Journal of Vegetation Science **22** (2011) 660–676

SPECIAL FEATURE: ECOINFORMATICS

Mapping gradients of community composition with nearest-neighbour imputation: extending plot data for landscape analysis

Janet L. Ohmann, Matthew J. Gregory, Emilie B. Henderson & Heather M. Roberts

Keywords
Canonical correspondence analysis; Constrained ordination; Gradient analysis; Gradient nearest neighbour; *k*NN; Landscape scenario analysis; Oregon; Species distribution modelling

Abbreviations
CCA, canonical correspondence analysis; NN, nearest neighbour; GNN, gradient nearest neighbour; RMSE, root mean square error

Nomenclature
USDA NRCS (2000)

Received 18 June 2010
Accepted 20 November 2010
Co-ordinating Editor: Dr. Robert Peet

Ohmann, J.L. (corresponding author, johmann@fs.fed.us): Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, Oregon, 97331, USA
Gregory, M.J. (matt.gregory@oregonstate.edu) & **Roberts, H.M.** (heather.roberts@oregonstate.edu): Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon, USA
Henderson, E.B. (emilie.henderson@oregonstate.edu): Institute of Natural Resources, Oregon State University, Portland, Oregon, USA

Abstract

Question: How can nearest-neighbour (NN) imputation be used to develop maps of multiple species and plant communities?

Location: Western and central Oregon, USA, but methods are applicable anywhere.

Methods: We demonstrate NN imputation by mapping woody plant communities for $> 100\,000\,km^2$ of diverse forests and woodlands. Species abundances on ~25 000 plots were related to spatial predictors (rasters) describing climate, topography, soil and geographic location using constrained ordination (CCA). Species data from the nearest plot in multi-dimensional CCA space were imputed to each map pixel. Maps of multiple individual species and community types were constructed from the single imputed surface. We computed a variety of diagnostics to characterize different qualities of the imputed (mapped) community data.

Results: Community composition gradients were strongly associated with climate and elevation, and less so with topography and soil. Accuracy of the imputation model for presence/absence of 150 species varied widely (kappa 0.00 to 0.80). Omission error rates were higher than commission rates due to low species prevalence, and areal representation of species was only slightly inflated. A map of 78 community types was 41% correct and 78% fuzzy correct. Errors of omission and commission were balanced, and areal representation of both rare and abundant communities was accurate. Map accuracy may be lower for some species than with other methods, but areal representation of species and communities across large landscapes is preserved. Because imputed vegetation surfaces are developed for all species simultaneously, map units contain suites of species known to co-occur in nature. Maps of individual species, and of community types derived from them, will be internally consistent at map locations.

Conclusions: NN imputation is a useful modelling approach where maps of multiple species and plant communities are needed, such as in natural resource management and conservation planning or models that project landscape change under alternative disturbance or climate scenarios. More research is needed to evaluate other ordination methods for NN imputation of plant communities.

Introduction

Many of today's most challenging issues in natural resource management and conservation planning span broad spatial scales, land ownership and administrative boundaries, as well as long and complex ecological gradients. In addition, landscape-scale issues typically require consideration of multiple interacting threats (e.g. wildfire, invasive species, climate change) and benefits

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

(e.g. wildlife habitat, watershed health, timber supply). Consequently, analysts and decision makers increasingly require basic quantitative and descriptive information about vegetation and land cover over large landscapes that is both highly detailed and spatially complete (i.e. mapped) (Spies et al. 2007).

In this paper we present an approach for developing detailed maps of plant community composition for large landscapes using nearest-neighbour (NN) imputation (see review by Eskelson et al. 2009). While awareness of NN methods is now fairly high in forest inventory circles (Eskelson et al. 2009; McRoberts et al. 2010), these methods are still relatively unknown among vegetation scientists, community ecologists and landscape modellers. We think NN imputation offers several advantages for landscape-level analyses that require information on the spatial distributions of a large number of species, or of plant community types defined by species relative abundances. In this paper we demonstrate how NN methods can be used to map species composition by presenting an analysis for a large region in western Oregon, USA. We focus on the advantages and limitations of NN methods for mapping continuous change in community composition, which has received little attention in the NN literature.

Nearest-neighbour imputation

Nearest-neighbour (NN) imputation is one means of filling in missing data by substituting values from "donor" observations (Eskelson et al. 2009). The value imputed to a location can be a value measured at another (donor) location, or an average value computed from multiple donor locations. In forest inventory, NN methods are used to estimate detailed forest characteristics of large areas at a reasonable cost. They are applied where limited (and less expensive) data (X variables) are available for all observations, and more detailed (and more expensive) data (Y variables) are available only for a sample. The Y variables typically are measures such as tree basal area, density and volume, derived from a sample of field plots or stand exams. The primary X variables often are satellite imagery and their derivatives (Tomppo 1991). The motivation behind NN methods is that two locations with similar Xs should also have similar Ys. Similarity (or distance) between locations (the basis of choosing donor plot(s)) can be evaluated in different ways.

In practice, the distance measure, number of nearest-neighbour plots (k), weighting of the plots in the calculations, choice of X and Y variables, and spatial scale (resolution and extent) of X and Y variables all can be varied to produce different variations of NN mapping. The nearest donor plot(s) to each map unit can be identified using Euclidean distance (kNN, Tomppo 1991), canonical correlation analysis (most similar neighbour, MSN, Moeur & Stage 1995), canonical correspondence analysis (GNN, Ohmann & Gregory 2002), an imputation implementation (Crookston & Finley 2008) of Random Forest (Breiman 2001) and others. When $k = 1$, vegetation attributes measured on the single nearest plot are imputed to each map unit. When $k > 1$, the mean, median, majority or other summary measure across k plots is imputed. When $k = 1$, the covariance structure of vegetation attributes is maintained within each map unit, and the variance structure of the imputations over the study area is similar to the observations. With $k > 1$, local prediction accuracy (RMSE) for individual variables may be stronger (lower RMSE), but covariance structure is not maintained and the range-of-variability of predicted values is reduced (Moeur & Stage 1995; McRoberts et al. 2002; Stage & Crookston 2007; Eskelson et al. 2009).

The most common implementation of NN in forest inventory applications has been kNN using Euclidean distance, $k > 1$, and satellite imagery for the X variables (Tomppo 1991; Franco-Lopez et al. 2001; McRoberts et al. 2010). MSN with $k = 1$ has been widely applied by federal land managers in the USA based on partial stand exam data (Moeur & Stage 1995), and GNN with $k = 1$ has been applied across much of the Pacific Northwest USA for purposes of landscape analysis and monitoring (Spies et al. 2007; Moeur et al. 2009). Virtually all forestry applications of NN have emphasized forest structure, with little attention given to species composition beyond general forest types (but see Ohmann & Gregory 2002; Ohmann et al. 2007; Hudak et al. 2008).

Error in NN predictions arises from measurement error, error inherent in the particular imputation method and pure error (variation in the Y variables not associated with available X variables that is associated with the underlying true but unknown model) (Stage & Crookston 2007). Choice of X and Y variables, distance measure and k all contribute to error, but no single choice gives best results for all applications, nor for all response variables within a given application, and models must be developed on a case-by-case basis (Eskelson et al. 2009). Very few studies have compared alternative NN methods specifically for mapping species composition. Furthermore, although diagnostic tools for evaluating NN models are becoming increasingly available (Crookston & Finley 2008; McRoberts 2009), few are specific to assessing distributions of species or community composition.

Relationships to other community-level modelling approaches

Although a large body of literature exists on modelling the distributions of individual species (species distribution

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

models, or SDMs), much less research has been devoted to mapping plant community composition (Franklin 2009). Community-level modelling includes predictive mapping of community types, species groups, axes or gradients of compositional variation and other ecological properties (Ferrier & Guisan 2006). Ferrier & Guisan (2006) described three strategies for modelling at the community level, each with advantages and limitations for different applications: (i) "assemble first, predict later," where data are first classified into community types that are then spatially modelled; (ii) "predict first, assemble later," in which species are modelled and mapped individually and then the stacks of species maps are classified or summarized; and (iii) "assemble and predict together." To maximize utility for a variety of landscape analyses that require spatial information on individual species, community types, or both, we sought a method that retains information on individual species identities in the final predictions. This ruled out strategy (i), as well as several methods within strategy (iii). Furthermore, "stacks" of species maps produced with strategy (ii), as well as any community types derived from them, may yield unrealistic combinations for a given map unit.

NN imputation, which falls in strategy (iii), is one of the few methods with the potential capacity to model all species simultaneously, to retain individual species identities in the spatial predictions, and (when $k=1$) to produce map units containing assemblages of species known to co-occur in nature. Thessler et al. (2005) used *k*NN to map ordination scores, but not species or community composition. Thessler et al. (2008) used *k*NN to distinguish floristically and structurally different forest types, but did not map individual species. Others have used ordination and gradient analysis in spatial prediction, but not within a NN imputation framework. For example, Guisan et al. (1999) used CCA to map individual species, and Schmidtlein et al. (2007) depicted continuous variation in community composition using ordination scores. We know of only a handful of studies that have used ordination within a NN framework for mapping individual species and communities (Gottfried et al. 1998; Ohmann & Gregory 2002; Dirnböck et al. 2003; Ohmann et al. 2007; Hudak et al. 2008).

