Moderate-Resolution Data and Gradient Nearest Neighbor Imputation for Regional-National Risk Assessment

Kenneth B. Pierce, Jr.; C. Kenneth Brewer; and Janet L. Ohmann

Kenneth B. Pierce, Jr., research ecologist, USDA Forest Service, Forestry Sciences Lab, Corvallis, OR 97331 (now Department of Fish and Wildlife, Olympia, WA); **C. Kenneth Brewer**, IAAA Program Leader, USDA Forest Service, Remote Sensing Applications Center, Salt Lake City, UT 84119; and **Janet L. Ohmann**, research ecologist, USDA Forest Service, Forestry Sciences Lab, Corvallis, OR 97331.

Abstract

This study was designed to test the feasibility of combining a method designed to populate pixels with inventory plot data at the 30-m scale with a new national predictor data set. The new national predictor data set was developed by the USDA Forest Service Remote Sensing Applications Center (hereafter RSAC) at the 250-m scale. Gradient Nearest Neighbor (GNN) imputation was designed by the USDA Forest Service Pacific Northwest Research Station (hereafter PNW) to assign a plot identifier, and, therefore, a link to associated plot data, to each pixel within a target raster. Gradient Nearest Neighbor was implemented at 30-m resolution in three separate multimillion-hectare regions of the Western United States. Concurrently, RSAC developed a set of spatial predictor surfaces at 250-m resolution for use in producing nationally consistent data products. These data have been used for modeling forest types and forest biomass for the conterminous United States and Alaska. These predictor data have also been used for large regional applications.

In this study, we substituted the 250-m predictor data for the 30-m predictor data used thus far in GNN. Our objective was to quantify the difference in performance using the lower spatial resolution predictors. We remodeled the same three regions that were mapped at 30 m with the 250-m data set and compared the error structure of the two modeling efforts. For species presence/absence models in the two areas with large environmental gradients, the Sierra Nevada and northeastern Washington, the species models performed substantially the same at the two resolutions. For the region with reduced environmental heterogeneity and moderate environmental gradients, coastal Oregon, species models did not work well with either the 30-m or 250-m studies. Models geared towards mapping forest structure did not perform as well as the 30-m models and may be insufficient for risk-assessment use.

Keywords: Gradient Nearest Neighbor, imputation, regional analysis, species distributions, vegetation mapping.

Introduction

A great wealth of resources has been expended to inventory our Nation's forests, and an equally substantial amount of effort has gone into acquiring remotely sensed data. As such, these two data types make up the ends of a continuum of detail. Plot inventory data are extremely sparse geographically but have a high level of information content regarding the resources at the inventory plot locations. Conversely, remotely sensed data cover the entire globe but with comparatively limited information at any single location. The common approach to leverage these two forms of data has been to create thematic maps of vegetation-related classes as well as response surfaces for other target variables of interest. With regard to vegetation mapping, these thematic maps typically describe dominant vegetation and include physiognomic, floristic, or structural characteristics, or all. These thematic maps often include some additional land use classes or map land use as a separate theme. The variables available for analysis are limited to the classes included in the map, and once the analysis is complete there is no ability to develop new attributes without a new mapping effort. Recently, a more flexible approach, single neighbor imputation, has been utilized to provide sample tree lists and plot-calculated variables for all the unsampled pixels in raster maps. Although not replacing traditional mapping methods, imputed maps can greatly enhance analytical flexibility and provide information in a familiar context that is often supported by extensive simulation modeling capability. Imputed maps are not intended to suggest that each pixel is in fact occupied by the imputed plot data, but rather given

current information, what do we expect. However, developing and mapping 30-m products over broad spatial extents is a lengthy process. This project was conducted in order to evaluate the differences in using a new national spatial predictor database at a coarser resolution (250 m) instead of the 30-m data used in the previous study.

Existing Methods for Regional-National Vegetation Mapping and Risk Assessment

Traditional Methods—

In recent years, many regions within the USDA Forest Service have implemented midlevel classification and mapping programs to provide thematic maps of existing vegetation for a wide variety of analysis applications. These programs are becoming more similar as they implement the USDA Forest Service direction established by the *Existing Vegetation Classification and Mapping Technical Guide* (hereafter technical guide) (Brohman and Bryant 2005). The vegetation classifications and mapping methodologies of these programs follow the technical guide's midlevel direction (Brewer and others 2005), and most use satellite remote sensing approaches to provide synoptic coverage of the mapping areas (e.g., Mellin and others 2004). Some of these programs also utilize summary databases populated by the USDA Forest Service Forest Inventory and Analysis (hereafter FIA) data to develop quantitative map unit descriptions including estimates of common inventory variables. This approach provides thematic map products with statistically sound estimates of inventory variables. These estimates are explicitly connected to the vegetation pattern depicted in the thematic map products. The approach is designed to support midlevel and broad-level analysis applications, as well as some project-level cumulative effects analyses.

