Tree Genetics & Genomes (2011) 7:399–408
DOI 10.1007/s11295-010-0341-7

ORIGINAL PAPER

Developing seed zones and transfer guidelines with multivariate regression trees

Andreas Hamann · Tim Gylander · Pei-yu Chen

Received: 22 March 2010 / Revised: 14 August 2010 / Accepted: 1 September 2010 / Published online: 12 October 2010

Abstract Managing seed movement is an important component of forest resource management to minimize maladaptation of planting stock in forest plantations. Here, we describe a new approach to analyze geographic patterns of adaptive and neutral genetic variation in forest trees and to link this genetic information to geographic variables for the delineation of seed zones and the development of seed transfer guidelines. We apply multivariate regression trees to partition genetic variation, using a set of environmental or geographic predictor variables as partitioning criteria in a series of dichotomous splits of the genetic dataset. The method can be applied to any type of genetic data (growth, adaptive, or marker traits) and can simultaneously evaluate multiple traits observed over several environments. The predictor variables can be categorical (e.g., ecosystem of seed source), continuous (e.g., geographic or climate variables), or a combination of both. Different sets of predictor variables can be used for different purposes: In two case studies for aspen and red alder, we show (1) how latitude, longitude, and elevation of seed sources in a provenance trial can be used to develop simple seed transfer guidelines; (2) how ecosystem classes and elevation as predictor variables can be used to delineate seed zones and breeding regions; and (3) how climate variables as predictors can reveal adaptation of genotypes to the environments in which they occur. Partitioning of genetic variation appears very robust regarding the choice of predictor variables, and we find that the method is a powerful aid for interpreting complex genetic datasets.

Keywords Genetic diversity · Tree improvement · Ecological genetics · Genecology · Aspen · Red alder

Introduction

Seed zones and seed transfer guidelines are essential tools in reforestation to ensure that seedlings are well adapted to the growing conditions of the planting site (e.g., Morgenstern 1996; Ying and Yanchuk 2006). Using planting stock for reforestation that originates within a restricted geographic area delineated as a seed zone aims at minimizing loss of productivity and forest health issues due to maladaptation. Alternatively, movement of seed can be restricted with seed transfer guidelines, also sometimes referred to as floating or flexible seed zones (Ying and Yanchuk 2006). Transfer guidelines avoid drawing fixed boundaries across continuous genetic clines by specifying a maximum distance and elevation movement from source location to a planting site to avoid maladaptation (e.g., Rehfeldt 1988, 1989).

Generally, there are two conceptual approaches to develop seed zones and seed transfer guidelines. The first aims at *maximizing tree growth* by comparing response functions of different genotypes over multiple test environments. The approach usually employs univariate or multivariate curve fitting techniques to analyze growth and adaptive traits as a function of environmental or geographic predictor variables (e.g., Lindgren and Ying 2000; Raymond and Lindgren 1990; Roberds et al. 1990; Wang et al. 2006b). The second approach aims at *minimizing risk*

Communicated by R. Burdon

Electronic supplementary material The online version of this article (doi:10.1007/s11295-010-0341-7) contains supplementary material, which is available to authorized users.

A. Hamann (✉) · T. Gylander · P.-y. Chen
Department of Renewable Resources, University of Alberta,
739 General Services Building,
Edmonton, AB T6G 2H1, Canada
e-mail: andreas.hamann@ualberta.ca

based on the assumption that local sources are optimally adapted to the environments in which they occur. Numerous techniques have been developed to match seed sources from various geographic locations to target environments, and these techniques use the environmental conditions of the seed source as matching criteria (e.g., Campbell 1979, 1986; Lesser and Parker 2006; Parker 1992; Rehfeldt 1988, 1989).

Both approaches pose major practical and research challenges. The response-function approach requires expensive series of test plantations over multiple environments, and it is generally difficult to fit mathematical response functions to traits that are influenced by multiple environmental variables, plus additional confounding factors related to test-site conditions (e.g., O'Neill et al. 2007). Further, it is often unclear how adaptive trait measurements correspond to fitness (e.g., we do not know what the optimal timing of leafout and budset is), and as a consequence, important adaptive traits cannot be taken into account in the same way as growth traits. Perhaps the largest drawback for the second, risk-avoidance approach is that the assumption of local optimality is not always valid (Mangold and Libby 1978; Mátyás 1990; Namkoong 1969) and that the approach will further lose relevance under observed and projected global climate change (Marris 2009; McKenney et al. 2009).

It is also difficult to translate genetic information into geographic zones or transfer guidelines—a second step that is required for both the risk-avoidance and growth-optimization approaches. Geographic information system (GIS)-based techniques have been developed to delineate seed-zone boundaries where response functions of differently adapted genotypes intersect or drop below a certain threshold (Hamann et al. 2000). For the risk-avoidance strategy, GIS-based seed-zone optimization techniques are available that assign groups of similarly adapted genotypes to their corresponding environments (O'Neill and Aitken 2004; Parker 1992). However, complications arise in this second step when genetic information for multiple growth and adaptive traits has to be integrated and then translated to practical seed zones and transfer guidelines that further have to account for administrative and operational planning realities of reforestation programs (Ying and Yanchuk 2006). To our knowledge, virtually all seed-zone systems in practical use have been developed by evaluating available genetic information and then subjectively deciding on reasonable transfer guidelines or seed-zone delineations that usually track ecological regions.

