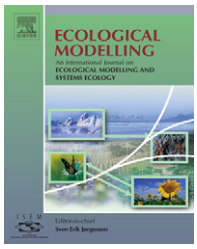
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Effects of sample survey design on the accuracy of classification tree models in species distribution models

Thomas C. Edwards Jr.[a,*], *D. Richard Cutler*[b], *Niklaus E. Zimmermann*[c], *Linda Geiser*[d], *Gretchen G. Moisen*[e]

[a] *USGS Utah Cooperative Fish and Wildlife Research Unit, College of Natural Resources, Utah State University, Logan, UT 84322-5290, USA*
[b] *Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900, USA*
[c] *Department of Landscape Research, Swiss Federal Research Institute WSL, Zuercherstrasse 111, CH-8903 Birmensdorf, Switzerland*
[d] *USDA Forest Service, Siuslaw National Forest, P.O. Box 1148, Corvallis, OR 97339, USA*
[e] *USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA*

ARTICLE INFO

Article history:
Published on line 24 July 2006

Keywords:
Model accuracy
Sample survey
Study design
Classification trees
Lichens
Accuracy assessment
Probability samples
Non-probability samples

ABSTRACT

We evaluated the effects of probabilistic (hereafter DESIGN) and non-probabilistic (PURPOSIVE) sample surveys on resultant classification tree models for predicting the presence of four lichen species in the Pacific Northwest, USA. Models derived from both survey forms were assessed using an independent data set (EVALUATION). Measures of accuracy as gauged by resubstitution rates were similar for each lichen species irrespective of the underlying sample survey form. Cross-validation estimates of prediction accuracies were lower than resubstitution accuracies for all species and both design types, and in all cases were closer to the true prediction accuracies based on the EVALUATION data set. We argue that greater emphasis should be placed on calculating and reporting cross-validation accuracy rates rather than simple resubstitution accuracy rates. Evaluation of the DESIGN and PURPOSIVE tree models on the EVALUATION data set shows significantly lower prediction accuracy for the PURPOSIVE tree models relative to the DESIGN models, indicating that non-probabilistic sample surveys may generate models with limited predictive capability. These differences were consistent across all four lichen species, with 11 of the 12 possible species and sample survey type comparisons having significantly lower accuracy rates. Some differences in accuracy were as large as 50%. The classification tree structures also differed considerably both among and within the modelled species, depending on the sample survey form. Overlap in the predictor variables selected by the DESIGN and PURPOSIVE tree models ranged from only 20% to 38%, indicating the classification trees fit the two evaluated survey forms on different sets of predictor variables. The magnitude of these differences in predictor variables throws doubt on ecological interpretation derived from prediction models based on non-probabilistic sample surveys.

* *Corresponding author at*: USGS Utah Cooperative Research Unit, 5290 Old Main Hill, Utah State University, Logan, UT 84322-5290, USA. Tel.: +1 435 797 2529; fax: +1 435 797 4025.
E-mail address: tce@nr.usu.edu (T.C. Edwards Jr.).
0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.016

1. Introduction

Species distribution models rely heavily on the collection of site-specific data from the hypothesized spatial and environmental ranges of target species. Once the data are collected, many different types of analytical methods can be used to ascertain relationships between measured environmental variables and species occurrence. One of the more common forms of analysis applied to data of this type is that of classification models, which are used to discriminate between nominal response values using a set of environmental predictors. Applications of classification models to species distribution modelling abound in the published literature, and range from predicting the distribution or characteristics of plant species (Austin et al., 1983, 1990; Frescino et al., 2001; Zimmermann and Kienast, 1999) to habitat relationships of terrestrial animal species (McNoleg, 1996; Jaberg and Guisan, 2001; Lawler and Edwards, 2002; Welch and MacMahon, 2005). An excellent overview of the state of species distribution modelling can be found in Scott et al. (2002).

Classification models can be created using various statistical approaches, including generalized linear models (GLM) (McCullagh and Nelder, 1989) such as logistic regression (Hosmer and Lemeshow, 2000), generalized additive models (GAM) (Hastie and Tibshirani, 1990; Yee and Mitchell, 1991), which are semi-parametric extensions of GLMs, and fully non-parametric methods such as classification trees (Breiman et al., 1984; De'ath and Fabricius, 2000). The latest generation of statistical classification procedures, including support vector machines, random forests (Breiman, 2001), and other ensemble classifiers (Steele, 2000), have strong potential for ecological classification but have not yet received much attention. Irrespective of the selected analytical methodology, the ecological object to be modelled – be it species presence, or some attribute of species presence such as a bird nesting site or habitat use – is typically surveyed, and a response, often nominal, is tallied and linked with a set of ecological predictor variables. Depending on the analytical method, the predictors can be nominal, ordinal, ratio, or interval scales, or mixtures of both.

Data for species distribution modelling are typically obtained by survey sampling. Ideally, survey sampling involves the random selection and measurement of samples from a defined, target population referred to as the sampling frame. Sampling frames can have many different characteristics. Often sampling frames are a defined spatial extent, considered a finite population, that is divided into N smaller units of some set size, from which a subset n_i is randomly selected and surveyed for the species of interest (see Edwards et al., 2004). These area samples (*sensu* Nusser et al., 1998) may, or may not, correspond with the actual sample unit, which may be something as simple as the presence or absence of a specific species within the area being sampled. Estimates derived from these types of sampling designs are considered design-based, and carry with them the power of inferential statistics.

Many ecological studies, however, deal with attempts to survey and build a classification model for ecological events best described as rare or uncommon on the landscape (Engler et al., 2004; Edwards et al., 2005). Rare ecological events may not be truly amenable to randomization procedures *per se*, especially when the goal is to generate sufficient observations for a classification model (see Edwards et al., 2005). For example, the random selection of small areas within some defined spatial extent like a conservation reserve, which is then surveyed to locate bird nests, might not be a fruitful exercise if too few bird nests are found. While such a design would allow for inferential statistics to be determined for, say percent of sites occupied (see Edwards et al., 2004), it may not provide sufficient tallies of nest presence, thereby precluding attempts to build a classification model. Consequently, ecologists often actively search for the event of interest using non-probability sampling procedures (see Cochran, 1977). One of the more common of these non-probability sampling efforts, termed purposive sampling, occurs when ecologists actively seek the event of interest, such as an active breeding nest of a bird, or a specific plant species.

