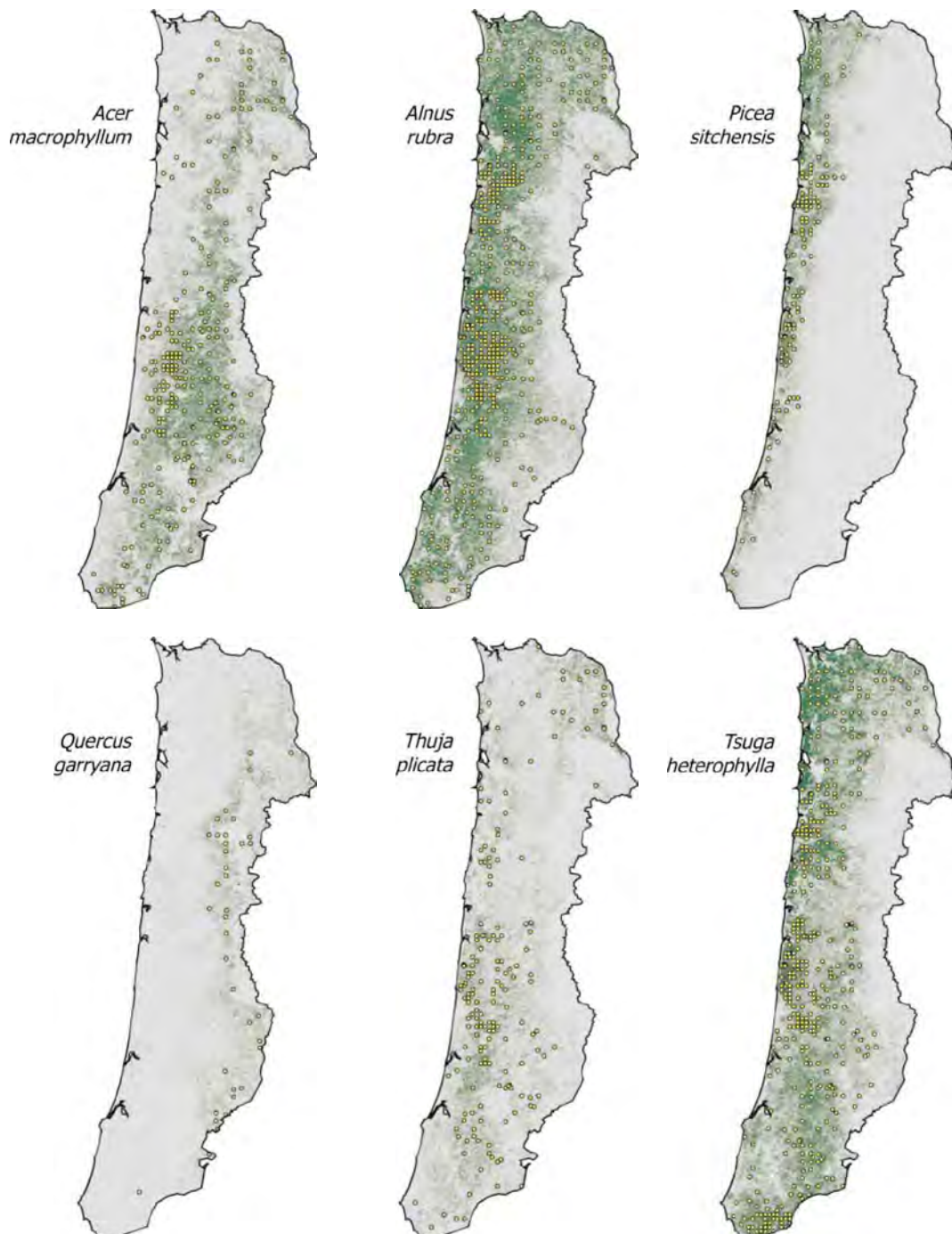
**Fig. 7.** Predicted vegetation classes from the gradient nearest neighbor method. See Table 6 for definitions of vegetation classes.



and, thus, had potential to yield biased results. We excluded from GNN those plots with obvious mismatches between ground and spectral data and that straddled distinct boundaries in forest condition. Furthermore, for the plot-based estimates nonforest land uses were determined in the field,



**Fig. 8.** Predicted occurrence of tree species from the gradient nearest neighbor method (shaded in green). Yellow circles are field plot locations where the species was observed.



whereas we applied an independently derived map of nonforest areas to the GNN predictions.