This practice has generally resulted in sensible guidelines that have been successfully used for many decades, and the methodology that we propose in this study is not meant to replace sound judgment. Using two case studies, we illustrate how multivariate regression tree analysis can be used to support subjective decisions by integrating and visualizing complex genetic information. The technique can be applied for both risk-avoidance and growth-optimization strategies and executes the genetic analysis and the geographic mapping of seed zones in one step. By using climatic variables or potentially other relevant environmental factors instead of geographic variables, the technique can further help us to understand how genotypes are adapted to their source environments. This is relevant not only for delineating seed zones under current environments but also for developing climate change adaptation strategies.

In two case studies for aspen (*Populus tremuloides* Michaux.) and red alder (*Alnus rubra* Bong.), we show (1) how latitude, longitude, and elevation of seed sources in a provenance trial can be used to develop simple seed transfer guidelines with multivariate regression trees; (2) how ecosystem classes and elevation as clustering criteria can be used to delineate seed zones and breeding regions; and (3) how climate variables as predictors can reveal adaptation of genotypes to their source environments.

Methods

Genetic trials and measurements

The aspen case study is based on a provenance trial series established by an industry cooperative in 1998 at five locations in western Canada and includes seed collections from British Columbia to Minnesota. At each test site, provenances were planted in a randomized complete block design with six replications in five-tree row plots. Trees were evaluated in 2006 for height after eight growing seasons in the field. For this analysis, we use means of 42 provenances in five-tree row plots as basic experimental units. Means of row plots were summarized at the provenance level as least-squares means, taking advantage of the blocked experimental design. A normalized version of these data is included as a sample dataset in the Electronic Supplement 1.

For the red alder case study, we re-analyzed previously published data from a trial series planted in 1994 by the BC Ministry of Forests with 55 provenances. To illustrate how multivariate regression tree analysis can handle various types of genetic information, we analyzed data from only a single site of this series, but we included multiple quantitative traits as well as genetic marker traits (Hamann et al. 1998, 2000). Measurements include provenance means for 4-year height and survival, the timing of leaf abscission in fall of 1998, the timing of bud break in 1999 (average Julian day calculated from observations at weekly intervals), and the extent of wilting of leaves during an exceptional drought period in July/August of 1999 (scored on a scale of 0 = no wilting to 5 = heavy wilting). Because of spatial heterogeneity of the test sites, least-squares means for each provenance were calculated after removal of spatial autocorrelations (Hamann et al. 2002). We further evaluate allele frequencies from six polymorphic allozyme loci for a subset of 19 provenances, using the frequency of the most frequent allele at each locus for analysis (Table 1 in Hamann et al. 1998).

Climatic and spatial datasets

To characterize the long-term climate conditions at the source locations of provenances, we used interpolated climate data for the 1961–1990 normal period that was generated with the Parameter Regression of Independent Slopes Model (PRISM; Daly et al. 2008). We enhanced this spatial database for use with sample data by applying lapse-rate based elevation adjustment functions to all climate variables that are based on temperature measurements (e.g., degree days, heat–moisture indices, or mean annual temperature). This improves climate variable estimates by accounting for the difference between the recorded elevation of the sample and the elevation that was used to estimate climate variable values with PRISM for the sample location (Hamann and Wang 2005; Wang et al. 2006a). Estimation of biologically relevant climate variables, lapse-rate elevation adjustments, and data extraction from grids for sample locations were carried out with a custom software application that we make freely available (Mbogga et al. 2009).[1]

Ecosystem delineations of western Canada served as candidate seed zones for multivariate regression tree analysis. We use "Ecoprovinces" of the National Ecological Framework for Canada (Selby and Santry 1996) and for the USA we use the corresponding "Level 3" delineation of the United States Ecoregion System (EPA 2007). Note that we chose a coarse ecosystem delineation for illustrative reasons. The Natural Regions System of Alberta (NRC 2006) and the Biogeoclimatic Ecosystem Classification system for British Columbia (Meidinger and Pojar 1991) would also be appropriate for analysis but too detailed for display at the map scale we use in this paper. Species distribution maps were added for illustration, using the maps of Hamann et al. (2005) for red alder and Little's (1971) maps for aspen, intersected with remotely sensed deciduous forest land cover (Wulder et al. 2008).

Statistical analysis

Data for all traits, including allozyme frequencies, were standardized so that they are weighted equally in multivariate regression tree analysis. We subtracted the mean and divided by the standard deviation of each trait at each test site, so that all traits are expressed in units of standard deviations from a site mean of zero.