Our objective was to compare classification models that predict the presence of four lichens common in the Pacific Northwest, USA. These models were developed from environmental data collected from randomly versus purposively selected sample sites. We specifically evaluated two common sample designs used to collect the model building data; the first used presence–absence data collected on a randomly started, systematic grid. Such probability-based sampling efforts support design-based inference (see Gregoire, 1998), and are the basis of many current efforts to model and assess environmental and ecological systems (Olsen and Schreuder, 1997; Olsen et al., 1999). The second set of models was based on presence–absence data collected in a non-probability, or purposive, framework, where biologists used knowledge of lichen life histories to search for and "sample" for lichen presences. Unlike probabilistic samples, those collected in a purposive framework can have questionable inference, due principally to biases associated with the non-random selection of sample locations. Identical predictor variables were used for both sample design forms, allowing us to test whether the probability-based and purposive sampling resulted in the same final models. An independent data set was collected using probability sampling as well and used for validation. Effects of the two sample designs were evaluated by various measures of model accuracy and fit, and by performing both internal cross-validation and external, independent validation exercises. All data were collected in the same spatial area.

2. Methods

2.1. Study design and species

Data used in our analyses were collected from seven national forests and adjacent Bureau of Land Management (BLM) districts in the Cascades and Coast Ranges of Oregon and Washington (Fig. 1). All sample sites in the study region were surveyed at least once as part of a broader effort using epiphytic macrolichens as indicators of air quality (Geiser, 2004). Four common lichen species, *Lobaria oregana* (Tuck.) Müll., *L. pulmonaria* (L.) Hoffm., *Pseudocyphellaria anomala* Brodo & Ahti, and *P. anthraspis* (ACH.) H. Magn., were used in the analyses presented here. Each of these four species had sufficient num-

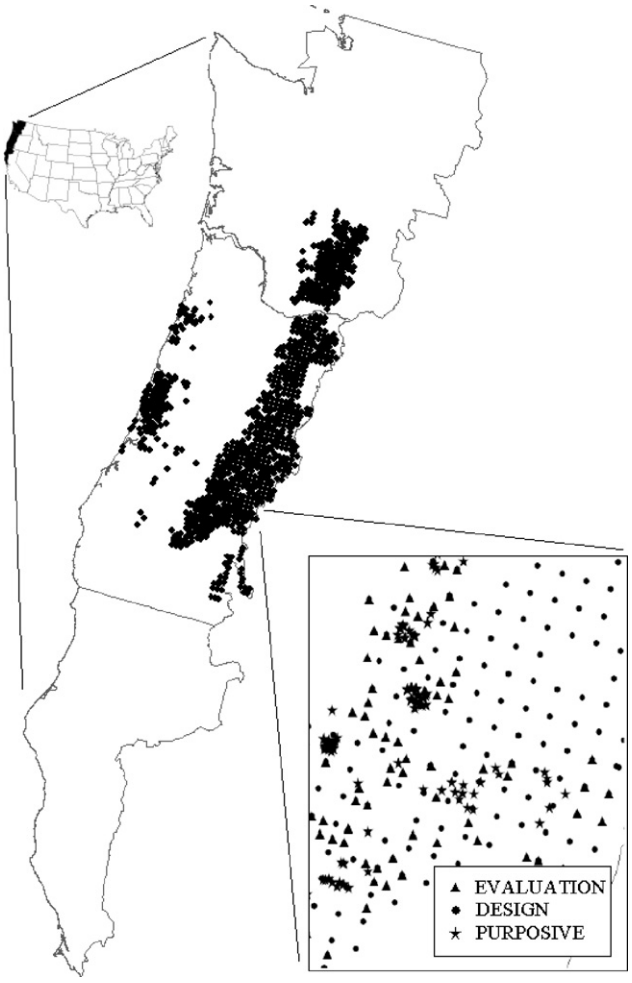


Fig. 1 – Depiction of the Northwest Forest Plan study region, Pacific Northwest, USA. Sample regions for the DESIGN, PURPOSIVE and EVALUATION study areas are in black. Note on inset the regularity of the DESIGN and EVALUATION sample plots, and the clumped nature of the PURPOSIVE plots.

bers of detections (>50) for developing the models used in our comparisons. All four species are large, foliose, broadly distributed cyanolichens that can be found on tree trunks, live branches, and leaf litter of conifers of the Pacific Northwest. All achieve their greatest biomass in riparian and late-seral forests. Eye-level habitat and large size makes them relatively easy to find and identify.

Two sample surveys common to predictive models in ecology were evaluated. The first (hereafter DESIGN) was a probability-based sampling effort. The DESIGN survey used data collected on the Current Vegetation Survey plots (CVS), a randomly started, systematic ~5.47 km grid overlaid on all Forest Service and BLM lands in the Pacific Northwest. Its principal application is the generation of estimates of forest resources (see Max et al., 1996). Sites were surveyed by field botanists trained and certified in the recognition and differentiation of regional epiphytic macrolichens.

The second survey was purposive, a non-probability design where samples are searched for and collected based on a desired need or outcome (hereafter PURPOSIVE). Its common use in ecological studies reflects the reality of ecological modelling exercises involving rare or uncommon events, such as the presence of a bird nest (e.g., Lawler and Edwards, 2002) or a plant species (e.g., Dreisbach et al., 2002). Because the survey sites were not randomly selected, introduced bias is a concern. Here, field botanists essentially searched for lichens using no underlying sample design, relying instead on their personal knowledge of lichen life history to guide their sampling efforts on the landscape. The botanists were searching for targeted lichen species in areas proposed for timber harvest or other management actions on public lands.