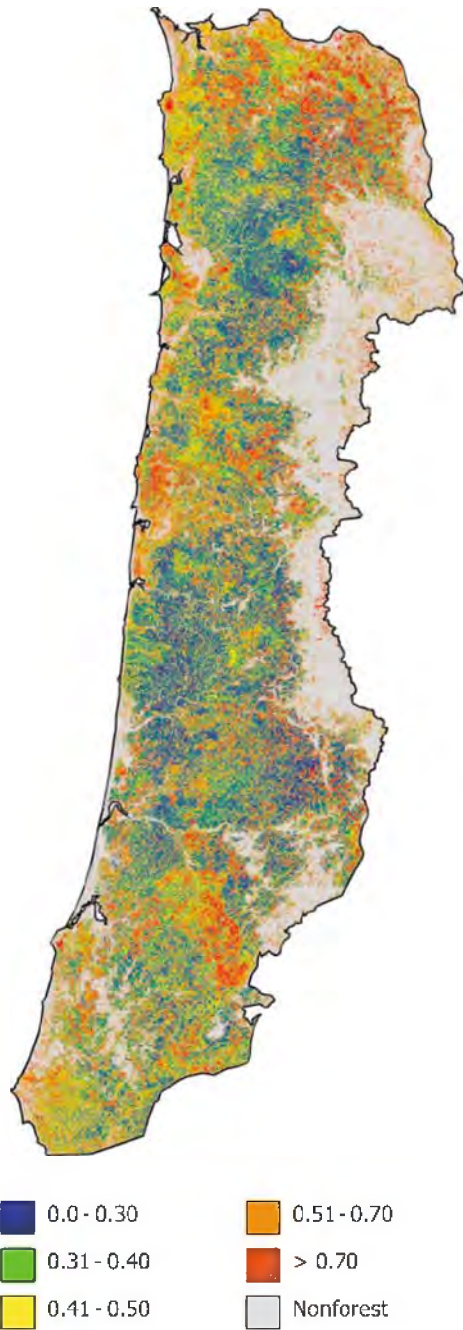
We evaluated GNN prediction accuracy from both regional (Table 5, Fig. 6) and site-level (Tables 6–8, Fig. 10) perspectives but have not evaluated the spatial distribution of error. The nearest-neighbor distances (Fig. 9) indicate potential error patterns only. Research is needed on the application of methods of spatial accuracy assessment (Lowell and Jaton 1999; Hunsaker et al. 2001) to GNN vegetation maps. The spatial distribution of error has particular implications

for applications of vegetation maps that utilize information about landscape pattern, such as models of wildlife habitat suitability.

#### Sources of error in GNN

Cross-validation methods quantify the collective effect of all sources of error on prediction accuracy. Errors attributable to the GNN method itself, which are of most interest to us, cannot be distinguished from other sources of error. Other important sources include errors in the mapped ex-

**Fig. 9.** Nearest-neighbor distances for the gradient nearest neighbor method for  $n = 823$  plots. Distance is Euclidean distance in eight-dimensional gradient space, based on the first eight axes in canonical correspondence analysis, with distance to each axis weighted by its eigenvalue.



planatory variables and georegistration errors among the mapped explanatory variables and plot locations. We excluded plots with obvious location errors from our analysis, but undiscovered errors contribute to overall prediction accuracy to an unknown degree. Errors and limitations associated with the use of Landsat TM imagery in forest vegetation mapping are described elsewhere (e.g., see Franklin 2001) and not enumerated here.