As suggested in the technical guide, these midlevel map products can be used as is or rescaled for a variety of base-level analysis applications including project support, risk analysis, and watershed analyis at the scale of 4^{th} and 5^{th} Hydrologic Units, (USDA FS 1995). Unfortunately, the geographic extent of these base-level analysis applications is normally too small to effectively use FIA data as the inventory data source. Given the extensive design of the FIA program without spatial intensification, each plot represents approximately 6,000 acres. This leaves forests and ranger districts faced with the difficult choice of using the midlevel map product as is with an inadequate sample size of associated FIA data or reverting to the use of often biased and outdated stand-exam data that cannot provide defensible statistical estimates and have no explicit relationship to the midlevel map data used for forest plan revision and project cumulative effects analyses. Alternatives to this untenable choice include several expensive and logistically difficult inventory approaches including intensifying the base grid to provide an adequate sample size or implementing a new traditional two-stage sample of the map features depicting vegetation pattern.

Imputation Methods—

A primary information need of land managers is consistent and continuous current vegetation data on each and every parcel of land in an analysis area sufficient to address the principal issues and resource concerns. As discussed above, where these data do exist they are normally based on a sampling inference procedure rather than wall-to-wall inventory data. Many of the analyses needed to address multiple resource issues at the project level are essentially analyses of vegetation pattern and process relationships through time and space. Inventory data based on traditional two-stage sampling or quantitative map unit descriptions from a systematic random grid are not sufficient to address the spatial or the temporal dimensions or both of these analyses. These data are not spatially explicit enough to identify important vegetation pattern relationships and do not provide adequate thematic detail (i.e., plot-level tree list data) for simulating vegetation change through time.

The ability to simulate these vegetation pattern relationships through space and time, particularly with a variety of management and disturbance alternatives, is important for effective land and resource planning. Despite the capability of simulation models and decision-support tools, comprehensive landscape-level planning is still difficult to implement because the inventory data are rarely complete or current or both. For planning purposes, it would be convenient to be able to operate as if detailed inventory

information were available for all units in the planning area (Moeur and Stage 1995).

As an alternative to historically common statistical approaches (e.g., regression estimates or stratum averages) to populating unsampled units with data, imputation can be used. Imputation involves estimating values for variables of interest (Y variables) by supplying realistic measurements from one or more sampled units to unsampled units with similar characteristics in auxiliary (X) variable space (Hassani and others 2004, LeMay and Temesgen 2005, McRoberts and others 2002, Moeur and Stage 1995, Ohmann and Gregory 2002, Temesgen and Gadow 2004). These auxiliary (X) variables typically include biophysical characteristics such as slope, aspect, precipitation, etc., as well as data from remotely sensed imagery such as aerial photography or satellite imagery.

Imputation of inventory data from sampled areas to similar unsampled areas produces data sets that function like wall-to-wall data for planning purposes. There are many methods and variations of imputation, both univariate and multivariate; however, multivariate approaches that impute a single plot tend to produce more realistic data sets for simulation modeling because they retain the original covariance structure of actual sample units. LeMay and Temesgen (2005) provided a brief summary of common imputation approaches and a detailed comparison of variable-space nearest neighbor (NN) methods for estimating basal area and stems per hectare using aerial auxiliary variables. LeMay and Temesgen (2005) also summarized variable-space nearest neighbor methods and compared them to other estimation methods. These summaries were the most comprehensive available in current literature when this paper was written.

In recent years, two modeling approaches have been developed that could potentially address this critical need through the imputation of inventory data. The first of these approaches, Most Similar Neighbor (MSN), was developed by Moeur and Stage (1995) to impute attributes measured on some sample units (e.g., stand polygons) to sample units where they were not measured. The MSN was originally designed to use a traditional two-stage inventory of forest stands, as described by Stage and Alley (1972), imputing stand data to unsampled stands. The second of these approaches, Gradient Nearest Neighbor (GNN) developed by Ohmann and Gregory (2002), follows the same general analytical logic, but is designed to use vegetation information from regional grids of field plots (similar to an intensified FIA grid) with remotely sensed imagery and other spatial data to produce a continuous raster surface by imputing data from sampled grid cells to unsampled grid cells.