The third sampling effort (hereafter EVALUATION) was probability-based, and served as an independent test data set for the DESIGN and PURPOSIVE models. The EVALUATION surveys covered national forests and BLM districts in three regions of the Pacific Northwest, including the southern Washington Cascades, the Oregon Coast Range, and the Umpqua Basin. Within each of the three areas a stratified random sample of sites on a randomly started ~2.7 km grid was obtained. The stratification criteria were reserve status (reserve and non-reserve) and age class (<80 and 80 + years) of the dominant tree species. Allocations to the four strata were 60% to reserve/80+, 20% to reserve/<80, and 10% to each of the non-reserve strata. These allocations reflected the priorities of a different research effort (see Edwards et al., 2005), but are essentially a proportional allocation for the EVALUATION study region.

A total of 840 plots were sampled in the DESIGN, and 299 in the PURPOSIVE effort. These two data sets constituted the training data. An additional 300 sites were surveyed in the EVALUATION design for model assessment purposes. To alleviate concern over the possible effects of the different sample sizes between the DESIGN and PURPOSIVE data sets, we randomly selected 100 samples of size 299 from the DESIGN data, fit a classification tree for the lichen species on each of these subsets, and predicted the occurrence of the lichen species both on the data sets of size 299 on which they were fit and for the EVALUATION data. Mean differences across all evaluation metrics (see below) ranged from 2.8% to 2.1%, well within any binomial error range and indicative of no sample size effect on our results.

Presence and absence of each lichen species was recorded on a 0.4 ha plot centered on the central (#1) subplot on each CVS site for the DESIGN surveys (details in Edwards et al., 2004). Plot size for the EVALUATION surveys was smaller, only 0.2 ha in size. Plot size for the PURPOSIVE surveys was variable according to the size of the proposed sale or management action. If the purpose of our study was to compare the estimated percent occupancy rates from the DESIGN and PURPOSIVE surveys versus the EVALUATION survey, the difference in the size of the sample units would be a concern given that larger plots will have higher probabilities of occupancy. However, the purpose of our analyses is to use the DESIGN and PURPOSIVE data to fit models predicting the likelihood of occurrence of the four identified lichen species, and the EVALUATION data for assessment. For this application it does not matter if the EVALUATION plot size is the same, larger, or smaller than the DESIGN and PURPOSIVE plot size.

Table 1 – Topographic, bio-climatic and vegetation variables used to model the probability of presence for four lichen species in the study area of the Pacific Northwest Forest Plan

Variable type/name	Description	Units
Topographic		
SLPE	Percent slope	Percent, 0–90
ASPE	Aspect	Degree 0, 1–360
ELEV	Elevation	m
Bio-climatic		
ETPJ	Potential evapotranspiration	mm
MIND	Monthly moisture index	cm
PREC	Precipitation	cm
RELH	Relative humidity	Percent
SFMM	Monthly potential global radiation	kJ
TAVE	Monthly average temperature	°C
TDAY	Monthly average daytime temperature	°C
TMAX	Maximum temperature	°C
TMIN	Minimum temperature	°C
VPAM	Ambient vapor pressure	Pa
VPSA	Saturated vapor pressure	Pa
Vegetation		
BDLCNT	Percent broadleaf cover	Percent, 0–100
CNFCNT	Percent conifer cover	Percent, 0–100
FORBIO	Live tree (>1 in DBH) biomass, above ground dry weight	tonnes/acre
VEGCNT	Percent vegation cover	Percent, 0–100

Both the DESIGN and PURPOSIVE training data sets were co-located in the Cascade Mountains of Oregon and Washington, while the independent EVALUATION data set was collected in the Oregon Coastal Range (Fig. 1). The EVALUATION data set was not co-located in the same spatial extent as the two training data sets to allow for an independent test of model extrapolative capabilities. All three data sets were within the known distribution ranges of all four of the surveyed lichens.

2.2. *Data structure and characteristics*

All plot locations were intersected with maps of topographic, bio-climatic, and vegetation variables (Table 1) in a geographic information system (GIS). The selected environmental variables were all hypothesized to have direct relationships to the presence of the modelled lichen species. Ninety meters resolution topographic variables (slope, aspect and elevation) were obtained by resampling the 30 m resolution National Elevation Data set (NED) (Gesch et al., 2002).

Bio-climatic variables were derived from the DAYMET 1 km daily-gridded weather surfaces that have been reduced to 18-year monthly and yearly climatological summaries (1981–1998) (Thornton et al., 1997; Thornton and Running, 1999). Preliminary analyses showed that correlations among the monthly values for the 11 sets of bio-climatic predictor variables were high. To address the issue of collinearity, a principal components analysis was carried out on each of the 11 sets of monthly bio-climatic predictors. In each case, the first principal component was an average of the 12 monthly measurements, while the second principal component was a contrast of values for 6 so-called summer months (April–September) to the 6 so-called winter months (October–March). For each set of 12 monthly variables, these two principal components explained over 95% of the variability, and in most cases the first two principal components explained over 99% of the variability in the sets of variables. Accordingly, we defined two new variables for each set of monthly bio-climatic predictors: (1) the average of the 12 monthly variables; (2) the difference between the sum of the summer monthly values and the winter monthly values, divided by 12. Hereafter we use the variable suffix "A" to denote the average of the 12 monthly measurements, and the suffix "D" to denote the difference derived variable. Thus, TMINA is the average minimum temperature for the 12 months and PRECD is the difference between summer and winter precipitation. See Edwards et al. (2005) for additional specifics on the derivation of the DAYMET-based bio-climatic variables.