**Table 6.** Error matrix and prediction accuracy for vegetation classes from gradient nearest neighbor method, based on numbers of  $n = 823$  plots.

| Observed class*     | Predicted class* |                 |                |                 |                |                   |                 |                 |                 |                     | % within one class |
|---------------------|------------------|-----------------|----------------|-----------------|----------------|-------------------|-----------------|-----------------|-----------------|---------------------|--------------------|
|                     | Open             | Broadleaf       | Mixed, small   | Mixed, medium   | Mixed, large   | Mixed, very large | Conifer, small  | Conifer, medium | Conifer, large  | Conifer, very large | % correct          |
| Open                | 11               | 9 <sup>†</sup>  | 6 <sup>†</sup> | 0               | 0              | 0                 | 4 <sup>†</sup>  | 0               | 0               | 0                   | 100                |
| Broadleaf           | 0 <sup>†</sup>   | 50              | 5 <sup>†</sup> | 31 <sup>†</sup> | 6 <sup>†</sup> | 0 <sup>†</sup>    | 2               | 4               | 1               | 0                   | 93                 |
| Mixed, small        | 0 <sup>†</sup>   | 5 <sup>†</sup>  | 8              | 14 <sup>†</sup> | 2              | 0                 | 6 <sup>†</sup>  | 20              | 2               | 0                   | 82                 |
| Mixed, medium       | 0                | 14 <sup>†</sup> | 3 <sup>†</sup> | 34              | 9 <sup>†</sup> | 0                 | 2               | 15 <sup>†</sup> | 6               | 0                   | 90                 |
| Mixed, large        | 0                | 2 <sup>†</sup>  | 1              | 14 <sup>†</sup> | 8              | 1 <sup>†</sup>    | 0               | 2               | 5 <sup>†</sup>  | 9                   | 71                 |
| Mixed, very large   | 0                | 0 <sup>†</sup>  | 0              | 0               | 1 <sup>†</sup> | 0                 | 0               | 0               | 2               | 2 <sup>†</sup>      | 60                 |
| Conifer, small      | 0 <sup>†</sup>   | 1               | 9 <sup>†</sup> | 17              | 0              | 0                 | 34              | 40 <sup>†</sup> | 3               | 2                   | 78                 |
| Conifer, medium     | 1                | 2               | 5              | 26 <sup>†</sup> | 1              | 0                 | 20 <sup>†</sup> | 118             | 17 <sup>†</sup> | 2                   | 94                 |
| Conifer, large      | 0                | 0               | 0              | 6               | 2 <sup>†</sup> | 0                 | 1               | 16 <sup>†</sup> | 40              | 26 <sup>†</sup>     | 92                 |
| Conifer, very large | 0                | 0               | 0              | 0               | 7              | 1 <sup>†</sup>    | 0               | 4               | 37 <sup>†</sup> | 69                  | 91                 |
| % correct           | 92               | 60              | 22             | 24              | 22             | 0                 | 49              | 54              | 35              | 63                  | 45                 |
| % within one class  | 92               | 96              | 84             | 84              | 72             | 100               | 87              | 86              | 88              | 88                  | 87                 |

\*Open, <1.5 m<sup>2</sup>/ha basal area (BA) and quadratic mean diameter of dominant trees (QMD) <50 cm or <10 m<sup>2</sup>/ha BA and QMD ≥50 cm; broadleaf, ≥1.5 m<sup>2</sup>/ha BA, ≥65% of which is broadleaf; mixed, mixed conifer-broadleaf, ≥1.5 m<sup>2</sup>/ha BA, 20–64% of which is broadleaf; conifer, ≥1.5 m<sup>2</sup>/ha basal area, <20% of which is broadleaf; small, 2.5–25.4 cm QMD; medium, 25.5–50.4 cm QMD; large, 50.5–75.4 cm QMD; very large, >75.4 cm QMD.

<sup>†</sup>Correct within one class, where class similarity is defined by both species composition and size class.