A landscape-level approach for mapping multiple species, communities and gradients

We suggest that NN imputation is a modelling approach worth considering for cases where information is needed on distributions of a large number of species, or of plant community types defined by species presence or relative abundance. NN maps are particularly useful where multiple analytical needs are to be met by a single map, such as within a landscape modelling framework, where interactions among vegetation components are important. Although much of our past work has emphasized forest structure (Ohmann & Gregory 2002; Pierce et al. 2009), in this paper we focus on how NN imputation can be used to map gradients in species composition. We present an analysis of gradients in vegetation composition over a large forested landscape in western and central Oregon, USA (Fig. 1), and use NN imputation to develop maps of multiple species and forest community types. Resulting maps can be used for many applications in forest management, conservation planning and research. We think our example of the NN imputation approach is unique in its application over large geographic areas at a relatively fine spatial resolution needed to support landscape analysis for natural resource planning and policy decisions (Spies et al. 2007; Moeur et al. 2009). Because our objective is to demonstrate NN imputation as a general approach, rather than to compare alternative methods, we present only one NN method (GNN), which is based on constrained ordination (direct gradient analysis) (ter Braak 1986).

Methods

Study area

Our study area is 109 279 km^2 in western and central Oregon, USA (Fig. 1). Our models apply only to the 81% of the region that is forest or woodland ($\geq$10% tree cover). Western Oregon has a maritime climate with mild, wet winters, cool dry summers and heavy precipitation. Precipitation increases and temperature decreases from

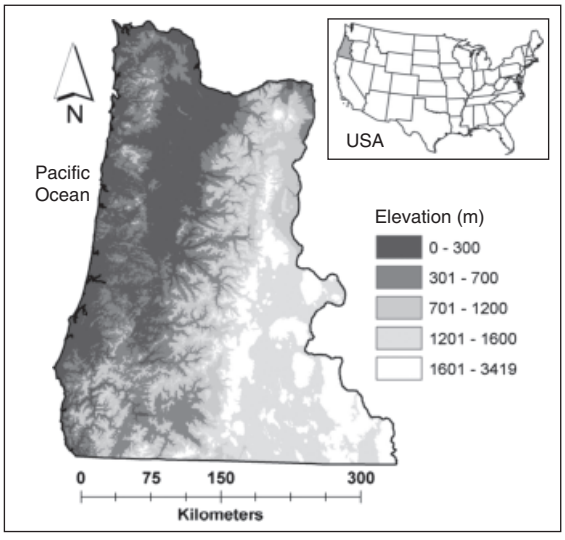

Fig. 1. Study area in western and central Oregon, USA, showing the elevation gradient.

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

south to north. East of the crest of the Cascade Mountains in central Oregon, temperatures fluctuate more widely and are more extreme, frost-free seasons are shorter and precipitation is much less. Elevations range from sea level to > 3000 m. Soil types are primarily inceptisols, spodosols and ultisols. Volcanic activity during the Pleistocene and Holocene has mantled large tracts at higher elevations in the Cascade Range and in central Oregon with pumice and ash. Ultramafic parent materials strongly influence community composition in parts of southwest Oregon.

Coniferous tree species dominate forest communities. Outside of the mixed-evergreen zone of southwest Oregon, where several evergreen hardwood trees co-dominate, broadleaf trees tend to occupy harsh sites or riparian habitats, or serve as pioneers. The forest zones of interior southwest Oregon represent northern extensions of the mixed-conifer forest of the Sierra Nevada and the mixed sclerophyll forest of the California Coast Ranges. In central Oregon, Pacific coastal elements mix with Rocky Mountain elements.

Fire is the predominant natural disturbance. Natural fire-return intervals ranged from 15 years in drier eastside pine forests, to 400 years in moist coastal forests, to 800 years in subalpine forests (Agee 1993). In the last 100 years, natural disturbance regimes have been supplanted by timber management and wildfire suppression. See Ohmann & Spies (1998) and Franklin & Dyrness (1973) for more detailed descriptions of the region.

Plot data

The *Y* variables (response variables) were abundances of woody species (trees and shrubs) measured on a sample of 24 937 field plots from regional forest inventories (Forest Inventory and Analysis (FIA), USFS; Current Vegetation Survey (CVS), USFS and BLM); the Region 6 Ecology Program, USFS; and the interagency Josephine-Jackson County Fuel Mapping Project. The FIA and CVS plots were installed on systematic grids. We used only plots on forest land, defined as ≥10% tree cover or being assigned to a forest or woodland plant association. Plot sizes ranged from 500 m^2 (Ecology plots) to a cluster of subplots distributed over about 1 ha (inventory plots). Cover of shrub and herbaceous species was visually estimated to the nearest 5% on all plots. Tree species cover was visually estimated to the nearest 5% on Ecology plots, and computed from other tallied attributes (tree diameter, height and live crown ratio) on the inventory plots. We excluded extremely rare (< 20 occurrences) and non-native species from model development. Herbaceous species were excluded because of concerns about inconsistent sampling protocols and species identification across the plots. Vegetation data for the FIA and CVS plots are available from FIA (http://www.fs.fed.us/pnw/fia/), and the Ecology Program plots are available from http://ecoshare.info/.

Explanatory variables (spatial predictors)

The *X* variables (explanatory variables or spatial predictors) were rasters representing climate, topography, soil parent material and geographic location (Table 1), resampled to 30-m resolution for modelling. Climate variables were derived from PRISM data (Daly et al. 2008), which were 1971–2000 normals with native spatial resolution of 800 m. We did not use satellite imagery in the model. Previous experience showed that including Landsat spectral data, which is correlated primarily with local-scale variation in forest structure, reduces prediction accuracy for species composition in models spanning large regions (Ohmann & Gregory 2002; Ohmann et al. 2007). In addition, the plots were measured over a span of many years, which precludes use of imagery from a single date. For these reasons, disturbance and successional status are not taken into account, and the models should be considered an approximation of potential distributions of species and plant communities.

Constrained ordination (direct gradient analysis) with CCA

Because our intent was to demonstrate NN imputation as a general approach for spatial modelling of community composition rather than to compare NN methods, we limited our analysis to a single method: GNN (based on CCA) with $k = 1$. Although previous research (Hudak et al. 2008; Grossmann et al. 2010, unpubl. report, http://www.fsl.orst.edu/lemma/pubs) suggests that Random Forest NN slightly outperforms GNN and other NN methods for mapping tree species composition, implementation of Random Forest NN for large rasters and large numbers of plots is computationally prohibitive with current software and hardware. Both studies demonstrated that several NN methods (including GNN) performed similarly well.

In developing the GNN variation of NN imputation (Ohmann & Gregory 2002), we looked to ordination for a distance measure that would allow analysis of multispecies plant communities and also facilitate interpretation of ecological gradients. Despite known limitations, we chose CCA because it is a constrained ordination method that allows prediction, based on a linear combination of explanatory variables. CCA also accommodates sparse data matrices and nonlinear responses of species along environmental gradients. Although the validity of the Gaussian model of species distributions that underlies

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

Table 1. Spatial predictors (explanatory variables) used in CCA and GNN imputation.

Variable subset	Code	Description
Climate	ANNPRE	Mean annual precipitation (natural logarithm, mm)
	ANNTMP	Mean annual temperature (°C)
	AUGMAXT	Mean maximum temperature of the hottest month (August) (°C)
	CONTPRE	Percentage of annual precipitation falling during the growing season (June–August)
	DECMINT	Mean minimum temperature of the coldest month (December) (°C).
	SMRTP	Growing season moisture stress, the ratio of mean temperature (°C) to precipitation (natural logarithm, mm), May–September
	STRATUS	Percentage of hours in July with cloud ceiling of marine stratus < 1524 m and visibility < 8 km. (unpubl. data from Chris Daly, resolution 795 m)
Topography	ELEV	Elevation (m)
	ASP	Cosine transformation of aspect (degrees)
	SOLAR	Cumulative potential relative radiation during the growing season (Pierce et al. 2005)
	SLOPE	Slope (%)
	TPI	Topographic position index, calculated as the difference between a cell's elevation and the mean elevation of cells within a 450-m radius window
Parent material	ASH	Total depth of ash deposition (feet), primarily from Mt. Mazama in the eastern Cascades (unpubl. data from Mike Simpson)
	ALLUV	Unconsolidated material deposited by rivers (categorical)
	SILICIC	Contains rocks with minerals high in silica (categorical)
	ULMA	Rocks with ultramafic minerals including serpentine (categorical)
	SAND	Sandy sediments (categorical)
Location	UTME	Universal Transverse Mercator easting (m)
	UTMN	Universal Transverse Mercator northing (m)

correspondence analysis and its derivatives remains under debate (Austin 2002), violations (such as skewed species distributions) are not thought to cause problems for CCA (McCune & Grace 2002). Furthermore, within an imputation framework the critical feature affecting prediction is the relative positions of plots in ordination space. Resulting maps should be satisfactory if plots with similar species composition are closer together in ordination space and dissimilar plots are further apart.

Another potential concern with CCA is its use of multiple regressions of community gradients on environment variables, which limits the kinds of gradient that can be captured by the ordination. However, a previous analysis of similar data showed that axis scores from detrended correspondence analysis (DCA), an unconstrained ordination method, and detrended CCA were strongly correlated, suggesting that much of the variation in vegetation was related to the measured explanatory variables (Ohmann & Spies 1998). Furthermore, CCA is robust to multicollinearity that can cause problems in other regression applications (Palmer 1993).