Vegetation variables came from the BLM Interagency Vegetation Mapping Project[1] (IVMP) and the USDA Forest Service, Forest Inventory and Analysis program[2] (FIA). Variables from the IVMP were derived from Landsat Thematic Mapper imagery, and included percent cover of all vegetation, conifer, and broadleaf species. Maps were obtained at a resolution of 25 m, and were resampled using a bilinear interpolation to 90 m resolution within a GIS. The two vegetation variables obtained from the FIA were forest type and above ground live tree biomass, both modelled for the continental US, Alaska, and Puerto Rico at 250 m resolution. These nationwide maps were constructed by modeling forest type and biomass collected on FIA sample plots as nonparametric functions of numerous ancillary predictor layers, including: 16-day Moderate Resolution Imaging Spectrometer (MODIS) composites and associated vegetation indices and MODIS percent tree cover; vegetative diversity and type synthesized from the National Land Cover Dataset; topographic variables derived from Digital Elevation Models; monthly and annual climate parameters; other ancillary variables (J. Blackard, USFS, personal communication, 2005).

2.3. *Statistical modelling and assessment*

We used classification trees (Breiman et al., 1984) to relate the DESIGN and PURPOSIVE lichen presences to the modelled topographic, bio-climatic, and vegetation predictor variables. We followed the approach suggested by De'ath and Fabricius (2000), pruning trees by cross-validation and the 1-SE rule. Separate models were built for each species in each of the DESIGN and PURPOSIVE surveys. We next applied the resultant sampling design ($n = 2$) and species-specific ($n = 4$) classification tree models (total models = 8) to our spatially explicit predictors within the GIS, and modelled the probability of presence of each of the four lichen species for the EVALUATION sample plots. Observations from the EVALUATION study area were

[1] http://www.or.blm.gov/gis/projects/vegetation/ivmp.
[2] http://www.fia.fs.fed.us/.

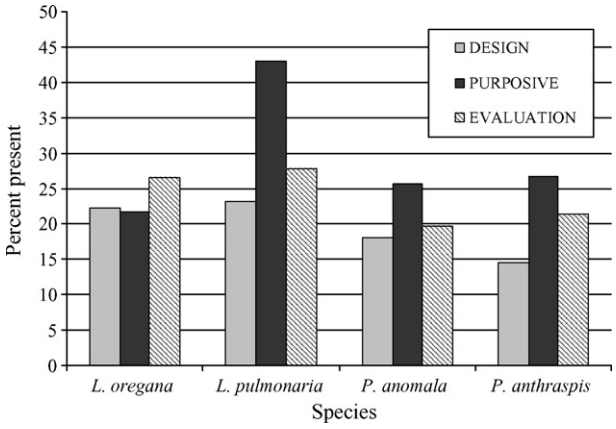


Fig. 2 – Percent of sample plots at which each lichen species was present in the DESIGN ($n = 840$), PURPOSIVE ($n = 299$), and EVALUATION ($n = 300$) study sites, area of the Pacific Northwest Forest Plan, Pacific Northwest, USA. Number of detections for each species and sampling form are shown above each histogram bar.

then compared against the predicted DESIGN and PURPOSIVE probabilities for that plot, allowing us to link the presence or absence of each of the four common species to an estimate of the probability of presence. All trees were fit using the *rpart* library of functions in the *R* statistical package[3] (Ihaka and Gentleman, 1996).

Three general classes of model accuracy were used to compare the models based on the DESIGN and PURPOSIVE data. These classes were: (1) resubstitution (model) accuracy rates for each of the eight models; (2) 10-fold cross-validation (Manly, 1997) estimates of accuracy for each of the eight models; (3) a prediction accuracy rate for the independent EVALUATION survey, based on a probability of presence threshold of $P > 0.5$. Threshold-dependent measures of accuracy reported and assessed included percent correct classification rate (PCC), sensitivity, specificity, and kappa (after Fielding and Bell, 1997). Cut-off for these metrics was assumed to be 0.5. The areas under the receiver operator characteristic (AUC), plus SE, were also calculated from formulae provided by Hanley and McNeil (1982).

3. Results

L. pulmonaria, *P. anomala* and *P. anthraspis* occurred with the greatest frequency in the PURPOSIVE compared to the DESIGN and EVALUATION surveys (Fig. 2). *L. oregana* occurred with the greatest frequency in the EVALUATION survey, and with similar frequency in the PURPOSIVE and DESIGN surveys.

Classification tree models for the four lichen species ranged from 88.0% to 91.1% and 83.8% to 92.6% accurate (PCC) for the DESIGN and PURPOSIVE surveys, respectively (Table 2). Measures of tree specificity and sensitivity between the survey forms were similar between species, too, indicating that the trees were maximizing classification to the extent possible given the predicator variables. In short, measures of accuracy as gauged by resubstitution rates were similar for each species irrespective of sample survey form.

Cross-validated accuracy estimates were lower for all species and sampling forms, with percent correct classification rates ranging from 79.7% to 85.6% for the DESIGN and from 55.7% to 78.6% for the PURPOSIVE design (Table 2). As with the PCC, sensitivities were similar in the magnitude of difference between the DESIGN and PURPOSIVE, but lower overall relative to PCC. Specificity decreased dramatically when DESIGN and PURPOSIVE were compared by species. Similar patterns in decline occurred for kappa and AUC as well.

Tree structure differed considerably both among and within the modelled species, depending on the sampling form (Table 3). One indication of this variability in model complexity is based on the number of levels required to build the trees. For example, the number of tree levels for *L. oregana* was 11 for the DESIGN, but only 4 for the PURPOSIVE sampling form. *P. anomala* had 10 levels for the DESIGN and 7 for PURPOSIVE. Both *L. pulmonaria* and *P. anthraspis* had 8 and 9, and 7 and 8, levels for the DESIGN and PURPOSIVE sampling forms, respectively. No real pattern is apparent.

Pattern was lacking in tree variable selection as well. While differences among the variables used by the tree classifier were expected among species, the magnitude of differences in variables used to discriminate presence within a species varied tremendously by sampling form (Table 3). One indication of this variability is the overlap in tree-selected variables. Overall, 18 of the 21 tree-selected variables (85.7%) overlapped among the species and survey forms. However, variable overlap was considerably lower when survey forms were compared within a species. Overlap based on sampling form was only 3 of 11 (27.3%) for *L. oregana*, 5 of 13 (38.5%) for *L. pulmonaria* and *P. anomala*, and 3 of 15 (20.0%) for *P. anthraspis*, indicating the sampling forms were classifying on different sets of predictor variables.