Study Objectives

Our primary objective was to quantify the difference in GNN model performance using the lower spatial resolution predictors. We remodeled the same three regions that were mapped at 30 m with the 250-m data set and compared the error structure of the two modeling efforts. As explained below, two effects occur when 30-m data are replaced with 250-m data, and both involve averaging across multiple pixels.

One of the reasons for implementing this study is that development of spatial products and modeling at 30 m could take several years to complete a large portion of the Western United States. We anticipate much faster turnaround if 250-m modeling proves sufficient. The 250-m data products could potentially be available for large areas such as the Western United States in 1 to 2 years of production.

Methods

The Study Area

Three western regions covering temperate steppe, coastal forest, and Mediterranean ecosystems were mapped using GNN imputation for a Joint Fire Sciences Program study (Figure 1). The original study examined the feasibility of mapping wildland fuels and vegetation structure to provide data for fire and fuels management planning (Pierce and others 2009, Wimberly and others 2003). The 2.86-million-ha coastal forest site was located in the coast range of Oregon extending as far inland as the western edge of the Willamette Valley (Figure 1). The forests are primarily coniferous with hardwoods occupying riparian and disturbed areas. The 4.1-million-ha Mediterranean site was located in the central Sierra Nevada occupied by savannah,

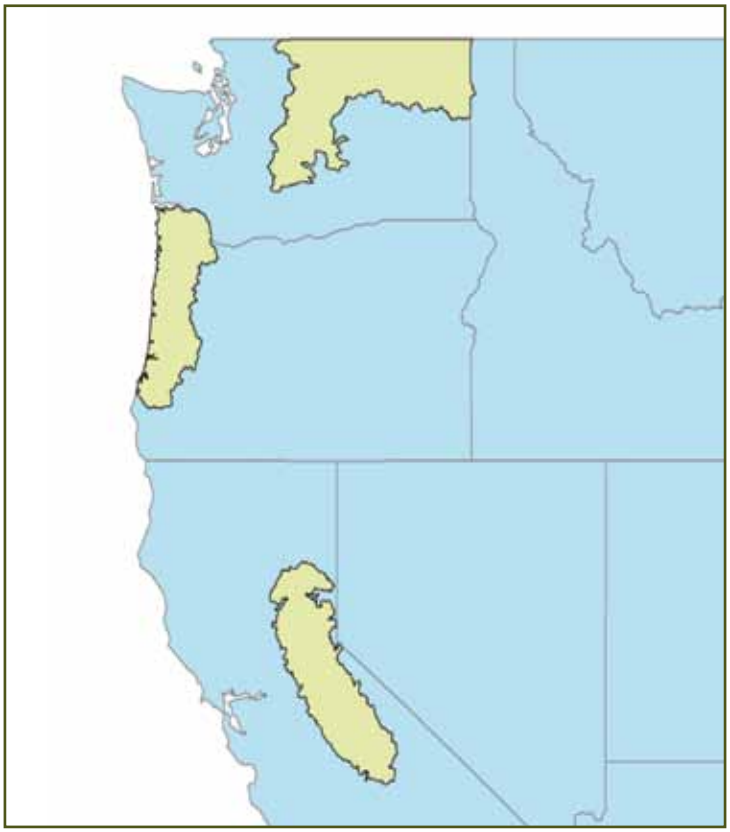

Figure 1—The Joint Fire Sciences project covered three western sites in contrasting ecosystems.

chaparral, mixed conifer, and alpine woodlands vegetation types. The site stretches from the northern border of Sequoia National Park north through the Plumas National Forest. The 5-million-ha temperate steppe in northeastern Washington was bounded on the west by the Cascade crest and on the south by the Columbia and Spokane Rivers (Figure 1). The temperate steppe site is dominated by a combination of mixed coniferous forest and extensive shrub steppe.

Vegetation Data from Field Plots

Vegetation data from regional inventories were derived in each of the three regions from multiple sources including Forest Inventory and Analysis (FIA) plots, Current Vegetation Survey/Pacific Northwest Region (R6) plots (CVS), Pacific Southwest Region (R5), Bureau Land Management (BLM), research Ecology Plots in North Cascades National Park (NCNP) (provided by Dave L. Peterson) and Yosemite National Park (provided by Jan Van Wagtendonk). The FIA, R5, and CVS plots were installed on systematic grids. The CVS/R6 and R5 plots covered the national forests whereas FIA installed plots on all ownerships. The FIA and CVS inventory plots used five subplot arrays within a 1-ha area. Small trees, snags, coarse woody debris line-intercept transects, and ground cover were sampled on each subplot.