We used the forward step-wise procedure of CCA in CANOCO version 4.5 (Microcomputer Power, Ithaca, NY, US) to identify a reduced set of explanatory variables. We added explanatory variables to the model in the order of greatest additional contribution to explained variation. Variables were added if significant ($P < 0.01$), where significance was determined by a Monte Carlo permutation test using 99 permutations (H0: additional influence of variable on vegetation is not significantly different from random), and if adding the variable did not cause any variance inflation factors to exceed 20. We excluded UTME and UTMN from the step-wise procedure because they are strongly correlated with several of the climate variables and do not directly influence species distributions. However, we added UTME and UTMN to the final CCA model used in the imputation, to encourage selection of NN plots that are closer in geographic space as well as in gradient space. Response variables in CCA were cover (percentage) by species, square-root transformed to dampen the influence of dominant species, and we did not down-weight rare species.

Nearest-neighbour imputation

To implement GNN, we used the R package vegan (*Vegan: community ecology package*. Version 1.8-8. http://cran.r-project.org/, http://vegan.r-forge.r-project.org/.) to run the final CCA model, a C++ program (GNNRun) with fast neighbour-finding (Finley & McRoberts 2008) to conduct the imputation, and Python and R scripts to compute a suite of model diagnostics. The GNNRun program is available on request. Alternatively, NN methods (including GNN) can be implemented using the R package yaImpute (Crookston & Finley 2008), and some diagnostics can be computed using the R package nnDiag (k-nearest neighbor diagnostic tools. Version 0.0-5. http://cran.r-project.org/, http://blue.for.msu.edu/NAFIS/software.html.).

Neighbour-finding was based on Euclidean distance within CCA ordination space defined by the first eight CCA axes, with axes scores weighted by their eigenvalues, and using scores that are linear combinations of the explanatory variables. For each map pixel, distance was calculated to each of the 24 937 plots, and the nearest plot was identified and associated with the pixel.

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

The model output is in ArcGIS grid format, where each pixel is assigned a unique plot identifier and summary attributes from the plot are joined from a database table.

Developing maps of species distributions and community types from the imputation model

Every pixel in the map was imputed with data from a single field plot, which allowed us to generate a distribution map for any species observed on at least one of the plots. In addition, as an example of how imputed maps can be used in mapping plant community types, we classified all plots into Ecological Systems (Comer et al. 2003). Ecological Systems are related to the hierarchical US National Vegetation Classification System, but not nested within it. They are a conceptual aggregation of plant associations that generate groupings more floristically detailed than Formation, but less than Alliance and Association, and were defined partly with the intent of creating "mappable" units. Ecological Systems emphasize existing dominant vegetation types, and are defined as groups of plant community types that tend to co-occur within landscapes with similar ecological processes, substrates and environmental gradients (Comer et al. 2003). The Ecological Systems classification has been adopted by the Gap Analysis Program for vegetation mapping and conservation planning at the regional to national level (Kagan et al. 2008; Grossmann et al. 2010). We built keys to classify plots into Ecological Systems based on plant community composition, primarily cover of individual species, with the list of possible Systems constrained to those that occur within the physiographic provinces of our study area. We developed the logic behind our keys from the NatureServe descriptions of the Ecological Systems (http://www.natureserve.org/getData/USecologyData.jsp), indicator species lists and classification guidelines provided by NatureServe and the LANDFIRE program (http://www.landfire.gov), and expert opinion of ecologists in our region.

Evaluation of the CCA ordination and the imputed maps

To evaluate how well the constrained ordination represented among-plot distances in the original data set, we computed after-the-fact correlations between ordination distances and distances in the unreduced species space using relative Euclidean distance (McCune and Grace 2002), implemented in PC-ORD version 5 (MjM Software, Gleneden Beach, OR, US).

To evaluate the imputed map, we computed a variety of diagnostics that address different aspects of model error and uncertainty for various measures of species composition. The multiple model diagnostics provide different characterizations of a single imputation model. They allow users to evaluate the maps for their particular application, and provide a basis for comparison with other studies. Note that mapped species distributions contained in the imputed spatial predictions are not expressed as probabilities, where a threshold value (e.g. 0.50) can be chosen to classify a species as present. Therefore, the effects of varying the threshold probability on model accuracy measures such as sensitivity and specificity (Franklin 2009) could not be evaluated, and metrics such as area under the curve (AUC) of the receiver operating characteristic (ROC) (Hanley & McNeil 1982) could not be computed.

Several diagnostics were computed at the local- (plot-) scale based on a modified leave-one-out cross-validation, described in Ohmann & Gregory (2002) as the second-nearest-neighbour procedure. For the locations of the 24 937 plots in the model, we compared observed to predicted values from the second-nearest-neighbour plot (the NN would be the plot itself). Although technically not the same as a true leave-one-out analysis, we have observed that individual plots do not have a measurable effect on CCAs for our very large sample sizes. From the cross-validation we computed RMSEs and correlations for continuous variables such as species richness; kappa statistics (Cohen 1960) for species presence/absence; and kappas and confusion matrices for plant community types (Ecological Systems). In assessing model accuracy for Ecological Systems, we also applied fuzzy set methods (Gopal & Woodcock 1994; Congalton & Green 1999), which recognize that thematic mapping involves placing a continuum of vegetation change into discrete classes and that there can be different magnitudes of error among classes. We defined our fuzzy sets based on similarity among the Systems in several dimensions: seral relationships, geographic proximity (Systems that tend to occur near one another were considered more similar); and similarities in moisture regimes, elevational limits, species composition, structure and soil types. Ecological System pairs that were similar in multiple dimensions were designated as "fuzzy correct" for fuzzy accuracy assessment (Grossmann et al. 2010). For example, the NP Dry-Mesic DF-WH Forest was classified as "fuzzy similar" to the NP Mesic-Wet DF-WH Forest.

We computed several diagnostics of the imputed map for various measures of species composition using R scripts that are "wrappers" for R functions from published R packages. The scripts compare observed and predicted (imputed) values. We calculated several measures of species diversity based on the "diversity" function in the R package vegan (*Vegan: community ecology package*. Version 1.8-8. http://cran.r-project.org/, http://vegan.r-forge.r-project.org/). We calculated the distance within

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

species multivariate space between observed and imputed community composition for each plot location using distance metrics in the R packages vegan and labdsv (Version 1.4-1. http://cran.r-project.org/, http://ecology.msu.montana.edu/labdsv/R.). We present Bray-Curtis and the binomial metric to illustrate complementary dimensions of plant community composition. The Bray-Curtis metric tends to place plots close together when they contain the same dominant species, whereas the binomial metric places plots far apart when their species lists differ, even with respect to minor species. We calculated the median percentage improvement in terms of these distance measures as the difference between the observed–predicted distance versus the median distance from each observed plot to all other observed plots. We also assessed species lists for observed and imputed communities for errors of commission (false positives, species predicted to occur at a given plot location that were absent in the observed data) and omission (false negatives, species present in the original data that are absent in the imputed prediction for that location). The R scripts, which are available on request, reference the yaImpute object and so can be used with the R package yaImpute (Crookston & Finley 2008).

At the regional scale, we compared area distributions for species and communities predicted from our model to sample- (probability-) based estimates from the FIA plots, as a way of evaluating areal representation across a broader area. Model uncertainty also was depicted spatially using a map of NN distance.

Results

Primary gradients in species composition and environment

A total of 158 woody species, 39 trees and 119 shrubs, were recorded on the 24 937 plots. Median richness (alpha diversity) on the plots was 10 species, the same as the median alpha diversity for all pixels in the imputed map. Species turnover across plots in our large region was high. Prevalence (frequency of species occurrence on plots) was low for most species (median 1.5%), with trees generally more prevalent (median 4.2%) than shrubs (median 0.9%). In total, 54% of tree species and 25% of shrub species were present on at least 5% of the plots (Table 2). Only one tree (*Pseudotsuga menziesii*) and one shrub (*Mahonia nervosa*) were present on $> 50\%$ of the plots.

Nearest-neighbour distance, an indicator (rather than a direct measure) of model uncertainty and plot support for the imputation, is shown in Fig. 2. Each pixel value represents the distance from that pixel to the NN plot imputed by GNN, where distance is Euclidean, in eight-dimensional gradient space with axes weighted by their eigenvalues.

The first four CCA axes were readily interpretable and were most strongly associated with broad-scale climate (Fig. 3). Eigenvalues were 0.571 (axis 1), 0.401 (axis 2), 0.264 (axis 3) and 0.171 (axis 4), with subsequent axes < 0.10. Although we used eight axes in the imputation, later axes have very little influence on neighbour selection because distances are weighted by the eigenvalues. After-the-fact correlations between ordination distances and distances in the unreduced species space indicated that axis 1 explained the most variation in the species data ($r^2 = 0.338$), followed by axis 2 ($r^2 = 0.081$) and axis 3 ($r^2 = 0.012$). After-the-fact correlations were very similar for CCA models with and without UTME and UTMN. Axis 1 was positively associated with elevation, and with a gradient of maritime climate along the coast (low axis scores) to more continental conditions east of the Cascades (high axis scores). The direction of change was nearly longitudinal (correlated with UTME) (Figs 3 and 4). Axis 1 also reflected the steeper and more dissected topography in western Oregon (Figs 3 and 4). Axis 2 was predominantly a gradient in moisture, especially during the summer growing season, and was structured latitudinally (correlated with UTMN) (Figs 3 and 4). Axis 3 reflected a latitudinal gradient in the seasonality of precipitation (Fig. 4), and axis 4 (not shown) indicated frequency of summer stratus conditions. Several soil parent materials were significant in the ordination but were not important until later axes. Only a few parent materials explained significant amounts of variation and their effects were limited to certain regions: sand near the Coast, ultramafic in southwest Oregon and ash depth in central Oregon.