Assessment of the DESIGN and PURPOSIVE tree models on the EVALUATION validation data set shows lower overall prediction accuracy for the PURPOSIVE tree models compared to DESIGN models (Table 4). This is consistent across all four lichen species, with 11 of the 12 possible comparisons having significantly lower accuracy rates. Differences in specificity are particularly large, typically about 50%. Differences in sensitivity are not as severe, but still indicate an overall pattern of lower accuracy in the PURPOSIVE tree models. Measures of kappa and AUC were significantly lower for the PURPOSIVE compared to the DESIGN survey, too.

Resubstitution DESIGN and PURPOSIVE PCC, sensitivity, and specificity accuracy rates were statistically greater than the EVALUATION accuracy rates for 36 of 40 possible tests of species, sampling form and accuracy rates (Table 5). Both the DESIGN and PURPOSIVE sampling forms consistently overestimated resubstitution accuracy irrespective of the species and accuracy rate when compared to the independent EVALUATION data. A different pattern emerges when evaluating cross-validation accuracy rates, where only 12 of 40 possible tests showed difference (Table 6). However, there were extreme differences among the DESIGN and PURPOSIVE sampling forms, with 10 of 20 possible PURPOSIVE tests indicating statistical differences between PURPOSIVE cross-validation

[3] http://www.r-project.org/.

Table 2 – Percent correct classification (PCC), sensitivity and specificity accuracy rates, and kappa and AUC, for the DESIGN and PURPOSIVE classification tree models of four lichen species in the Pacific Northwest, USA

Species	Survey form	Resubstitution					10-Fold cross-validation				
		PCC	Specificity	Sensitivity	Kappa	AUC	PCC	Specificity	Sensitivity	Kappa	AUC
L. oregana	DESIGN	88.6	95.6	64.2	0.644	0.906	79.2	88.5	49.5	0.368	0.734
	PURPOSIVE	85.6	95.7	49.2	0.515	0.761	78.9	89.7	40.0	0.324	0.601
L. pulminaria	DESIGN	88.0	93.6	69.1	0.650	0.882	81.1	90.6	49.5	0.429	0.788
	PURPOSIVE	78.3	74.7	83.0	0.565	0.721	52.2	61.2	40.3	0.015	0.498
P. anomala	DESIGN	90.1	96.1	63.2	0.640	0.856	84.2	93.3	42.8	0.403	0.763
	PURPOSIVE	81.6	93.2	48.0	0.462	0.782	69.6	82.4	32.5	0.157	0.532
P. anthraspis	DESIGN	91.1	97.4	64.5	0.592	0.835	86.5	93.9	43.9	0.412	0.794
	PURPOSIVE	82.0	93.2	51.2	0.491	0.778	64.5	80.8	20.0	0.009	0.449

Accuracy measures were estimated using both resubstitution and 10-fold cross-validation techniques.

Table 3 – Representation of classifcation trees for the DESIGN and PURPOSIVE sampling surveys and lichen species

Variable type/name	*L. oregana*		*L. pulmonaria*		*P. anomala*		*P. anthraspis*	
	DESIGN	PURPOSIVE	DESIGN	PURPOSIVE	DESIGN	PURPOSIVE	DESIGN	PURPOSIVE
Topographic								
ASPE	10				8			
ELEV	1,2,5,9,10		8				6	6
SLPE			4		10			
Bio-climatic								
ETPJA							5	
ETPJD	3					5		
MINDA		3						
MINDD			8		4,6			
PRECA	8	1					7	3
PRECD				5			5	2
RELHA	8	2,4	3	3,9	3	7	2	
RELHD	7				2			7
SFMMD					9	1		4
SFMMA						3		
TEMPA			1,2	9	1	4		1
TEMPD	2,11	3		1	3,5	2	4	
VPAMA			6				3	
VPAMD			6	2,8			1	
Vegetation								
BDLCNT	6			4,7		6		8
CNFCNT			7		7			2,5
FORBIO	4,6		5	4	9	7		3
VEGCNT	7		7	6			3	

Numbers represent the discrete levels in a tree, with higher numbers indicating splits further from the initial tree node. Empty cells indicate the variable was not relevant to the classification tree model.

accuracy rates and the accuracy rates of the PURPOSIVE models evaluated on the EVALUATION data. In contrast, only 2 of 20 possible DESIGN tests indicated statistical difference between the DESIGN cross-validation accuracy and the DESIGN models evaluated on the EVALUATION data. Cross-validation accuracy rates better reflected the true model accuracy than simple resubstitution accuracy rates even though there were differences between the DESIGN and PURPOSIVE sampling forms.

4. Discussion

Predictive statistical models are commonly used tools to estimate the likelihood of species presence (Edwards et al., 1996; McNoleg, 1996; Frescino et al., 2001; Lawler and Edwards, 2002; Moisen et al., 2006). These tools are essential for the ecological understanding, and management and conservation, of plant and animal species. Levins (1966) was one of the first to note that models must balance between the often competing wishes for generality in application and the specificity desired for prediction. Best and Stauffer (1986), commenting on bird-habitat relationship models, suggested that the specificity desired of predictive models is not likely to be achieved given variability in time and space of both the response and predictor variables. Van Horne and Wiens (1991) regarded an ideal model as one that simultaneously maximizes ecological realism, generality, and simplicity of use. Depending on the objective of the study, one or two of the three is often sacrificed

Table 4 – Percent correct classification (PCC), sensitivity and specificity accuracy rates, and kappa, for the DESIGN and PURPOSIVE classification tree models of four lichen species in the Pacific Northwest, USA, evaluated on the independent EVALUATION data