Because vegetation data were derived from multiple inventories with different sampling protocols, all individual tree records were converted to per-hectare values. For plots with multiple vegetation or land cover conditions, only the forested portion was used with expansion factors adjusted accordingly. Plot-level summary variables were calculated for each plot.

Stand-summary variables included total basal area, basal area by species, trees per hectare, quadratic-mean diameter, snags per hectare, percentage of tree canopy cover, and down-wood volume. Different inventories collected down-wood data using different sampling schemes

Table 1—Comparison of spatial data between the 30-m and 250-m analyses

250-m resolution	30-m resolution (Pierce and others 2009)
Elevation, slope, and aspect	Elevation, slope, and aspect
DAYMET shortwave radiation	Potential Relative Radiation
Ecological region layers	NA
USGS mapping zones, Bailey's ecoregions, and unified ecoregions for Alaska	NA
STATSGO data layers (lower 48)	NA
Available water capacity, soil bulk density, soil permeability, soil PH, soil porosity, soil plasticity, depth to bedrock, rock volume, soil types, and soil texture	None
MODIS Vegetation indices such as EVI, NDVI	Landsat TM band ratios, 3/4, 4/5, 5/7
MODIS Vegetation continuous fields	NA
MODIS fire points developed from the MODIS active fire maps	NA
MODIS 8-day composites, multiple dates from the spring, summer, and fall	TM-5 and ETM-7 data
NA	Within footprint spectral variability
All MODIS 32-day composite imagery that is cloud-free between the years 2001–2003	NA
USGS NLCD layers, percentage tree cover, percent herbaceous cover, and percentage bare ground	NA
DAYMET temperature and precipitation, mins, maxs, aves, SD (all, yearly, monthly)	DAYMET temperature and precipitation, mins, maxs, aves (all)

and minimum sizes. As a result, we focused primarily on species basal area.

Moderate-Resolution Predictor Data

Beginning in 2003, RSAC, in cooperation with the FIA remote sensing band, developed a national predictors database to support FIA national mapping efforts (e.g., national forest type maps, and national biomass maps) (Blackard and others 2008, Ruefenacht and others 2008). The original database included about 60 layers consisting primarily of MODIS imagery. The national predictors database has been extensively used with additional data layers added each year. The current version has more than 700 layers, which includes DAYMET climate data, additional MODIS imagery, derived MODIS-based vegetation indices, STATSGO soil layers, topography, and several derived thematic products (Table 1). These data cover the entire conterminous United States as well as Alaska and Puerto Rico.

To prepare these data for Gradient Nearest Neighbor analysis, all layers were sampled with all plot locations. With 30-m data, 13-pixel footprints were used to sample the spatial data to account for plots occupying more area than a single 30-m pixel. With 250-m data, only a single pixel was necessary to represent each plot because the plot area is less than the area of one 250-m pixel. All predictor images were split into separate bands, comprising over 700 individual predictor layers. For example, some STATSGO soil layers comprised 11 soil horizons. Those 11 horizons were split into separate predictor layers. Before statistical modeling, the 700 bands were reduced to those without a correlation of > 0.90 with any other remaining predictor. For each of the three study areas, this left about 100 to 150 predictor layers. These predictors were entered as the predictor matrix in a stepwise canonical correspondence analysis. In each case, 10 to 15 predictors were sufficient to explain the primary gradients in vegetation composition.

Satellite Imagery

For the 30-m study, Landsat Thematic Mapper imagery (TM) was mosaiced and histogram matched. For the 250-m study, several products derived from MODIS imagery were used (Table 1). These products have a spatial extent covering our entire study regions, so no image matching was necessary.

Biophysical Environment Data

For both studies, biophysical data were derived from digital elevation models, including slope, aspect, and elevation. Both studies used climate data derived from DAYMET, although the 250-m data set used many more calculated variables. These variables included year-to-year variability by month. The DAYMET data are provided at 1-km resolution, so both the 30-m and 250-m DAYMET data were interpolated or resampled to their respective resolutions.

The 250-m data set also included STATSGO data layers, such as available water-holding capacity, soil bulk density, soil permeability, and soil pH. No analog for these data existed in the 30-m data set.