**Table 7.** Prediction accuracy (kappa coefficient of agreement; see Cohen 1960) by the gradient nearest neighbor method for vegetation classes.

| Vegetation class*   | $\kappa$ | $\kappa$ , correct within one class* |
|---------------------|----------|--------------------------------------|
| Open                | 0.51     | 0.98                                 |
| Broadleaf           | 0.49     | 0.94                                 |
| Mixed, small        | 0.12     | 0.67                                 |
| Mixed, medium       | 0.20     | 0.80                                 |
| Mixed, large        | 0.17     | 0.72                                 |
| Mixed, very large   | -0.03    | 0.75                                 |
| All mixed           | 0.30     | na <sup>†</sup>                      |
| Conifer, small      | 0.32     | 0.84                                 |
| Conifer, medium     | 0.43     | 0.86                                 |
| Conifer, large      | 0.31     | 0.87                                 |
| Conifer, very large | 0.54     | 0.88                                 |
| All conifer         | 0.59     | na                                   |

\*See Table 6 for definitions of vegetation classes and for classes that are within one class.

<sup>†</sup>na, not applicable.

Prediction error in GNN also is introduced by temporal differences between the satellite imagery and field plot measurement. We reduced this source of error by excluding plots that had been heavily disturbed (i.e., clear-cut) between the dates of imagery and ground measurement. In addition, we used two imagery dates (1988 and 1996) and paired plots with the imagery date closest to plot measurement for assigning the spectral values. This reduced the maximum temporal mismatch from 12 to 4 years, but the corresponding reduction in error is unknown. The validity of using multiple imagery dates rests on the assumption that given spectral values at different points in time result from similar vegetation. We minimized violations of this assumption by applying a histogram equalization function among scenes within each year and between the 2 years. Gradual changes in forest vegetation (tree growth and mortality) over as much as a 4-year period were not accounted for in our predictions.

#### Advantages of GNN for ecological analysis and integrated forest assessment

Vegetation maps produced with GNN have several advantages for ecological analysis, simulation modeling, and integrated forest assessment. These advantages derive from the use of imputation or direct gradient analysis and are shared with other methods that employ these techniques. However, CCA has only recently been employed in predictive vegetation mapping (Hill 1991; Gottfried et al. 1998; Guisan et al. 1999), and we know of only one other case (Gottfried et al. 1998) where CCA and imputation have been used together to predictively map vegetation. Firstly, information content is high, because each pixel is attributed with a list of trees by species, size, and density. Of particular note is the species-level detail contained in the map. Because the vegetation data are preserved at this most basic level, user-defined classification schemes can be applied, maps constructed, and accuracy assessed for specific analytical purposes. Furthermore, many vegetation simulation models require input data in the form of tree lists. Secondly, because we impute a single nearest-neighbor plot to each pixel, the

covariance among predicted species and structures within map units is ecologically realistic. Despite the importance of realistic correlation structures for ecological applications, few previous studies have addressed the distortion of correlation structure in predictive vegetation mapping (but see Moeur and Stage 1995; Nilsson 1997). Thirdly, the range of variability present in the sampled stands is maintained in the mapped predictions. If the ground sample is representative of the entire regional landscape, then the GNN procedure will reflect the inherent variability of the region. Fourthly, direct gradient analysis contributes to knowledge about regional ecological gradients. Our CCA model, which used both environmental and Landsat TM imagery, explained substantially more variation than one based on Landsat TM imagery alone.