Species	Survey forms evaluated	Accuracy measure				
		PCC	Specificity	Sensitivity	Kappa	AUC
L. oregana	DES on EVAL	76.7	87.3	47.5	0.368	0.754
	PURP on EVAL	63.2	76.8	26.2	0.032	0.567
	z-Score	3.646	3.383	5.083	4.600	3.652
	P	<0.001	<0.001	<0.001	<0.001	<0.001
L. pulmonaria	DES on EVAL	82.0	92.2	55.4	0.514	0.832
	PURP on EVAL	50.0	46.1	60.2	0.048	0.411
	z-Score	8.790	14.106	−1.192	6.698	9.130
	P	<0.001	<0.001	0.233	<0.001	<0.001
P. anomala	DES on EVAL	85.0	92.5	54.2	0.496	0.797
	PURP on EVAL	56.0	63.1	27.1	−0.076	0.443
	z-Score	8.214	9.263	7.030	4.712	6.475
	P	<0.001	<0.001	<0.001	<0.001	<0.001
P. anthraspis	DES on EVAL	81.7	91.1	46.9	0.410	0.781
	PURP on EVAL	67.7	80.1	21.9	0.020	0.517
	z-Score	3.996	3.885	6.681	6.361	4.832
	P	<0.001	<0.001	<0.001	<0.001	<0.001

DES on EVAL, DESIGN on EVALUATION; PURP on EVAL, PURPOSIVE on EVALUATION, positive z-scores indicate higher accuracy rates for models based on the DESIGN sampling survey; negative indicates lower rates.

at the expense of the others. From a conservation perspective, where maximum prediction is often the goal, ecologists will often sacrifice realism for model generality and usability.

Model generality and usability is assumed represented by many different types of metrics (see Fielding and Bell, 1997), of which resubstitution accuracy is one of the more common. In our study, resubstitution accuracies were approximately the same for the models built on the DESIGN and PURPOSIVE surveys, leading to the possible conclusion that both survey forms led to models of similar utility. However, when evaluated on the EVALUATION data, predictive accuracies were quite different, with the DESIGN outperforming the PURPOSIVE surveys.

Table 5 – Proportional tests comparing DESIGN and PURPOSIVE resubsitution accuracy rates vs. EVALUATION accuracy rates for classification tree models built for four species of lichens in the Pacific Northwest, USA

Species	Accuracy rate	DESIGN		PURPOSIVE	
		z-Score	P	z-Score	P
L. oregana	PCC	4.447	<0.001	7.377	<0.001
	Specificity	4.052	<0.001	7.454	<0.001
	Sensitivity	5.024	<0.001	7.494	<0.001
	Kappa	3.971	<0.001	5.603	<0.001
	AUC	4.066	<0.001	3.657	<0.001
L. pulmonaria	PCC	2.414	0.016	8.794	<0.001
	Specificity	0.793	0.427	8.812	<0.001
	Sensitivity	4.173	<0.001	7.333	<0.001
	Kappa	2.084	0.037	7.631	<0.001
	AUC	1.489	0.137	6.640	<0.001
P. anomala	PCC	2.213	0.027	8.095	<0.001
	Specificity	2.167	0.030	10.314	<0.001
	Sensitivity	2.708	0.007	6.761	<0.001
	Kappa	1.944	0.051	6.805	<0.001
	AUC	1.435	0.151	6.446	<0.001
P. anthraspis	PCC	3.854	0.002	4.754	<0.001
	Specificity	3.635	<0.001	5.318	<0.001
	Sensitivity	5.300	<0.001	9.497	<0.001
	Kappa	2.217	0.027	5.670	<0.001
	AUC	1.258	0.208	4.959	<0.001

Data for tests from Tables 2 and 4.

Table 6 – Proportional tests comparing DESIGN and PURPOSIVE cross-validation accuracy rates vs. EVALUATION accuracy rates for classification tree models built for four species of lichens in the Pacific Northwest, USA

Species	Accuracy rate	DESIGN		PURPOSIVE	
		z-Score	P	z-Score	P
L. oregana	PCC	1.068	0.285	−5.321	<0.001
	Specificity	1.002	0.316	−5.598	<0.001
	Sensitivity	−0.536	0.591	−3.942	<0.001
	Kappa	0.000	1.000	3.300	0.009
	AUC	−0.488	0.625	0.607	0.543
L. pulmonaria	PCC	−0.308	0.757	−1.703	0.088
	Specificity	−0.972	0.331	−4.707	<0.001
	Sensitivity	−1.374	0.169	3.710	0.002
	Kappa	−1.245	0.213	−0.437	0.662
	AUC	−1.228	0.219	1.772	0.076
P. anomala	PCC	−0.942	0.345	−4.373	<0.001
	Specificity	−0.501	0.616	−7.614	<0.001
	Sensitivity	−3.619	<0.001	−0.528	0.597
	Kappa	−1.205	0.228	2.902	0.004
	AUC	−0.708	0.435	1.695	0.089
P. anthraspis	PCC	1.535	0.128	0.691	0.489
	Specificity	1.522	0.128	−0.263	0.792
	Sensitivity	−2.822	0.004	3.118	0.002
	Kappa	−0.578	0.562	−0.133	0.894
	AUC	0.296	0.762	−1.233	0.217

Data for tests from Tables 2 and 4.

Clearly, the resubstitution accuracy estimates reported here do not represent how poorly the models fit on the PURPOSIVE data perform compared to the models fit on the DESIGN data. Moreover, the selected predictors from the classification trees differed considerably between the DESIGN and PURPOSIVE surveys, leading to different ecological interpretations as well. These differences in both model predictive generality and ecological interpretation should make modellers more cautious in their interpretation than is currently observed in the literature.