The Gradient Nearest Neighbor (GNN) Method of Predictive Vegetation Mapping

Imputation is a process in which values are assigned to unmeasured locations from either measured values or a statistical summary of a few selected measured values such as a mean (Moeur and Stage 1995). Unlike regression-based predictions, the assigned value is not a product of predictor variables and coefficients. The predictor variables are used to rank sample plots as to their similarity to a target location (30-m or 250-m-square pixel). In Gradient Nearest Neighbor imputation (GNN) we used the loadings for the ordination axes and their eigenvalues from a Canonical Correspondence Analysis (CCA) to relate the target locations to the locations of sample plots. This is achieved by calculating a Euclidean distance in eight-axis ordination space between the target pixel and each plot using the ordination loadings to weigh the spatial variables and the eigenvalues to weigh each axis. The distance in gradient space between the target location and each plot is used to rank the sample plots for potential assignment at each target pixel in our study regions (Ohmann and Gregory 2002). Because we use a multivariate response, which is analogous to a community representation of a plot, the variables selected must be of similar type. Our primary modeling scenarios have been species modeling, which involves using the total basal area of each individual species, and structure modeling, which has used basal area in different size classes of hardwoods and conifers, snag density, coarse woody debris volume, and canopy cover.

When GNN is run as a single neighbor imputation, the closest plot in gradient space is assigned to each pixel in the landscape. Once the assignment has been made, any attribute calculated for all plots can be mapped, maintaining the original covariance structure for any/all other attributes to be mapped. Additionally, the ranking of potential neighbors provides a sample neighborhood of similar plots from which natural variability and sample sufficiency can be evaluated (Pierce and others 2009).

Gradient Nearest Neighbor Results Summary from Previous Studies

In the 30-m study, coastal Oregon had favorable results for structure models but substandard results for species models. This pattern was reversed in both Washington and California, where species models worked well, but structure models did a poor job with attributes such as quadratic mean diameter and trees per hectare. The purpose of the original study was to map wildland fuels and vegetation structure. Wildland fuel components, such as coarse woody debris and snag density, were not mapped adequately in any of the three sites (However, canopy-related fuels variables were more satisfactory.) This was partially due to remote sensing not directly detecting wildland fuels, which, in general, are below the canopy. Another factor is that coarse woody debris data are collected on only a relatively short transect on FIA plots such that the resulting sample size is too small to characterize individual plots. The original design of the coarse woody debris sampling was to create estimates for a region.

We expect that species models will perform more favorably than structure models with 250-m data. In the previous study, remote sensing imagery was very important for mapping structure and has a much finer spatial grain than our climate data. Although we had climate data at 30 m, it was interpolated from 1-km-resolution DAYMET data, which are interpolated from weather stations plus higher resolution covariates (elevation, topography, etc.). Therefore, climate data would change gradually over the course of a kilometer whereas TM data can change abruptly from one

30-m pixel to the next. Therefore, our change in resolution loses much more information for predictions relying heavily on imagery than those relying on climate.

Model Evaluation and Accuracy Assessment

To evaluate the results of the 30-m study compared with the 250-m study results, we used two primary approaches. For continuous variables, we calculated 2nd nearest neighbor correlations (analogous to r-squares from cross-validation), (see Ohmann and Gregory 2002). For discrete variables, such as species presence/absence, we used standard confusion matrices. Producers accuracy, users accuracy, and Kappa statistics were calculated for each species and compared for the two studies (Wilkie and Finn 1996). The Kappa statistic accounts for the probability of randomly assigning a plot to its correct class. As such, the random probability of assigning a species with a frequency of 0.9 to any plot is very high; Kappa statistics tend to be low as the probability of improving upon randomness decreases. Producers accuracy is the proportion of sample plots with a species present in which the species is predicted to occur. Users accuracy is the proportion of plots with the predicted species occurrence that actually had the species in the inventory.

Any discrepancies derived from changing the scale of the analysis stems from two primary effects, which are both aspects of spatial averaging. First, with the 30-m product, the predicted value for a plot is the average of the 13 pixels imputed to the plot footprint. In this way, the predicted value is subject to averaging over those 13 imputations. With the 250-m imputation, only a single plot is imputed for the same ~1 ha space. Therefore, the 250-m results will be slightly more variable. The second effect involves the spatial averaging of the predictor data set. The predictor variables are the averages of the 13 pixels within the individual predictor layers. In addition, we calculate two texture indices for the remote sensing data, which provides an estimate of variability within the footprint. With the 250-m data, a single pixel covers an area considerably larger than a 1-ha plot (62 500 m^2 vs. 11 700 m^2). Thus, the spectral values are averaged over a larger area, and we have no estimate of within-plot variability.

Results

Gradients in Species Composition

Daubenmire (1952) noted the axiomatic relationship between climate and vegetation. Our CCA modeling results are consistent with this observation and suggest that climatic variables, as well as topographic and edaphic variables indirectly related to temperature and moisture, strongly influence the patterns of species composition. The gradients described by the three CCA models were composed of the dominant patterns in temperature and precipitation. In Washington, elevation, precipitation frequency, and brightness in MODIS 8-day composites separated warmer drought-tolerant conifers from both high-elevation wet and dry forests. In California, species were separated by September growing degree days, September average air temperature, April cooling degree days, and water vapor pressure variability in July. In Oregon, the dominant environmental variables were August mean precipitation, June cooling degree days, soil permeability, and June standard deviation of water vapor pressure.