#### Species response models in multispecies mapping

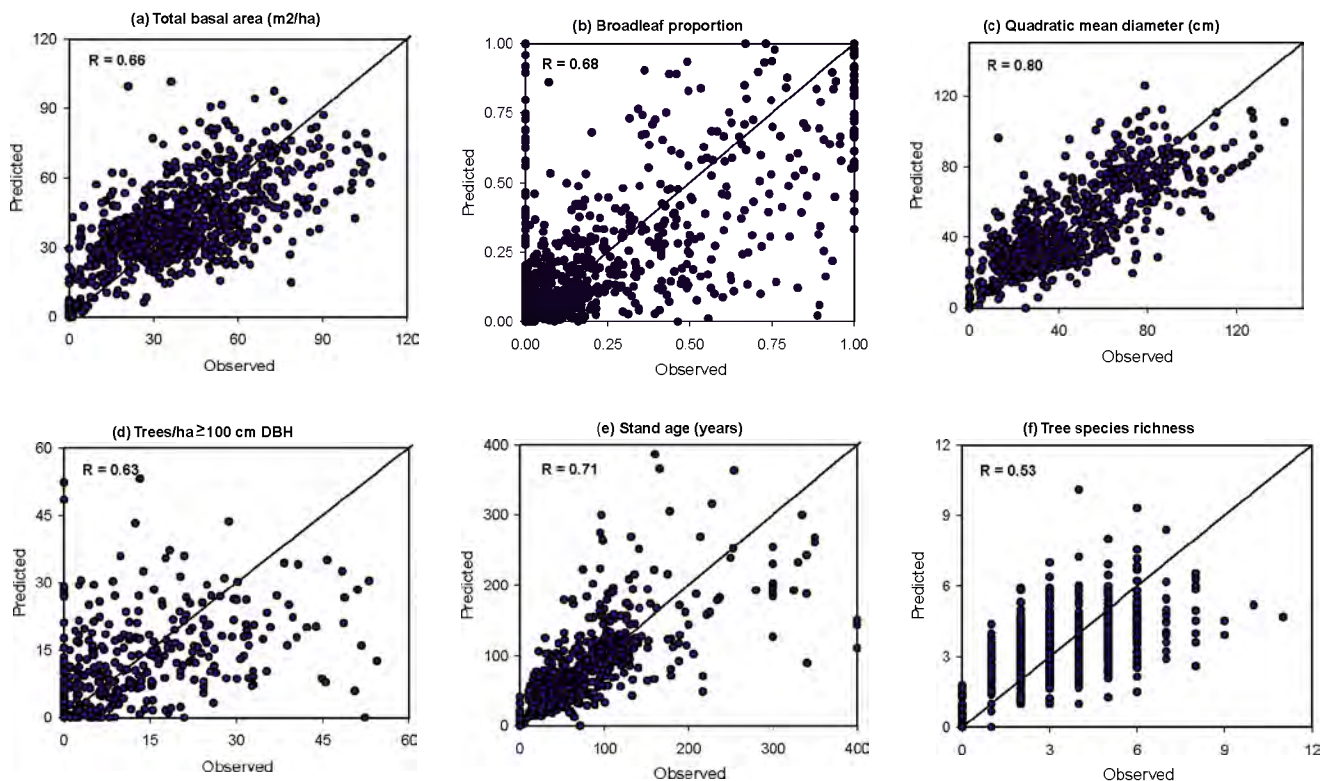
The weighted averaging algorithm in CCA implies Gaussian (unimodal) response curves of species to the environment. Use of statistical methods that assume unimodal responses for the prediction of individual species has been criticized on the basis of empirical evidence of other response patterns in nature (Austin et al. 1994). However, because our study objectives required that we simultaneously predict multiple species and structures, it was necessary to use a model that assumes a single type of response for all species. Methods that model responses of single species lose information about the co-occurrence of multiple species within samples (Gottfried et al. 1998), whereas CCA makes use of this information in the weighted averaging algorithm. Single-species models often will yield better predictions than a multispecies model for the same species (Guisan et al. 1999). However, our approach insures that predicted plant communities are realistic assemblages of species and structures. It is likely that if all individual species distributions were predicted independently and then assembled into communities, unrealistic collections of species would result (Moeur and Stage 1995).

#### Application of the GNN method and mapped predictions

The GNN method predicts extremely well at the regional level and moderately well to poorly for specific sites, similar to other Landsat TM image classifications in our region. However, there exists a danger that the fine spatial resolution and detailed information content of the GNN predictions may imply a higher level of precision than actually exists. We stress that vegetation maps produced with GNN are appropriately used for strategic-level planning and policy analysis, not to guide local management decisions.