Plots (not shown) and species were arrayed on the first two CCA axes in readily interpretable ways (Fig. 3). Species associated with drier forests east of the Cascades were concentrated in the upper-right quadrant, species of high-elevation forests along the crest of the Cascades in the lower-right, coastal species associated with maritime climate in the lower-left, and species found in mixed-evergreen forests and *Quercus* woodlands of southwest Oregon and the interior valleys of western Oregon in the upper-left.

Imputed maps of individual species

Although all of the 158 species can be mapped from the single GNN imputation model, for brevity we show predicted presence for five tree species (Fig. 5). The species represent a variety of habitat associations that span the dominant gradients captured by CCA axes 1 and 2 (species are circled in Fig. 3). *Juniperus occidentalis* is the

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

Table 2. Diagnostics for predicted species presence/absence from GNN imputation, listed by descending prevalence. All but areal representations are calculated from cross-validation using terminology and methods from Franklin (2009). Only those shrub species present on ≥5% of model plots are shown. Areal representation is area where species is predicted present in the model compared to area present in the systematic sample of 2853 plots. Species with positive values are over-predicted and negative values are under-predicted.

Species group	Code	Species	Prevalence	From cross-validation			Areal representation (%)
				False negative rate	False positive rate	Kappa	
Trees	PSME	*Pseudotsuga menziesii*	0.759	0.050	0.150	0.799	4.8
	TSHE	*Tsuga heterophylla*	0.373	0.203	0.129	0.667	7.8
	ABGRC	*Abies grandis* and *A. concolor*	0.322	0.279	0.128	0.594	6.6
	CHCH7	*Chrysolepis chrysophylla*	0.238	0.497	0.161	0.340	3.6
	PIPO	*Pinus ponderosa*	0.226	0.264	0.077	0.659	2.0
	CADE27	*Calocedrus decurrens*	0.194	0.393	0.102	0.500	4.4
	ARME	*Arbutus menziesii*	0.169	0.349	0.072	0.577	3.1
	ACMA3	*Arctostaphylos manzanita*	0.165	0.501	0.098	0.401	5.9
	THPL	*Thuja plicata*	0.155	0.461	0.086	0.451	4.0
	PICO	*Pinus contorta*	0.135	0.244	0.041	0.710	−0.6
	PILA	*Pinus lambertiana*	0.120	0.518	0.071	0.409	3.4
	ABPRSH	*Abies procera, A. shastensis* and *A. magnifica*	0.114	0.322	0.042	0.636	1.1
	TABR2	*Taxus brevifolia*	0.114	0.612	0.080	0.307	4.6
	CONU4	*Cornus nuttallii*	0.112	0.576	0.074	0.349	6.5
	ABAM	*Abies amabilis*	0.106	0.279	0.035	0.682	1.6
	TSME	*Tsuga mertensiana*	0.102	0.239	0.028	0.733	0.9
	ALRU2	*Alnus rubra*	0.095	0.615	0.058	0.336	−1.5
	PIMO3	*Pinus monticola*	0.092	0.509	0.053	0.436	1.6
	LIDE3	*Lithocarpus densiflorus*	0.090	0.221	0.022	0.757	2.4
	QUCH2	*Quercus chrysolepis*	0.079	0.417	0.034	0.554	2.0
	QUKE	*Quercus kelloggii*	0.052	0.421	0.022	0.565	1.7
	UMCA	*Umbellularia californica*	0.032	0.476	0.017	0.502	1.2
	CHLA	*Chamaecyparis lawsoniana*	0.031	0.474	0.016	0.502	1.1
	QUGA4	*Quercus garryana*	0.031	0.541	0.016	0.449	0.3
	PRUNU	*Prunus* spp.	0.030	0.812	0.024	0.171	1.8
	ABLA	*Abies lasiocarpa*	0.023	0.537	0.012	0.463	0.2
	JUOC	*Juniperus occidentalis*	0.022	0.449	0.010	0.540	1.4
	PISI	*Picea sitchensis*	0.022	0.299	0.009	0.664	0.3
	PIEN	*Picea engelmannii*	0.017	0.700	0.012	0.288	0.2
	CELE3	*Cercocarpus ledifolius*	0.013	0.549	0.007	0.447	1.3
	LAOC	*Larix occidentalis*	0.011	0.636	0.006	0.379	0.5
	PIAL	*Pinus albicaulis*	0.010	0.667	0.005	0.362	0.1
	PIJE	*Pinus jeffreyi*	0.008	0.521	0.006	0.440	0.3
	PIAT	*Pinus attenuata*	0.005	0.797	0.004	0.194	0.1
	FRLA	*Fraxinus latifolia*	0.005	0.894	0.004	0.111	0.0
	POTR5	*Populus tremuloides*	0.005	0.831	0.004	0.161	0.3
	CHNO	*Chamaecyparis nootkatensis*	0.003	0.704	0.002	0.296	0.0
	POBAT	*Populus balsamifera* ssp. *trichocarpa*	0.003	0.937	0.002	0.071	−0.2
	PIBR	*Picea breweriana*	0.001	0.677	0.001	0.372	0.1
	ALRH2	*Alnus rhombifolia*	0.001	1.000	0.001	0.000	−0.2
	SESE3	*Sequoia sempervirens*	0.001	0.714	0.001	0.266	0.0
Shrubs	MANE2	*Mahonia nervosa*	0.511	0.201	0.199	0.600	1.2
	RUUR	*Rubus ursinus*	0.408	0.354	0.239	0.407	3.7
	GASH	*Gaultheria shallon*	0.323	0.258	0.121	0.622	1.8
	VAPA	*Vaccinium parvifolium*	0.314	0.285	0.134	0.580	7.0
	CHUM	*Chimaphila umbellata*	0.306	0.343	0.152	0.505	2.2
	ACCI	*Acer circinatum*	0.301	0.315	0.126	0.562	−0.2
	HODI	*Holodiscus discolor*	0.235	0.460	0.135	0.409	1.6
	COCOC	*Corylus cornuta* var. *californica*	0.233	0.418	0.123	0.462	2.7

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

Table 2. Continued

Species group	Code	Species	Prevalence	From cross-validation			Areal representation (%)
				False negative rate	False positive rate	Kappa	
	RHMA3	*Rhododendron macrophyllum*	0.204	0.371	0.096	0.533	0.1
	CHME	*Chimaphila menziesii*	0.153	0.659	0.121	0.219	7.5
	VAME	*Vaccinium membranaceum*	0.145	0.358	0.063	0.576	1.0
	TODI	*Toxicodendron diversilobum*	0.138	0.370	0.059	0.571	1.0
	LOHI2	*Lonicera hispidula*	0.130	0.434	0.067	0.497	6.2
	AMAL2	*Amelanchier alnifolia*	0.121	0.590	0.079	0.333	2.3
	RUPA	*Rubus parviflorus*	0.120	0.742	0.092	0.170	0.6
	PAMY	*Paxistima myrsinites*	0.119	0.494	0.063	0.448	0.4
	VAOV2	*Vaccinium ovatum*	0.106	0.315	0.037	0.648	0.4
	ARPA6	*Arctostaphylos patula*	0.099	0.466	0.047	0.496	− 2.6
	PUTR2	*Purshia tridentata*	0.097	0.178	0.021	0.796	− 1.1
	MAAQ2	*Mahonia aquifolium*	0.095	0.588	0.061	0.353	3.3
	SYAL	*Symphoricarpos albus*	0.090	0.587	0.053	0.369	− 2.2
	RUSP	*Rubus spectabilis*	0.084	0.452	0.039	0.514	− 0.9
	RULA2	*Rubus lasiococcus*	0.079	0.537	0.048	0.412	3.2
	CEVE	*Ceanothus velutinus*	0.078	0.583	0.047	0.375	− 2.7
	FRPU7	*Frangula purshiana*	0.076	0.656	0.053	0.292	4.6
	ARNE	*Arctostaphylos nevadensis*	0.061	0.588	0.037	0.378	− 1.7
	RICE	*Ribes cereum*	0.055	0.566	0.032	0.408	− 1.2
	VAOV	*Vaccinium ovalifolium*	0.053	0.535	0.032	0.425	1.5

dominant tree species in the dry woodlands at the easternmost fringe of the study area. *Tsuga mertensiana* dominates the highest forested zone along the western slopes of the Cascades. *Picea sitchensis* characterizes forest in a narrow band along the coast. *Lithocarpus densiflorus* is a dominant evergreen broadleaf tree characteristic of the mixed evergreen forests of southwest Oregon. *Quercus garryana*, a deciduous broadleaf tree, is common in *Quercus* woodlands of the interior valleys of western Oregon and east of the Cascades in the northeastern portion of our study area.

Based on cross-validation, alpha diversity at the plot locations averaged 10 species for both observed (plots) and predicted (mapped). However, on average the predictions at plot locations included four species omission errors and four commission errors. Overall, when expressed as a rate, errors of omission (false negatives) were greater than errors of commission (false positives), because of the very low species prevalence (Table 2). The error of omission rate was 66% for all species (53% for trees and 71% for shrubs). The error of commission rate was 3% for all species (4% for trees and 3% for shrubs).