Species Mapping Performance of GNN

Species performance from confusion matrices are listed for all species occurring with a frequency of at least 5 percent in the plot data set (Tables 2 and 3). California had the highest average Kappa statistics at 0.53 for the 30-m study and 0.48 for the 250-m study (Table 2). Washington was second with 0.46 and 0.43 (Table 3), followed by Oregon with 0.32 and 0.25 (Table 4). Patterns for producers and users accuracy were similar across sites, as were the actual values. In each case, the 30-m study had higher producers accuracy than the 250-m study by about 22 percent, whereas for users accuracy, the 250-m study was actually about 3 percent higher with an average across sites of 55 percent.

Structure Mapping Performance of GNN

To date we have only developed structure models for the Washington and Oregon study sites. Second nearest neighbor correlations for structure variables were generally low in both sites. In Washington, total basal area had an r-square of 0.06 compared to 0.17 for the 30-m analysis, 0.04 for snags per hectare compared to 0.16, and 0.01 for quadratic

Table 2—California species presence/absence results, N = 1,835 plots for species of ≥5 percent frequency on plots

	Kappa		Producers		Users		
Species	30 m	250 m	30 m	250 m	30 m	250 m	Plots
Abies concolor	0.51	0.47	0.88	0.67	0.64	0.68	761
Calocedrus decurrens	0.51	0.52	0.86	0.69	0.61	0.67	643
Pinus ponderosa	0.56	0.52	0.88	0.70	0.63	0.66	632
Quercus kelloggii	0.56	0.48	0.86	0.61	0.59	0.61	510
Pinus jeffreyi	0.49	0.46	0.82	0.61	0.54	0.60	488
Pinus lambertiana	0.42	0.36	0.80	0.54	0.49	0.50	468
Abies magnifica	0.63	0.55	0.85	0.65	0.63	0.65	389
Pseudotsuga menziesii	0.59	0.52	0.86	0.61	0.58	0.62	370
Pinus contorta	0.60	0.42	0.84	0.59	0.58	0.48	340
Quercus chrysolepis	0.54	0.43	0.78	0.50	0.52	0.51	264
Pinus monticola	0.52	0.39	0.70	0.48	0.49	0.43	189
Quercus wislizeni	0.53	0.53	0.71	0.59	0.47	0.52	108
Juniperus occidentalis	0.40	0.42	0.64	0.50	0.34	0.42	102
Quercus douglasii	0.53	0.67	0.65	0.69	0.48	0.69	89
Average	0.53	0.48	0.79	0.60	0.54	0.57	

Table 3—Washington species presence/absence results, N = 763 plots for species of ≥5 percent frequency on plots

	Kappa		Producers		Users		
Species	30 m	250 m	30 m	250 m	30 m	250 m	Plots
Pseudotsuga menziesii	0.43	0.43	0.97	0.89	0.85	0.86	1,835
Pinus ponderosa	0.53	0.52	0.91	0.74	0.65	0.69	971
Pinus contorta	0.32	0.41	0.82	0.63	0.50	0.61	827
Larix occidentalis	0.54	0.51	0.89	0.70	0.61	0.67	809
Abies lasiocarpa	0.52	0.50	0.87	0.65	0.58	0.64	707
Picea engelmannii	0.41	0.35	0.81	0.54	0.49	0.52	643
Abies grandis	0.57	0.55	0.87	0.68	0.58	0.64	587
Thuja plicata	0.53	0.51	0.82	0.61	0.51	0.58	393
Tsuga heterophylla	0.49	0.44	0.78	0.53	0.45	0.48	277
Pinus monticola	0.35	0.27	0.68	0.37	0.33	0.33	240
Abies amabilis	0.67	0.62	0.86	0.66	0.60	0.66	219
Tsuga mertensiana	0.58	0.49	0.81	0.52	0.50	0.52	179
Pinus albicaulis	0.55	0.47	0.77	0.50	0.47	0.51	152
Populus tremuloides	0.15	0.16	0.40	0.23	0.16	0.20	146
Betula papyrifera	0.29	0.26	0.51	0.30	0.25	0.29	119
Average	0.46	0.43	0.78	0.57	0.50	0.55	

mean diameter compared to an almost equally random 0.05 for the 30-m analysis. In Oregon, where we had quite good results for structure with 30-m data, we mapped basal area with an r-square of 0.09 compared to 0.59 for the 30-m analysis, 0.03 for snags per hectare compared to 0.09, and 0.08 for quadratic mean diameter compared to 0.69.