Vegetation maps produced with GNN, but using a slightly different model specification, are now being used to initialize current landscape conditions for input to simulation modeling as part of the CLAMS (Spies et al. 2002). Because we developed the GNN method in the context of this research study, we did not formally evaluate operational considerations. Nevertheless, it appears that GNN could be successfully applied to other regions where a representative sample of georegistered field plots and mapped spectral and environmental data are available.

**Fig. 10.** Comparison of predictions from gradient nearest neighbor method to ground observations on  $n = 823$  field plots. (a) Total tree basal area ( $\text{m}^2/\text{ha}$ ). (b) Proportion of total tree basal area that is broadleaf. (c) Quadratic mean diameter (cm) of all dominant and codominant trees. (d) Number of trees per hectare  $\geq 100$  cm DBH. (e) Mean age (years) of dominant and codominant trees. (f) Tree species richness (number of species).



**Table 8.** Prediction accuracy (proportion of plots and kappa coefficient of agreement (see Cohen 1960) for  $n = 823$  plots for presence-absence of tree species.

| Tree species                 | Proportion of plots correctly classified | $\kappa$ |
|------------------------------|--|----------|
| <i>Acer macrophyllum</i>     | 0.56                                     | 0.24     |
| <i>Alnus rubra</i>           | 0.66                                     | 0.24     |
| <i>Picea sitchensis</i>      | 0.83                                     | 0.53     |
| <i>Pseudotsuga menziesii</i> | 0.89                                     | 0.22     |
| <i>Quercus garryana</i>      | 0.87                                     | 0.34     |
| <i>Thuja plicata</i>         | 0.58                                     | 0.21     |
| <i>Tsuga heterophylla</i>    | 0.61                                     | 0.22     |

## Acknowledgments

This study would not have been possible without the efforts of countless individuals in the regional forest inventory programs of the USDA Forest Service and USDI Bureau of Land Management. Jim Alegria provided data and counsel on use of the NRI inventory data and on accuracy assessment methods. Alissa Moses compiled the multiple-scene and multiple-date Landsat imagery with input from Tom Maersperger and Karin Fassnacht. Jonathan Brooks and K. Norm Johnson provided their map of nonforest land uses. Barbara Marks wrote early versions of the GNN programs. Tom Spies offered insightful suggestions throughout all

phases of this study and generously provided data from the Old Growth Study. We thank Ray Czaplewski, Janet Franklin, Mark Hanus, Andy Hudak, Tom Spies, Dale Weyermann, and Mike Wimberly for thoughtful comments on earlier versions of the manuscript. We dedicate this paper to John Gray, who was instrumental in conceptualization and data base development during early stages of this study.

## References

- Austin, M.P., Nicholls, A.O., Doherty, M.D., and Meyers, J.A. 1994. Determining species response functions to an environmental gradient by means of a  $\beta$ -function. *J. Veg. Sci.* **5**: 215–228.
- Beers, T.W., Dress, P.E., and Wensel, L.C. 1966. Aspect transformation in site productivity research. *J. For.* **64**: 691–692.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**: 37–46.
- Cohen, W.B., and Spies, T.A. 1992. Estimating structural attributes of Douglas-fir/western hemlock forest stands from Landsat and SPOT imagery. *Remote Sens. Environ.* **41**: 1–17.
- Cohen, W.B., Spies, T.A., and Fiorella, M. 1995. Estimating the age and structure of forests in a multi-ownership landscape of western Oregon, U.S.A. *Int. J. Remote Sens.* **16**: 721–746.
- Cohen, W.B., Maersperger, T.K., Spies, T.A., and Oetter, D.R. 2001. Modeling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. *Int. J. Remote Sens.* **22**: 2279–2310.
- Cohen, W.B., Spies, T.A., Alig, R.J., Oetter, D.R., Maersperger, T.K., and Fiorella, M. 2002. Characterizing 23 years (1972–