The average kappa for all 158 species was 0.31, but values ranged widely, from 0.00 to 0.80 (Table 2). Kappas generally were better for more prevalent species: the average kappa for 51 species present on at least 5% of plots was 0.49. Model performance also was better for those species having distributions with clearly defined limits in elevation or climate. For example, *Picea sitchensis* is confined to the zone of summer fog along the Pacific Coast and is present on only 2% of the plots, but had a kappa of 0.66.

In terms of areal representation, the model slightly over-predicted the distribution of individual species on the landscape when compared to estimates from a subsample of 2853 FIA plots established on a systematic grid (Table 2). On average, individual species were predicted to be present on 1.5% more of the total forest area than expected based on the sample-based estimates. Tree species were slightly more over-predicted on average (1.7%) than were shrubs (1.4%). However, the magnitudes of difference were greater for shrubs than for trees (Table 2).

Imputed maps of community composition and types (Ecological Systems)

The median Bray-Curtis distance between observed and imputed community composition for the plots was 0.48, and the median binomial distance was 51.6. The median improvement in the model over the median distance to all other plots was 59% for Bray-Curtis and 47% for binomial.

Fifty forest Ecological Systems were sampled by forest plots in the study area, and an additional 28 non-forest Systems were sampled by very small numbers of plots having very low tree cover (data not shown). The imputed Systems for part of the western Cascades is

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

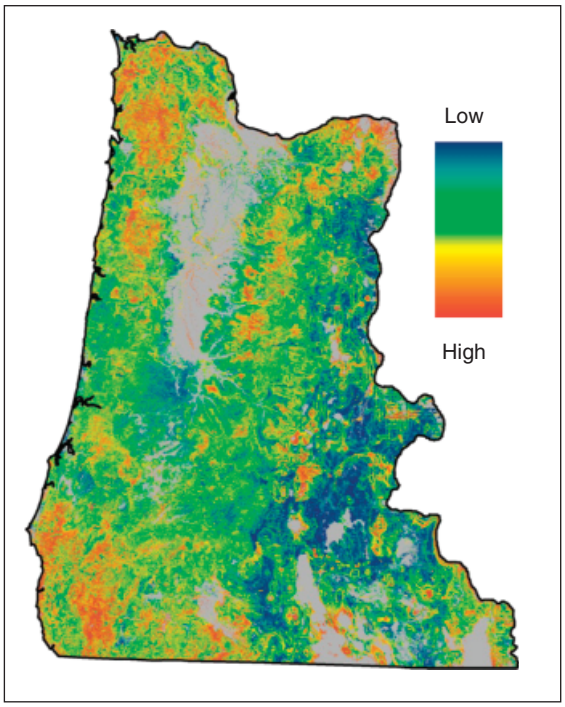

Fig. 2. Nearest-neighbour distance, an indicator of model uncertainty and plot support. Distance is calculated for each map pixel from values for the spatial predictors. Distance is Euclidean, in eight-dimensional CCA space with axes weighted by their eigenvalues. Grey areas are nonforest.

shown in Fig. 6. Overall prediction accuracy from cross-validation for all 78 Ecological Systems was 41%, and fuzzy accuracy was 78%. For the forest Systems that occupied at least 1% of the forest area in the model, overall and fuzzy accuracies were 35% and 71%, and average kappa and fuzzy kappa were 0.32 and 0.76 (Table 3). In terms of areal representation, the differences between model- and sample-based estimates of forest area in each Ecological System was slight ($\leq$ 2% difference for all of the most abundant Systems) (Table 3).

Discussion

Evaluating the ordination within the context of NN imputation

Total variation explained (TVE) in the CCA, calculated as the sum of all constrained eigenvalues divided by the total variation (inertia), was 10.2% – at the low end of the 10–50% range typically reported for CCA (Palmer 1993). Although low TVEs often are attributed to unmeasured explanatory variables, Økland (1999) demonstrated that eigenvalues from polynomial distortion axes can contribute 30–70% of total inertia. The percentage is larger for data sets with high beta diversity, as in this study and as reported previously for a similar data set (Ohmann & Spies 1998). Furthermore, TVEs cannot be compared among studies based on different data sets or ordination methods.

McCune & Grace (2002) recommend the after-the-fact correlation of among-plot distances in the constrained ordination space with distances in the original species space as a more appropriate measure of explained variation. This measure of model fit has intuitive appeal for NN imputation, where the critical feature affecting outcome (nearest neighbour selection) is how well the constrained ordination retains the relative positions of plots as defined by their species relative abundances. The resulting imputation (map) should be satisfactory if plots with similar species composition are closer together in ordination space and dissimilar plots are farther apart. The Bray-Curtis and binomial distances calculated between predicted (imputed) and observed composition provide a measure of this aspect of model fit.

By presenting only one variation of NN imputation, with a distance measure based on CCA, we do not mean to advocate CCA over potential alternatives. We encourage additional research on other distance measures in NN imputation where the objective is to map multi-species communities. Other ordination methods, such as non-metric multidimensional scaling (NMS; Kruskal & Wish 1978) (with an additional step required for prediction) or multidimensional fuzzy set ordination (Roberts 2009), may prove superior for ecological explanation and avoid the pitfalls of CCA. NN imputation based on Random Forest (Crookston & Finley 2008) appears promising for species mapping (Hudak et al. 2008), but further development is needed to make the algorithm computationally feasible for large sample sizes. In addition, Random Forest results are less interpretable ecologically and imputation "distance" is less intuitive.

Limitations of the plot data for regional vegetation modelling and mapping

Combining large plot data sets for vegetation modelling and mapping across large regions presents many challenges, a discussion of which is beyond the scope of this paper. However, it is worth mentioning the reliability of plant species identification on the plots. We purposefully limited our analysis to trees and shrubs, which generally are more reliably identified in the field than herbaceous species. Nevertheless, our plot data undoubtedly contain species omissions, misidentifications and other errors. In NN imputation with $k = 1$, these errors are transferred directly to the map rather than "averaged out," and map users need to be alert for anomalies.

Disturbance and successional status were not taken into account in the GNN model. The plots were measured

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

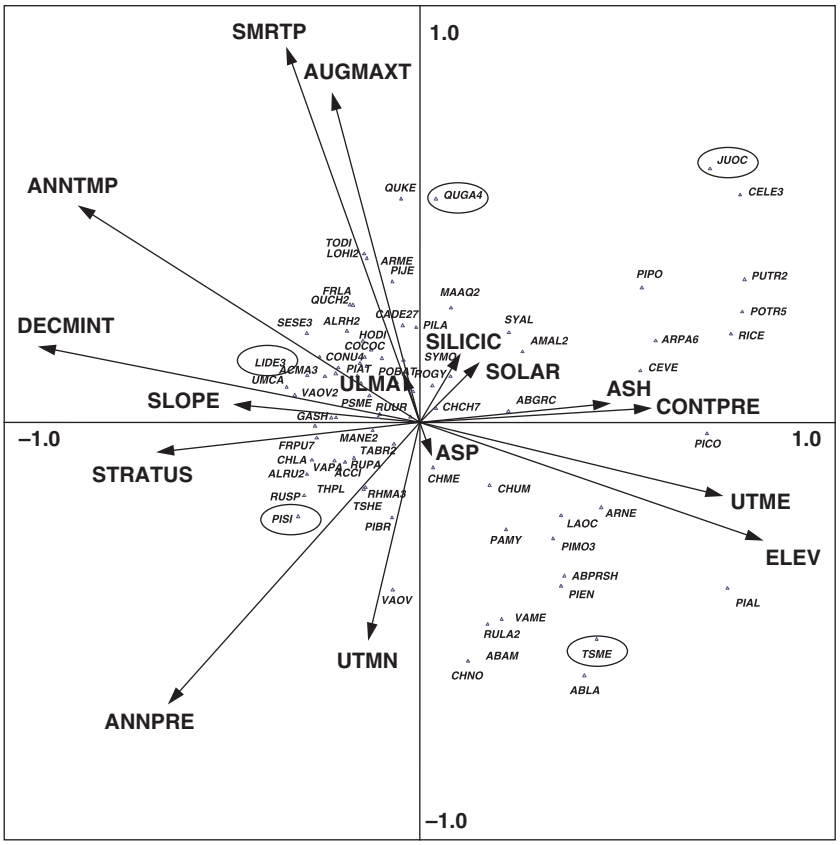

Fig. 3. Biplot for CCA axes 1 and 2, showing explanatory variables (TPI, ALLUVIAL and SAND not shown) and species centroids (all trees, and shrubs present on at least 5% of plots). Species scores are linear combinations of plot scores. See Table 2 for species codes. Spatial predictions for five circled species are shown in Fig. 5.

over a range of many years, and established in forests with a range of ages and disturbance histories. The models should be considered an approximation of potential distributions of species and plant communities. Even on heavily disturbed sites in the region, few woody species are totally eliminated, and ordinations of long gradients are influenced more by species presence than by abundance (Ohmann & Spies 1998).