Discussion

Accuracy of GNN Vegetation Maps

For the Washington and California study sites, species distributions were modeled equally well with the 250-m data and the 30-m data. Both the Kappa statistics and visual inspection of species maps indicated essentially the same

Table 4—Oregon species presence/absence results, N = 2,325 plots for species of ≥5 percent frequency on plots

	Kappa		Producers		Users		
Species	**30 m**	**250 m**	**30 m**	**250 m**	**30 m**	**250 m**	**Plots**
Pseudotsuga menziesii	0.22	0.19	0.97	0.95	0.91	0.91	684
Alnus rubra	0.32	0.18	0.89	0.70	0.67	0.66	440
Tsuga heterophylla	0.27	0.27	0.85	0.63	0.52	0.60	320
Acer macrophyllum	0.25	0.26	0.78	0.49	0.42	0.47	228
Thuja plicata	0.11	0.06	0.56	0.24	0.26	0.25	150
Picea sitchensis	0.52	0.42	0.83	0.50	0.50	0.53	128
Abies grandis (*concolor*)	0.39	0.29	0.63	0.34	0.34	0.33	52
Quercus garryana	0.47	0.33	0.62	0.32	0.43	0.38	40
Average	0.32	0.25	0.77	0.52	0.51	0.52	

pattern when moving from 30-m to 250-m data. Because the gradient models were largely composed of climate variables, there is actually little loss in predictor data information when using the 250-m data. This is because the climate data for both the 250-m and 30-m studies were interpolated from the same 1-km-resolution data. Species performance in Oregon was not as good, though it was also less accurate for the 30-m data. In both the 30-m and 250-m studies, we saw some definite differences between the results for Oregon and the results for Washington and California. Both California and Washington are precipitation limited, receive most of their precipitation during winter, and have large elevation gradients. The Coast Range in Oregon has much higher precipitation, milder temperatures, and lower overall topographic variation resulting in less orographic precipitation. Coastal Oregon has also had a long history of timber management and, therefore, has a large patchwork of even-aged stands.

Sources of Error in GNN—

The GNN and all nearest neighbor techniques, are particularly susceptible to errors introduced by natural variability at spectrally and environmentally similar sites. Whereas regression techniques model a trend and the departure from that trend, imputation retains the full range of variability within a data set. As such, for a certain location, a regression model with little predictive capability will predict the mean plus some small departure based on predictor variables and coefficients, whereas imputation will find the most environmentally similar site and select it. The tendency for imputation to impute values similar to the actual target values is constrained by the strength of the relationship between available spatial predictor variables and the target response variables.

Other sources of error include (1) residual spatial error of predictor data sets and plot locations as well as plot registration, (2) temporal mismatches between inventory dates and imagery dates, and (3) the lack of adequate spatial data on disturbance and management history across large regions.

Advantages of GNN for Risk Assessment—

There are several advantages to GNN for risk assessment. The GNN technique retains the covariance structure for multiple attributes by imputing whole plots and provides mapped estimates of natural variability and sample sufficiency (Pierce and others 2009). Comparative risk assessment requires spatially explicit data with estimates of variability (Borchers 2005) in order to create probability surfaces for different management scenarios. For instance, what is the probability of the desired outcome given two different management choices, and are they statistically different? Without an estimation of uncertainty, this type of analysis can't be performed. By using a set of multiple potential neighbors, the variability in potential neighbors for a selected attribute can be mapped. In addition, by using the frequency distribution of all interplot distances in gradient space, thresholds for closeness in gradient space can be assigned and the number of candidate plots within a threshold calculated. This gives an indication as to whether

or not the inventory can provide adequate information for a certain pixel, and, as such, a map depicting the sampling support can be created.

Species Response Models in Multispecies Mapping— One of the key areas of interest in natural resource risk assessment is the interactions among species. The location of invasive species and the presence of host species are two data surfaces of interest to managers. Mapping with single neighbor imputation ensures that the assemblages of tree species mapped are consistent with actual inventoried assemblages. This has both benefits and limitations. The benefit is robust assemblages of species as they currently exist. The limitation is that prediction for new interactions can not be inferred on the basis of these maps. Single-species models are probably best suited for predicting suitable habitat for an individual species, or rather the present distribution of habitat consistent with currently occupied habitat.