How representative our results are of other work is unclear given the paucity of studies that have had the opportunity to compare results using both resubstitution (training) and independent test data sets. Fewer studies yet have evaluated the differences in final model structures due to the underlying probabilistic or non-probabilistic sampling designs that led to collection of the training data. It is possible that some of the differences between the classification trees fit on the DESIGN and PURPOSIVE data sets could be due to sampling error, or to some collinearity that remains among the predictor variables even given our attempts to reduce collinearity. We note, however, that variables strongly associated with the presence (or absence) of lichen species consistently appear in the classification trees, a result that would not be apparent if either sampling error or collinearity were primarily determining the classification tree structures. This suggests that some portion of the differences in model structure is due to the underlying sample survey form. While these differences in tree structures may have minimal impact on purely predictive studies, they do have impact on the ecological interpretations behind any model-derived distribution pattern (Austin et al., 2006).

Even differences in the definition of what constitutes an "independent" data set confound attempts to fully understand effects of sampling design on model accuracy. Muñoz and Felicísimo (2004), for example, defined as "independent" the simple splitting of their original data into two subsets, one for training and one for "independent" validation. Not surprisingly, differences in their reported prediction accuracies between the 10-fold cross-validation and the 70:30 splitting of their data for validation were minimal. This is because any biases in the collection of the original data will naturally carry through, no matter what type of randomization is used to split the data, thereby precluding consideration of the test data as truly independent. Such splitting does not, in our opinion, constitute a truly "independent" evaluation, especially if the underlying sample survey was non-probabilistic.

One seeming pattern is an observed decline in accuracies from resubstitution to internal (e.g., cross-validation) and external (i.e., independent) validation exercises. Fielding and Haworth (1995) evaluated the generality of bird-habitat models for three bird species in Scotland and documented lower test than resubstitution accuracies. Other studies that performed some type of internal validation of the training data, whether jackknife or cross-validation, all show some decline in resubstitution accuracies during the validation process (see Frederick and Gutiérrez, 1992; Martin and Morrison, 1999; Leathwick et al., 2006). We believe that for many types of classification tools, including the classification trees used here, resubstitution accuracy estimates alone are generally poor estimators of true prediction accuracy rates. We further suggest that resubstitution accuracy rates should not be published in isolation if the ultimate intent of the model is

prediction, as they can lead to false impressions about model usability.

Instead, we believe greater emphasis should be placed on estimating accuracies through cross-validation techniques (Fielding and Bell, 1997, Table 1) when it is difficult to perform a truly independent evaluation (see Chatfield, 1995). For the models fit on the DESIGN data, the 10-fold cross-validation estimates of model accuracy were very good estimators of the true predictive accuracies for the EVALUATION data. In the two instances in which the cross-validated accuracies were significantly different from the predictive accuracies for the EVALUATION data, the cross-validation accuracy estimates underestimated the true accuracies of the models. Similar results comparing jackknifed accuracies against an independent data set were reported by Call et al. (1992). Accuracies were similar between the jackknifed and independent data, and both were lower than the resubstitution accuracies.

From a statistical perspective, non-probability sampling efforts like purposive sampling can lead to bias in estimates and uncertainty in the confidence about the estimate. This does not imply that a purposive sample never represents the population of interest. It does, however, cast doubt on how well the sample represents the population, a concern when statistical models are used to address ecological questions like species distributions or community structuring. In one such example, Kodric-Brown and Brown (1993) replicated a study by Glover (1989) on fish species in Australian desert springs, correcting sampling biases due to incomplete sampling that were acknowledged by Glover as inherent in the first study. This correction led to completely different patterns of species distribution and community structuring. Similarly, Austin and Heyligers (1989) noted that failure to adequately sample across the range of environmental variables influencing the distribution of plant communities can lead to erroneous conclusions about species distribution patterns. In either circumstance, attention to designs that minimize non-probabilistic biases is paramount, yet it is unclear if ecologists fully appreciate the magnitude of the biases that can arise when randomization as an element of design is ignored and samples are collected in a non-probabilistic manner.

Acknowledgments

The authors acknowledge the contributions of the following people in field collection, identification and data entry: S. Berryman, M. Boyll, C. Derr, A. Ingersoll, D. Glavich, K. Gossen, A. Mikulin, J. Riley, and R. Ulrich. Special thanks to P. Halonen and B. Ryan (deceased) for help with difficult identifications, and to the many others who assisted with field collections. Field work was funded by the PNW Region Air Program and the Forest Service Survey and Manage Program. We are also grateful for the assistance from a large number of individuals working on the Survey and Manage Program in the Pacific Northwest Forest Plan. Specifically, we thank T. Brumley, N. Diaz, B. Rittenhouse, and N. Middlebrook for guidance on Forest Service and Bureau of Land Management Survey and Manage Issues. Three anonymous reviewers, and T.B. Murphy, helped improve the manuscript with their insights. Funding was provided by the State of Oregon Bureau of Land Management through a cooperative research arrangement with the USGS Forest and Rangeland Ecosystem Science Center (FRESC), Corvallis, Oregon, and the USGS Utah Cooperative Fish and Wildlife Research Unit, Utah State University. The ideas behind this manuscript were developed during the "2004 Generalized Regression Analyses And Spatial Predictions: Grasping Ecological Patterns From Species To Landscape" workshop held in Riederalp, Switzerland. The ideas also benefitted from the liberal involvement of several excellent bottles of cognac.

REFERENCES

Austin, M.P., Cunningham, R.B., Good, R.B., 1983. Altitudinal distribution in relation to other environmental factors of several eucalypt species in southern New South Wales. Aust. J. Ecol. 8, 169–180.

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niche of five eucalyptus species. Ecol. Monogr. 60, 161–177.

Austin, M.P., Heyligers, P.C., 1989. Vegetation survey design for conservation: Gradsect sampling of forests in north-eastern New South Wales. Biol. Conserv. 50, 13–32.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D., Luoto, M., 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory, this issue.