The map of Ecological Systems is just one example of how species abundance data can be used to classify the plots, followed by mapping the classification across the landscape based on the imputed plot locations. Other classifications can be applied and mapped without having to develop a new model. Ecological Systems presented several challenges for modelling. The classification system was defined without specific experience of what can be reliably modelled with typical plot and spatial data. In the classification key, many of the Systems are distinguished by minor shifts in abundance among one or a few tree or shrub species. The large number of classes, and the similarity among several of the Systems, results in a fairly low per cent correct, although many of the confusion errors are minor, as reflected in the higher fuzzy accuracies (Table 3).

Strengths and limitations of imputed maps of multiple species and plant communities

Statistical properties and sampling considerations for NN imputation

The NN imputation methods are non-parametric, in that they do not rely on any underlying probability distribution, as well as multivariate in that multiple *X*-variables can be used and multiple *Y*-variables can be predicted for any given map unit. These are significant advantages for spatial prediction of ecological communities. On the down side, it has been demonstrated that NN methods may require a relatively large number of reference plots, at least in forest inventory applications where reliable estimates of bias and variance are required (Magnussen et al. 2009). However, the effects of sampling (number of plots and their geographic distribution) on NN imputation of species composition have not been investigated.

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

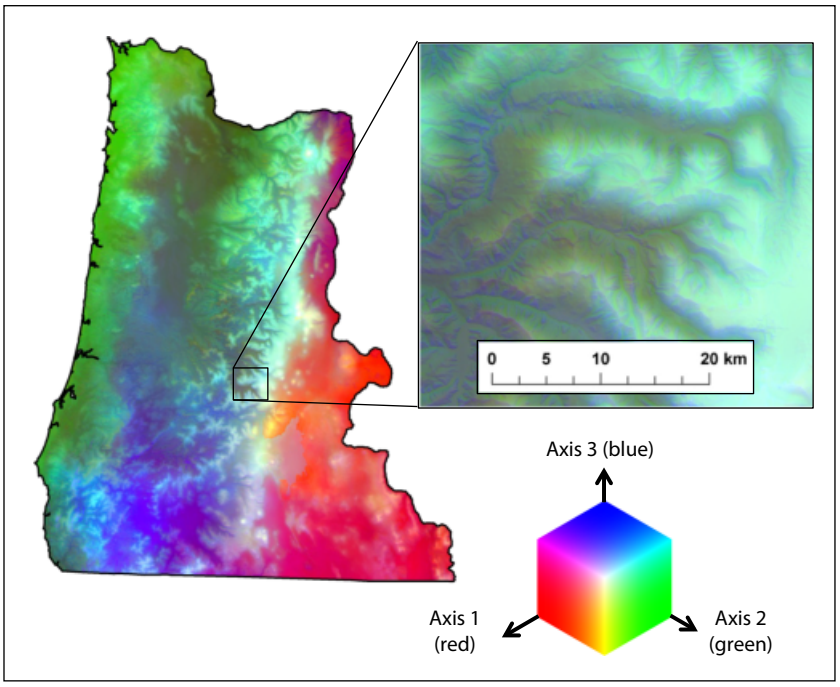

Fig. 4. Map of dominant gradients (scores on CCA axes 1–3), shown as an RGB image composite. Each axis is symbolized with a different colour. Map pixels with low scores on the axis are symbolized with low colour intensity and high scores with high colour intensity. When the three axes are composited, areas with high scores on all three axes display as white (high saturation of all three colours). Axis 1 is most strongly correlated with elevation and temperature, maritime to continental climate and slope; axis 2 with moisture (annual precipitation and growing season moisture stress); axis 3 with summer precipitation.

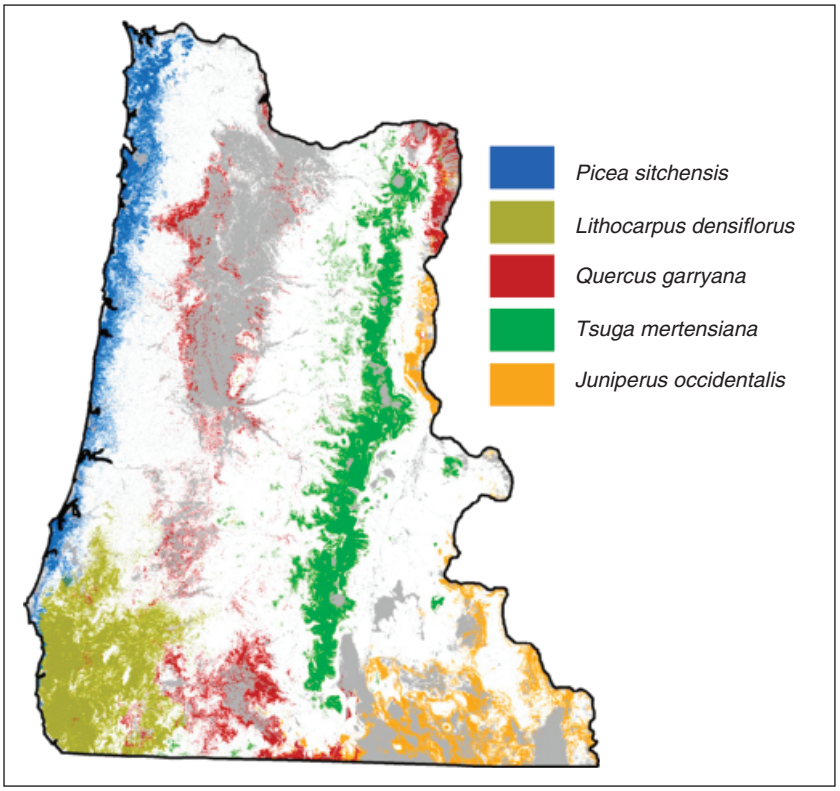

Fig. 5. Predicted presence of five tree species. Where species overlap, only the species listed first in the legend is visible. Grey areas are nonforest.

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

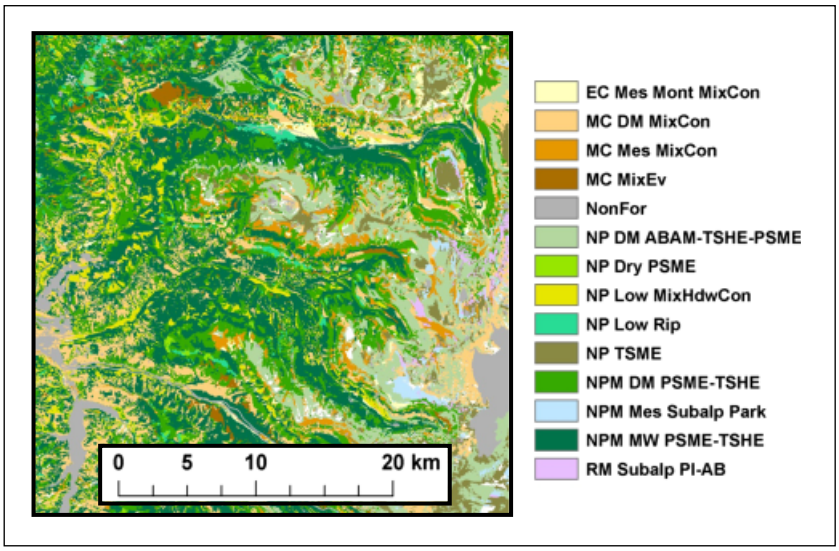

Fig. 6. Map of forest and woodland Ecological Systems in the western Cascade Mountains (same location as enlarged area in Fig. 4). See Table 3 for full names of Ecological Systems. Only those Systems that occupy at least 1% of forest area are shown.

When NN imputation is performed with $k = 1$, it is critical that field plots sample the range of variation in community composition present on the landscape, because imputed (mapped) values are limited to values observed on the plots. Species combinations that occur in nature but are not represented in the plot sample will not be present in the imputed map; but conversely, neither will the imputation contain novel communities that do not occur in nature (i.e. errors). Many studies have demonstrated that imputation with $k = 1$ preserves the variance structure of the reference (plot) data. In addition, McRoberts et al. (2007) found that areal estimates obtained using both probability-based and model-based approaches were unbiased. Our results suggest that model-based GNN with $k = 1$ provides areal distributions of community types (Ecological Systems) that are consistent with sample-based estimates (Table 3), despite being based on reference data that are not geographically balanced. Although our sample size is quite large and spans the range of variation in the study area, only the FIA and CVS plots are on systematic grids, the CVS plots are four times the sampling density of FIA, and federal lands are much more densely sampled (by CVS and Ecology plots) than nonfederal lands (FIA plots).

Diagnostics of NN model performance

In this paper we present several measures of predictive performance that describe a single GNN imputation model, but that address different aspects of model validity. We highlight measures of species composition (species, community characteristics such as diversity, and community types), as these have received relatively little attention in the NN literature beyond the mapping of generalized forest types (e.g. Thessler et al. 2008; Tomppo et al. 2009). We hope that we have provided vegetation scientists, map developers and map users with a basis for weighing the strengths and limitations of NN imputation against other available modelling methods and map products. We advise caution in comparing our model accuracy for species presence/absence to other published models, because the kappa statistic can be sensitive to species prevalence (Franklin 2009). More research is needed to develop ways of expressing imputed maps in terms of probability surfaces, to enable calculation of metrics such as AUC that depend on varying the threshold probability for classification.