Risk Assessment Applications of GNN Predictions— Single-neighbor imputation using GNN provides a very flexible wall-to-wall data set that includes any variable that can be derived from those measured on all inventory plots. This includes the ability to derive new variables or vegetation classifications after the initial modeling. A GNN imputation also provides a link to the full tree lists allowing for almost any kind of ecological modeling. The inclusion of multiple neighbors provides uncertainty data for Monte Carlo simulations or analyses seeking to show the uncertainty associated with different scenarios. As new risks or identification of new data needs arise, imputation maps are ready to adapt to new needs without the necessary production of a new model. However, at the 250-m scale, the variables that are correlated with broad climate patterns, specifically species distributions, will probably be characterized the best. Structure attributes, such as coarse woody debris and quadratic mean diameter, can be mapped, but the mapped variability will likely overwhelm the utility of such products.

Acknowledgments

We would like to thank Jerry Beatty and Terry Shaw for support and comments on this project. We also thank Matt Gregory and Bonnie Rufenacht for technical assistance in conducting these analyses.

Literature Cited

Blackard, J.; Finco, M.; Helmer, E. [and others]. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate-resolution information. Remote Sensing of Environment. 112: 1658-1677.

Borchers, J.G. 2005. Accepting uncertainty, assessing risk: decision quality in managing wildfire, forest resource values, and new technology. Forest Ecology and Management. 211: 36.

Brewer, C.K.; Schwind, B.; Warbington, R. [and others]. 2005. Existing vegetation mapping protocol. In: Brohman, R.; Bryant, L., eds. Existing vegetation classification and mapping technical guide. Gen. Tech. Rep. WO-67. Washington, DC: U.S. Department of Agriculture, Forest Service, Washington Office, Ecosystem Management Coordination Staff. 306 p.

Brohman, R.; Bryant, L., eds. 2005. Existing vegetation classification and mapping technical guide. Washington, DC: U.S. Department of Agriculture, Forest Service. [Not paged].

Daubenmire, R. 1952. Forest vegetation of northern Idaho and adjacent Washington, and its bearing on concepts of vegetation classification. Ecological Monographs. 22: 301–330.

Hassani, B.T.; LeMay, V.; Marshall, P.L. [and others]. 2004. Regeneration imputation models for complex stands of Southeastern British Columbia. Forestry Chronicle. 80: 271–278.

LeMay, V.; Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science. 51: 109–119.

McRoberts, R.E.; Nelson, M.D.; Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sensing of Environment. 82: 457–468.

Mellin, T.C.; Krausmann, W.; Robbie, W. 2004. The USDA Forest Service Southwestern Region mid-scale existing vegetation mapping project. In: Remote sensing for field users: Proceedings of the 10th Forest Service remote sensing conference. American Society of Photogrammetry and Remote Sensing: 1–6.

Moeur, M.; Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science. 41: 337–359.

Ohmann, J.L.; Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere. 32: 725–741.

Pierce, K.B.; Ohmann, J.L.; Wimberly, M.C. [and others]. 2009. Mapping wildland fuels and forest structure for land management: a comparison of nearest neighbor imputation and other methods. Canadian Journal of Forest Research. 39(10): 1901–1916.

Ruefenacht, B.; Finco, M.V.; Nelson, M.D. [and others]. 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. Photogrammetric Engineering and Remote Sensing. 74: 1379-1388.

Stage, A.R.; Alley, J.R. 1972. An inventory design using stand examinations for planning and programming timber management. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest & Range Experiment Station. 32 p.

Temesgen, H.; Gadow, K.V. 2004. Generalized height-diameter models: an application for major tree species in complex stands of interior British Columbia. European Journal of Forest Research. 123: 45–51.

U.S. Department of Agriculture, Forest Service [USDA FS]. 1995. Ecosystem analysis at the watershed scale: Federal guide for watershed analysis. Version 2.2. Portland, OR: USDA Forest Service, Regional Ecosystem Office. http://www.reo.geo/lirary/reports/watershed.pdf. [Date accessed: February 2010].

Wilkie, D.S.; Finn, J.T. 1996. Remote sensing imagery for natural resources monitoring: a guide for first-time users. New York: Columbia University Press. 295 p.

Wimberly, M.C.; Ohmann, J.L.; Pierce, K. [and others]. 2003. A multivariate approach to mapping forest vegetation and fuels using GIS databases, satellite imagery, and forest inventory plots. In: Proceedings of the 2nd International Wildland Fire Ecology and Fire Management Congress. Orlando, FL: American Meteorological Society: http://ams.confex.com/ams/FIRE2003/techprogram/paper_65758.htm.

Continue