- 1995) of stand replacement disturbance in western Oregon forests with Landsat imagery. *Ecosystems*, **5**: 122–137.
- Curran, P.J., and Hay, A.M. 1986. The importance of measurement error for certain procedures in remote sensing at optical wavelengths. *Photogramm. Eng. Remote Sens.* **52**: 229–241.
- Daly, C., Neilson, R.P., and Phillips, D.L. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**: 140–158.
- Franklin, J. 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geogr.* **19**: 474–499.
- Franklin, J.F., and Dyrness, C.T. 1973. Natural vegetation of Oregon and Washington. USDA For. Serv. Gen. Tech. Rep. PNW-8.
- Franklin, S.E. (Editor). 2001. Remote sensing for sustainable forest management. CRC Press, Boca Raton, Fla.
- Gauch, H.G. 1982. Multivariate analysis in community ecology. Cambridge University Press, New York.
- Gottfried, M., Pauli, H., and Grabherr, G. 1998. Prediction of vegetation patterns at the limits of plant life: a new view of the alpine-nival ecotone. *Arct. Alp. Res.* **30**: 207–221.
- Guisan, A., and Zimmerman, N.E. 2000. Predictive habitat distribution models in ecology. *Ecol. Modell.* **135**: 147–186.
- Guisan, A., Weiss, S.B., and Weiss, A.D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* **143**: 107–122.
- Harmon, M.E., and Marks, B. 1995. Programs to estimate the solar radiation for ecosystem models. Department of Forest Science, Oregon State University, Corvallis, Oreg. Available from <http://www.fsl.orst.edu/lter/data/tools/software/solarrad.cfm> [cited 4 April 2002].
- He, S.H., Mladenoff, D.J., Radeloff, V.C., and Crow, T.R. 1998. Integration of GIS data and classified satellite imagery for regional forest assessment. *Ecol. Appl.* **8**: 1072–1083.
- Hill, M.O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *J. Biogeogr.* **18**: 247–255.
- Hunsaker, C., Goodchild, M., Friedl, M., and Case, T. 2001. Spatial uncertainty in ecology. Springer-Verlag, New York.
- Isaacs, E.H., and Srivastava, R.M. 1990. Applied geostatistics. Oxford University Press, Oxford, U.K.
- Iverson, L.R., and Prasad, A.M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecol. Monogr.* **68**: 465–485.
- Jongman, R.H.G., ter Braak, C.J.F., and van Tongeren, O.F.R. (Editors). 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen, the Netherlands.
- Kauth, R.J., and Thomas, G.S. 1976. The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. In *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*. Purdue University, West Lafayette, Ind. pp. 4B-41 – 4B-50.
- Lillesand, T.M., and Kiefer, R.W. 1994. Remote sensing and image interpretation. 3rd ed. John Wiley & Sons, New York.
- Lister, A., Riemann, R., and Hoppus, M. 2000. Use of regression and geostatistical techniques to predict tree species distributions at regional scales. In *Proceedings of the 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs*, 2–8 Sept. 2000, Banff, Alta. University of Colorado, Boulder, Colo. Available from <http://www.colorado.edu/research/cires/banff/pubpapers/107/> [cited 4 April 2002].
- Little, E.L., Jr. 1979. Checklist of United States trees. U.S. Dep. Agric. Agric. Handb. 541.
- Lowell, K., and Jaton, A. (Editors). 1999. Spatial accuracy assessment: land information uncertainty in natural resources. Ann Arbor Press, Chelsea, Mich.
- Max, T.A., Schreuder, H.T., Hazard, J.W., Oswald, D.O., Teply, J., and Alegria, J. 1996. The Pacific Northwest Region vegetation inventory and monitoring system. USDA For. Serv. Res. Pap. PNW-RP-493.
- Moeur, M., and Hershey, R.R. 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. In *Spatial accuracy assessment: land information uncertainty in natural resources*. Edited by K. Lowell and A. Jaton. Ann Arbor Press, Chelsea, Mich. pp. 419–429.
- Moeur, M., and Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. *For. Sci.* **41**: 337–359.
- Nilsson, M. 1997. Estimation of forest variables using satellite image data and airborne lidar. Ph.D. thesis, Swedish University of Agricultural Sciences, Umeå, Sweden.
- Ohmann, J.L., and Spies, T.A. 1998. Regional gradient analysis and spatial pattern of woody plant communities of Oregon forests. *Ecol. Monogr.* **68**: 151–182.
- Ohmann, J.L., McComb, W.C., and Zumrawi, A.A. 1994. Snag abundance for cavity-nesting birds on nonfederal forest lands in Oregon and Washington. *Wildl. Soc. Bull.* **22**: 607–620.
- Palmer, M. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology*, **74**: 2215–2230.
- Spies, T.A., and Franklin, J.F. 1991. The structure of natural young, mature, and old-growth Douglas-fir forests in Oregon and Washington. USDA For. Serv. Gen. Tech. Rep. PNW-GTR-285. pp. 90–109.
- Spies, T.A., Reeves, G.H., Burnett, K.M., McComb, W.C., Johnson, K.N., Grant, G., Ohmann, J.L., Garman, S.L., and Bettinger, P. 2002. Assessing the ecological consequences of forest policies in a multi-ownership province in Oregon. In *Integrating landscape ecology into natural resource management*. Edited by J. Liu and W.W. Taylor. Cambridge University Press, New York. In press.
- ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**: 1167–1179.
- ter Braak, C.J.F., and Prentice, I.C. 1988. A theory of gradient analysis. *Adv. Ecol. Res.* **18**. pp. 271–313.
- ter Braak, C.J.F., and Smilauer, P. 1998. CANOCO reference manual and user's guide to CANOCO for Windows: software for canonical community ordination (version 4). Microcomputer Power, Ithaca, N.Y.
- Tokola, T., Pitkanen, J., Partinen, S., and Muinonen, E. 1996. Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. *Int. J. Remote Sens.* **17**: 333–351.
- Tomppo, E. 1990. Designing a satellite image-aided National Forest Survey in Finland. In *Proceedings of the SNS/IUFRO Workshop on the Usability of Remote Sensing for Forest Inventory and Planning*, 26–28 Feb. 1990, Umeå, Sweden. International Union of Forest Research Organizations, Vienna. pp. 43–47.
- Van Deusen, P.C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. *Can. J. For. Res.* **27**: 379–384.
- Walker, G.W., and MacLeod, N.S. 1991. Geologic map of Oregon, scale 1:500,000, two sheets. U.S. Geological Survey, Denver, Colo.
- Weyermann, D., and Fassnacht, K. 2000. The Interagency Vegetation Mapping Project: estimating certain forest characteristics