Ultimately, choice of a modelling approach hinges on objectives, scale (both space and time), organisms studied and limitations and error structure of the modelling approach relative to objectives. Perhaps the greatest strength of NN models and maps is their utility for serving multiple objectives – either simultaneously within a single multi-faceted analysis, or for multiple users having a variety of objectives. In cases where maps of species composition are sought, we expect that users most often will find advantages in imputed maps based on $k = 1$ due to the maintenance of species co-occurrence within map units. We provide a suite of model diagnostics that evaluate various aspects of error and uncertainty for many different vegetation attributes. The onus is left on the user to evaluate model sufficiency and utility for each particular application. However, in practice we have observed that the general "look-and-feel" of the maps, and how ecologically realistic they are perceived to be,

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

Table 3. Accuracy for 24 forest Ecological Systems occupying at least 1% of forest area in the GNN model. Accuracy and fuzzy accuracy are from an error matrix constructed from cross-validation. Producer's accuracy is the percentage of plots of a given class that were correctly classified in the map. User's accuracy is the percentage of pixels of a given class that were correctly classified by plots located within it. Areal representation shows percentage of forest area in the imputed map, and percentage difference is the amount of over- or under-representation compared to the sample-based estimate from the FIA plots.

Ecological System*	Accuracy			Fuzzy accuracy			Areal representation	
	Producer's	User's	Kappa	Producer's	User's	Kappa	% of map	% diff.
NP QU	15.4	15.4	0.15	55.0	52.8	0.67	1.1	0.1
Cal Coast SESE3	29.5	30.0	0.29	57.5	55.8	0.65	1.0	0.7
CP JUOC	27.3	32.6	0.30	76.4	78.3	0.85	0.9	−0.2
EC Mes Mont MixCon	47.8	48.1	0.47	71.0	69.0	0.75	2.0	−0.4
KS LM Serp MixCon	32.8	31.9	0.32	50.3	49.5	0.57	0.6	0.4
MC DM MixCon	30.1	29.4	0.22	80.9	80.0	0.85	7.1	1.0
MC Mes MixCon	53.0	53.5	0.46	84.3	84.6	0.85	10.6	1.1
MC Low Mont QUKE-Con	37.8	37.5	0.35	79.1	78.5	0.84	3.0	0.5
MC ABSH	49.6	47.7	0.47	77.5	77.7	0.82	1.5	0.4
NP Dry PSME	11.7	12.7	0.10	53.9	53.4	0.66	3.2	−1.6
NPHM PISI	57.3	51.4	0.54	85.1	82.1	0.87	1.9	0.3
NPM DM PSME-TSHE	32.7	33.3	0.24	91.2	90.0	0.93	11.5	−2.1
NPM Mes Subalp Park	27.6	27.2	0.27	58.6	52.7	0.65	0.6	0.5
NPM MW PSME-TSHE	48.0	46.6	0.39	84.5	84.5	0.86	13.8	1.5
NP TSME	39.9	40.3	0.39	85.4	85.4	0.90	2.0	−0.8
MC MixEv	56.8	55.7	0.53	82.8	81.9	0.85	5.4	1.0
NRM PIPO	63.4	63.3	0.61	80.6	81.9	0.83	9.5	0.2
RM Poor-site PICO	50.5	49.1	0.48	56.0	55.4	0.57	4.0	−1.1
NP DM ABAM-TSHE-PSME	43.6	42.9	0.41	89.4	91.1	0.93	3.2	1.0
EC QU-PIPO	26.0	27.0	0.26	75.0	77.0	0.84	0.7	0.1
NP Broadleaf Landslide	7.5	9.1	0.08	73.9	75.0	0.84	1.5	−0.8
NP Low MixHdwCon	23.5	24.1	0.20	79.3	78.8	0.86	6.9	1.0
NP Low Rip	5.2	7.5	0.06	46.8	48.6	0.63	0.9	−0.2
All Ecological Systems	34.5	34.5	0.32	70.6	70.3	0.76	3.9	0.1

* NP QU = North Pacific Oak Woodland; Cal Coast SESE3 = California Coastal Redwood Forest; CP JUOC = Columbia Plateau Western Juniper Woodland and Savanna; EC Mes Mont MixCon = East Cascades Mesic Montane Mixed-Conifer Forest and Woodland; KS LM Serp MixCon = Klamath-Siskiyou Lower Montane Serpentine Mixed Conifer Woodland; MC DM MixCon = Mediterranean California Dry-Mesic Mixed Conifer Forest and Woodland; MC Mes MixCon = Mediterranean California Mesic Mixed Conifer Forest and Woodland; MC Low Mont QUKE-Con = Mediterranean California Lower Montane Black Oak-Conifer Forest and Woodland; MC ABSH = Mediterranean California Red Fir Forest; NP Dry PSME = North Pacific Dry Douglas-fir Forest and Woodland; NPHM PISI = North Pacific Hypermaritime Sitka Spruce Forest; NPM DM PSME-TSHE = North Pacific Maritime Dry-Mesic Douglas-fir–Western Hemlock Forest; NPM MW PSME-TSHE = North Pacific Maritime Mesic-Wet Douglas-fir–Western Hemlock Forest; NPM Mes Subalp Park = North Pacific Maritime Mesic Subalpine Parkland; NP TSME = North Pacific Mountain Hemlock Forest; MC MixEv = Mediterranean California Mixed Evergreen Forest; NRM PIPO = Northern Rocky Mountain Ponderosa Pine Woodland and Savanna; SN Subalp PICO = Sierra Nevada Subalpine Lodgepole Pine Forest and Woodland; RM Poor-site PICO = Rocky Mountain Poor-Site Lodgepole Pine Forest; NP DM ABAM-TSHE-PSME = North Pacific Dry-Mesic Silver Fir–Western Hemlock–Douglas-fir Forest; EC QU-PIPO = East Cascades Oak–Ponderosa Pine Forest and Woodland; NP Broadleaf Landslide = North Pacific Broadleaf Landslide Forest and Shrubland; NP Low MixHdwCon = North Pacific Lowland Mixed Hardwood Conifer Forest and Woodland; NP Low Rip = North Pacific Lowland Riparian Forest and Shrubland.

often outweigh quantitative model diagnostics in the eyes of many map users.

In NN maps based on $k = 1$, alpha diversity (species richness) in the imputed maps faithfully reproduces that observed in the plot sample, at both the local (plot and map unit (pixel)) and regional (median values across all plots and all pixels) scales. In our model, for any given location (local scale), on average the list of (ten) woody plant species was balanced between errors of omission and commission. Across the entire model region, where species prevalence was low and beta diversity was high, there were more errors of omission than commission. In terms of total forest area, the model on average predicted slightly more area occupied by individual species than expected from a systematic sample of plots. For the classification of Ecological Systems, errors of omission and commission were balanced, and differences in areal representation between the plot sample and the imputed map were quite small. This contrasts with our experience in mapping Ecological Systems with a Random Forest predictive model outside of the imputation context (Grossmann et al. 2010).

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

Applications to landscape analysis and modelling

Several key features of ($k = 1$) NN imputation maps distinguish them from alternative types of vegetation maps, and offer advantages for certain applications in landscape analysis and modelling. These key features are: (1) maps of many individual species, which can be post-classified into plant communities by the user; (2) species assemblages within imputed map units that are ecologically realistic; and (3) unbiased areal representation of species and communities, for both rare and common entities. Therefore, in aggregate, modelled data across large regions will contain the full diversity and range of variability present in the plot sample. On the down side, perhaps the most important limitation of $k = 1$ NN maps is that prediction accuracy as assessed at the local scale varies greatly among species and community types, and may be lower for some species and types than can be achieved with other methods. Much of this is a manifestation of "pure error" (Stage & Crookston 2007), which reflects natural variability in plant communities. Nevertheless, as pointed out by Temesgen et al. (2003), NN methods may perform well for complex stands with multiple species. In addition, quality of the maps strongly depends on a representative sample of plots, since $k = 1$ imputation will not interpolate nor extrapolate beyond the data.

Areal representation often is not reported for published studies (but see Riemann et al. 2010), yet can be critically important to applications in natural resource management and conservation planning. Landscapes commonly contain a few community types that comprise most of the area, but also many minor types that are of particular interest or value or that require special management attention. Modelling approaches vary in terms of ability to predict abundant versus rare species or communities, and improved accuracy for common types often is achieved at the expense of representation of less common types. Users may need to decide whether local accuracy or regional representation is more important to their application. Because map uncertainty varies widely among vegetation attributes, and with the scale at which error is assessed, this challenges modellers to communicate implications to users (and for users to convey their needs to modellers).

The attribution of each map unit with a suite of species that are known to co-occur in nature is especially useful for characterizing landscape conditions as input to landscape models where multi-species information is desired. For example, a study in southwest Oregon modelled the potential and actual distribution of sudden oak death in order to prioritize landscapes for early detection and eradication of disease outbreaks (Václavík et al. 2010). GNN maps of 14 species of tree and shrub were used to map the abundance and susceptibility of host vegetation as one input to the invasive species distribution models.

Landscape scenario modelling is increasingly employed to explore potential effects of alternative land management policies, changes to natural and human-caused disturbance regimes, a changing climate, or other assumptions. For example, a study assessed potential ecological and socioeconomic effects of forest policies over 100 years of change for a large multi-ownership region in coastal Oregon (Spies et al. 2007). GNN maps of forest composition and structure provided initial landscape conditions for modelling (Ohmann et al. 2007). Another large interagency project utilized GNN maps as the basis for landscape state-and-transition models across Washington and Oregon to compare likely outcomes of alternative management strategies and disturbance regimes over one to several centuries (Moeur et al. 2009). Results are being used to guide natural resource planning by state and federal agencies and conservation planning by private organisations.