Best, L.G., Stauffer, D.F., 1986. Factors confounding evaluation of bird-habitat relationships. In: Verner, J., Morrison, M.L., Ralph, C.J. (Eds.), Wildlife 2000: Modeling Habitat Relationships of Terrestrial Vertebrates. University of Wisconsin Press, Madison, Wisconsin, USA, pp. 209–216.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks/Cole. Monterey, California, USA.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Call, D.R., Gutiérrez, R.J., Verner, J., 1992. Foraging habits and home-range characteristics of California spotted owls in the Sierra Nevada. Condor 94, 880–888.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. J. Roy. Stat. Soc. A. Stat. 158, 419–466.

Cochran, W.G., 1977. Sampling Techniques, 3rd ed. John Wiley & Sons, New York, USA.

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

Dreisbach, T.A., Smith, J.E., Molina, R., 2002. Challenges of modelling fungal habitat: when and where do you find chanterelles? In: Scott, J.M., Heglund, P., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), Predicting Species Occurrence: Issues of Accuracy and Scale. Island Press, Covello, California, USA, pp. 475–481.

Edwards Jr., T.C., Deshler, E., Foster, D., Moisen, G.G., 1996. Adequacy of wildlife habitat relation models for estimating spatial distributions of terrestrial vertebrates. Conserv. Biol. 10, 263–270.

Edwards Jr., T.C., Cutler, D.R., Geiser, L., Alegria, J., McKenzie, D., 2004. Assessing rarity and seral stage association of species with low detectability: lichens in western Oregon and Washington forests. Ecol. Appl. 14, 414–424.

Edwards Jr., T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Alegria, J., 2005. Use of model-assisted designs for sampling rare ecological events. Ecology 86, 1081–1090.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and

endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.
Fielding, A.H., Haworth, P.F., 1995. Testing the generality of bird-habitat models. Conserv. Biol. 9, 1466–1481.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.
Frederick, G.P., Gutiérrez, R.J., 1992. Habitat use and population characteristics of the white-tailed ptarmigan in the Sierra Nevada, California. Condor 94, 889–902.
Frescino, T.S., Edwards Jr., T.C., Moisen, G.G., 2001. Modelling spatially explicit forest structural variables using generalized additive models. J. Veg. Sci. 12, 15–26.
Geiser, L., 2004. Manual for monitoring air quality using lichens on national forests of the Pacific Northwest. USDA Forest Service, Pacific Northwest Region, Technical Paper R6-NR-AQ-TP-1-04. http://www.fs.fed.us/r6/aq.
Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler, D., 2002. The national elevation dataset. Photogramm. Eng. Rem. S. 68, 5–12.
Glover, C.J.M., 1989. Fishes. In: Zeidler, W., Ponder, W.F. (Eds.), Natural History of Dalhousie Springs. Southern Australian Museum, Adelaide, Australia, pp. 89–111.
Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 143, 29–36.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall/CRC, New York, USA.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. John Wiley & Sons, New York, USA.
Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314.
Jaberg, C., Guisan, A., 2001. Modelling the influence of landscape structure on bat species distribution and community composition in the Swiss Jura Mountains. J. Appl. Ecol. 38, 1169–1181.
Kodric-Brown, A., Brown, J.H., 1993. Incomplete data sets in community ecology and biogeography: a cautionary tale. Ecol. Appl. 3, 736–742.
Lawler, J.J., Edwards Jr., T.C., 2002. Landscape patterns as predictors of nesting habitat: a test using four species of cavity-nesting birds. Landsc. Ecol. 17, 233–245.
Leathwick, J.R., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions, this issue.
Levins, R., 1966. The strategy of model building in population biology. Am. Sci. 54, 421–431.
Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall, New York, USA.
Martin, J.A., Morrison, M.L., 1999. Distribution, abundance, and habitat characteristics of the buff-breasted flycatcher in Arizona. Condor 101, 272–281.
Max, T.A., Schreuder, H.T., Hazard, J.W., Teply, J., Alegria, J., 1996. The Region 6 vegetation inventory and monitoring System. General Technical Report PNW-RP-493, USDA Forest Service, Pacific Northwest Research Station, Portland, Oregon, USA.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall/CRC, New York, USA.
McNoleg, O., 1996. The integration of GIS, remote sensing, expert systems and adaptive co-kriging for environmental habitat modeling of the Highland Haggis using object-oriented, fuzzy-logic and neural network techniques. Comput. Geosci. 22, 585–588.
Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., Edwards, T.C., Jr., 2006. Predicting tree species presence and basal area in Utah: a comparison of generalized additive models, stochastic gradient boosting, and tree-based methods, this issue.
Muñoz, J., Felicísimo, Á.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15, 285–292.
Nusser, S.M., Breidt, F.J., Fuller, W.A., 1998. Design and estimation for investigating the dynamics of natural resources. Ecol. Appl. 8, 234–245.
Olsen, A.R., Schreuder, H.T., 1997. Perspectives on large-scale natural resource surveys when cause-effect is a potential issue. Environ. Ecol. Stat. 4, 167–180.
Olsen, A.R., Sedransk, J., Edwards, D., Gotway, C.A., Liggett, W., Rathbun, S., Reckhow, K.H., Young, L.J., 1999. Statistical issues for monitoring ecological and natural resources in the United States. Environ. Monit. Assess. 54, 1–45.
Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B. (Eds.), 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covello, California, USA.
Steele, B.M., 2000. Combining multiple classifiers: an application using spatial and remotely sensed information for land cover type mapping. Remote Sens. Environ. 74, 545–556.
Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190, 214–251.
Thornton, P.E., Running, S.W., 1999. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agri. For. Meteorol. 93, 211–228.
Van Horne, B., Wiens, J.A., 1991. Forest Bird Habitat Suitability Models and the Development of General Habitat Models. Research 8. US Fish and Wildlife Service, US Department of the Interior, Washington, DC, USA.
Welch, N.E., MacMahon, J.A., 2005. Identifying habitat variables important to the rare Columbia spotted frog in Utah (USA): an information-theoretic approach. Conserv. Biol. 19, 473–481.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602.
Zimmermann, N.E., Kienast, F., 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. J. Veg. Sci. 10, 469–482.