- using Landsat TM data and forest inventory plot data. *In* Proceedings of the 8th Biennial Forest Service Remote Sensing Applications Conference, 10–14 April 2000, Albuquerque, N.M. CD-ROM. American Society of Photogrammetry and Remote Sensing, Bethesda, Md.
- Wimberly, M.C., and Spies, T.A. 2001*a*. Influences of environment and disturbance on forest patterns in coastal Oregon watersheds. *Ecology*, **82**: 1443–1459.
- Wimberly, M.C., and Spies, T.A. 2001*b*. Modeling landscape patterns of understory tree regeneration in the Pacific Northwest, USA. *Appl. Veg. Sci.* **4**: 277–286.
- Wolter, P.T., Mladenoff, D.J., Host, G.E., and Crow, T.R. 1995. Improved forest classification in the northern Lake States using multi-temporal Landsat imagery. *Photogramm. Eng. Remote Sens.* **61**: 1129–1143.
- Woodcock, C.E., Collins, J.B., Gopal, S., Jakabhazy, V.D., Li, X., Macomber, S., Ryherd, S.J., Harward, V.J., Levitan, J., Wu, Y., and Warbington, R. 1994. Mapping forest vegetation using Landsat TM imagery and a canopy reflectance model. *Remote Sens. Environ.* **50**: 240–254.