NN methods can be useful in characterizing large landscapes for conservation planning, as illustrated by recent application of GNN to map Ecological Systems in eastern Oregon and Washington for national conservation planning for the Gap Analysis Program (Kagan et al. 2008).

Conclusions

In conclusion, NN imputation is worth adding to the vegetation scientist's toolbox as one potential modelling approach when spatial predictions (maps) are desired, or for other "missing data" problems. Methods are applicable to any geographic location where sufficient plot and spatial data are available. Resulting models provide quantitative and spatial information on regional ecological gradients, as well as vegetation maps that are useful for a variety of applications. Although NN imputation has the limitations of any static, correlative approach, resulting maps of multiple species and plant community types are especially useful for initializing large landscapes as input to other models that portray landscape change over time and space, under varying assumptions such as a changing climate.

Acknowledgements

Our research would not have been possible without the efforts of numerous collectors and managers of the regional field plot data. We thank Dave Roberts and one anonymous referee for their helpful comments on an earlier version of this manuscript.

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

References

Agee, J.K. 1993. *Fire ecology of Pacific Northwest forests*. Island Press, Covelo, CA, US.

Austin, M.P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157: 101–118.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5–32.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

Comer, P., Faber-Langendoen, D., Evans, R., Gawler, S., Josse, C., Kittel, G., Menard, S., Pyne, S., Reid, M., Schulz, K., Snow, K. & Teague, J. 2003. *Ecological Systems of the United States: a working classification of U.S. terrestrial systems*. NatureServe, Arlington, VA, US.

Congalton, R.G. & Green, K. 1999. *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis Publishers, Boca Raton, FL, US.

Crookston, N.L. & Finley, A.O. 2008. yaImpute: an R package for *k*NN imputation. *Journal of Statistical Software* 23: 1–16.

Daly, C., Halbleib, M., Smith, J.L., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J. & Pasteris, P.P. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology* 28: 2031–2064.

Dirnböck, T., Dullinger, S., Gottfried, M., Ginzler, C. & Grabherr, G. 2003. Mapping alpine vegetation based on image analysis, topographic variables and Canonical Correspondence Analysis. *Applied Vegetation Science* 6: 85–96.

Eskelson, B.N.I., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L. & Hudak, A.T. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research* 24: 235–246.

Ferrier, S. & Guisan, A. 2006. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43: 393–404.

Finley, A.O. & McRoberts, R.E. 2008. Efficient *k*-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment* 112: 2203–2211.

Franco-Lopez, H., Ek, A.R. & Bauer, M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment* 77: 251–274.

Franklin, J. 2009. *Mapping species distributions*. Cambridge University Press, New York, NY, US, 320pp.

Franklin, J.F. & Dyrness, C.T. 1973. *Natural vegetation of Oregon and Washington*. US Department of Agriculture, Forest Service, General Technical Report PNW-8.

Gopal, S. & Woodcock, C.E. 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing* 60: 181–188.

Gottfried, M., Pauli, H. & Grabherr, G. 1998. Prediction of vegetation patterns at the limits of plant life: a new view of the alpine–nival ecotone. *Arctic and Alpine Research* 30: 207–221.

Grossmann, E.B., Ohmann, J.L., Kagan, J., May, H.K. & Gregory, M.J. 2010. Mapping ecological systems with a random forest model: tradeoffs between errors and bias. *USGS Gap Analysis Bulletin*, No. 17; February 2010; pp 16–22.

Guisan, A., Weiss, S.B. & Weiss, A.D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* 143: 107–122.

Hanley, J.A. & McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristics curve. *Radiology* 143: 29–36.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E. & Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112: 2232–2245.

Kagan, J.S., Ohmann, J.L., Gregory, M. & Tobalske, C. 2008. Land cover map for map zones 8 and 9 developed from SAGEMAP, GNN, and SWReGAP: a pilot for NWGAP. Gap Analysis Bulletin, No. 15; February 2008; pp 15–19.

Kruskal, J.B. & Wish, M. 1978. *Multidimensional scaling*. Sage Publications, Beverly Hills, CA, US, 93pp.

Magnussen, S., McRoberts, R.E. & Tomppo, E.O. 2009. Model-based mean square error estimators for *k*-nearest neighbor predictions and applications using remotely sensed data for forest inventories. *Remote Sensing of Environment* 113: 476–488.

McCune, B. & Grace, J.B. 2002. *Analysis of ecological communities*. MjM Software Design, Glenedon Beach, OR, US, 300pp.

McRoberts, R.E. 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment* 113: 489–499.

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.

McRoberts, R.E., Tomppo, E.O., Finley, A.O. & Heikkinen, J. 2007. Estimating areal means and variances of forest attributes using the *k*-Nearest Neighbors technique and satellite imagery. *Remote Sensing of Environment* 111: 466–480.

McRoberts, R.E., Tomppo, E.O. & Næsset, E. 2010. Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research* 25: 368–381.

Moeur, M. & Stage, A.R. 1995. Most Similar Neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science* 41: 337–359.

Moeur, M., Ohmann, J., Hemstrom, M., Burcsu, T. & Merzenich, J. 2009. Projecting watershed condition with Interagency Mapping and Assessment Project (IMAP) vegetation data and landscape models. In: Bayer, J.M. & Schei, J.L. (eds.) *PNAMP special publication: remote sensing applications for aquatic resource monitoring. Chapter 11*. pp.

Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.

83–91. Pacific Northwest Aquatic Monitoring Partnership, Cook, WA, US.

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, USA. *Canadian Journal of Forest Research* 32: 725–741.

Ohmann, J.L. & Spies, T.A. 1998. Regional gradient analysis and spatial pattern of woody plant communities of Oregon forests. *Ecological Monographs* 68: 151–182.

Ohmann, J.L., Gregory, M.J. & Spies, T.A. 2007. Influence of environment, disturbance, and ownership on forest vegetation of coastal Oregon. *Ecological Applications* 17: 18–33.

Økland, R.H. 1999. On the variation explained by ordination and constrained ordination axes. *Journal of Vegetation Science* 10: 131–136.

Palmer, M. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* 74: 2215–2230.

Pierce, K.B., Lookingbill, T.R. & Urban, D.L. 2005. A simple method for estimating potential relative radiation (PRR) for landscape-scale vegetation analysis. *Landscape Ecology* 20: 137–147.

Pierce, K.B. Jr., Ohmann, J.L., Wimberly, M.C., Gregory, M.J. & Fried, J.S. 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: 1901–1916.

Riemann, R., Wilson, B.T., Lister, A. & Parks, S. 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sensing of Environment* 114: 2337–2352.

Roberts, D.W. 2009. Comparison of multi-dimensional fuzzy set ordination with CCA and DB-RDA. *Ecology* 90: 2622–2634.

Schmidtlein, S., Zimmermann, P., Schupferling, R. & Weiss, C. 2007. Mapping the floristic continuum: ordination space position estimated from imaging spectroscopy. *Journal of Vegetation Science* 18: 131–140.

Spies, T.A., Johnson, K.N., Burnett, K.M., Ohmann, J.L., McComb, B.C., Reeves, G.H., Bettinger, P., Kline, J.D. & Garber-Yonts, B. 2007. Cumulative ecological and socioeconomic effects of forest policies in coastal Oregon. *Ecological Applications* 17: 5–17.

Stage, A.R. & Crookston, N.L. 2007. Partitioning error components for accuracy assessment of near-neighbor methods of imputation. *Forest Science* 53: 62–72.

Temesgen, H., LeMay, V., Froese, K.L. & Marshall, P.L. 2003. Imputing tree lists from aerial attributes for complex stands of south-eastern British Columbia. *Forest Ecology and Management* 177: 277–285.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167–1179.

Thessler, S., Ruokolainen, K., Tuomisto, H. & Tomppo, E. 2005. Mapping gradual landscape-scale floristic changes in Amazonian primary rain forests by combining ordination and remote sensing. *Global Ecology & Biogeography* 14: 315–325.

Thessler, S., Sesnie, S., Ramos, Z.S.B., Ruokolainen, K., Tomppo, E. & Finegan, B. 2008. Using k-NN and discriminant analyses to classify rain forest types in a Landsat TM image over northern Costa Rica. *Remote Sensing of Environment* 112: 2485–2494.

Tomppo, E. 1991. Satellite image-based national forest inventory of Finland. *International Archives of Photogrammetry and Remote Sensing* 28: 419–424.

Tomppo, E., Gagliano, C., De Natale, F., Katila, M. & McRoberts, R.E. 2009. Predicting categorical forest variables using an improved *k*-nearest neighbour estimator and Landsat imagery. *Remote Sensing of Environment* 113: 500–517.

USDA NRCS. 2000. *The PLANTS database (http://plants.usda.gov)*. National Plant Data Center, Baton Rouge, LA, US.

Václavík, T., Kanaskie, A., Hansen, E., Ohmann, J.L. & Meentemeyer, R. 2010. Modeling potential vs. actual distribution of sudden oak death in Oregon: prioritizing landscape contexts for early detection and eradication of disease outbreaks. *Forest Ecology and Management* 260: 1026–1035.

 Doi: 10.1111/j.1654-1103.2010.01244.x. Published 2011.