Multivariate regression tree analysis was implemented with the *MVpart* package v1.2–6 for the R programming environment (R Development Core Team 2008). Multivariate regression trees (MRT) are based on the same principles as Classification and Regression Trees (CART) but extended to more than one response variable (De'Ath 2002). MRT can be viewed as a constrained clustering methodology that is suitable for explanation as well as prediction. A set of clusters is grown by repeated binary splits of the genetic dataset. Splits are made using environmental predictor variables as criteria, so that the homogeneity of genetic response variables is maximized. Homogeneity is evaluated as sums of squares of traits around the multivariate mean of observations in a cluster (De'Ath 2002). As in CART, no assumptions are made about the mathematical nature of the relationship between response and predictor variables. For our application, it means that the reaction norms of genotypes over multivariate environmental gradients may take any form. Sample R code for this analysis is included in the Electronic Supplement 2.

Although not used in this study, we provide additional R code for missing value imputation in the genetic datasets (Electronic Supplement 3). Missing values are a common occurrence in provenance trial data (e.g., due to insufficient planting stock for a test series). Since the *MVpart* function requires a complete dataset, a bootstrapped regression tree analysis is first used to impute missing values based on similarity with other rows (genotypes) in the dataset. We use the bootstrap implementation *rfImpute* of the *randomForest* package v 4.5 for the R programming environment (Breiman 2001). Obviously, this procedure should not be used to fill a large number of missing values in unbalanced experimental designs. Similarly, several missing values in a data table row will degrade the quality of imputations, which have then to rely on a smaller number of remaining trait values to represent a genotype. In this case, samples should be removed from the analysis as in the second case study, where we excluded genotypes that were not screened for allozyme frequencies.

[1] Available for download at http://www.ualberta.ca/~ahamann/climate.html. An alternate download site is http://www.genetics.forestry.ubc.ca/cfcg/climate-models.html.

Results

Reaction norms over multiple test sites

Using the geographic variables latitude, longitude, and elevation as predictors, provenances of the reciprocal transplant experiment with aspen are split into four groups, which explains 52% of the genetic variation in the dataset consisting of height measurements at five test sites (Fig. 1a, b). The first two longitudinal splits separate the sources from northern British Columbia and Minnesota, and the third latitudinal split separates northern provenances in Alberta from more southern provenances in Saskachewan and Alberta. The bar charts in the dendrogram represent group means, which can be interpreted as average genetic reaction norms of similarly adapted genotypes. The group from northern BC (Fig. 1a, b, purple) performs below average at all test sites with the relatively poorest performance at the most southern test site (33). The five provenances from Minnesota (blue) perform above average at the three southern test sites and approximately average at the test sites in northern Alberta and British Columbia. The group of five provenances from northern Alberta (green) performs slightly below average at the southern test sites but is the best performer in northern British Columbia. The remaining provenances from central Alberta and Saskachewan (orange) are average performers at all test sites.

Using ecological regions and elevation as predictor variables (Fig. 1c, d), we obtain the same results as above with four groups. Adding another group explains a small amount of additional genetic variation (4%) and separates the Rocky Mountain Foothill provenances, which are weakly distinguished from Boreal Plains provenances by lower performance at several locations, including their local test site (33). Notably, elevation is not useful to further separate provenances within ecological regions, and the Alberta and Saskatchewan populations of the Boreal Plains cannot be genetically distinguished.

Multiple traits observed in a single environment

Red alder populations from coastal British Columbia show a latitudinal cline of genetic differentiation when analyzed with geographic variables as predictors (Fig. 2a, b). Multivariate regression tree analysis separates populations along three roughly equally spaced latitudinal divisions that explain 42% of the genetic variation in growth and adaptive traits measured at a southern test site. At this test site, northern provenances have reduced height and lower survival rates than local sources. They break and set bud earlier and show a higher degree of leaf wilting under drought conditions. The local sources (Fig. 2a, b, light pink) are distinguished by the latest dates of budbreak and leaf abscission and the least wilting under drought.

Using ecological regions and elevation as predictor variables (Fig. 2c, d), we obtain similar results as above with four groups. Notably, provenances from the Queen Charlotte Islands remain in the same group as the Mid Coast provenances of the mainland. In southern British Columbia, we find a new grouping that separates populations in the Georgia Depression (South Coast and Vancouver Island East) from more western sources. The Vancouver Island West group can be further divided into low-elevation sources that are typically located right on the west coast and more inland sources above 165 m elevation. The coastal, low-elevation group is distinguished by very late dates of budbreak. The new Georgia Depression group is distinguished by another adaptive trait, showing the least amount of wilting under drought conditions.

The third example for regression tree analysis with this dataset adds genetic marker information, which substantially changes the group partitioning (Fig. 2c, e). Now, most of the genetic variation is accounted for by separating the island from mainland sources. Provenances from Queen Charlotte Island and Vancouver Island generally show above-average values for the most frequent allele at all loci, indicating below-average heterozygosity. Further, one locus for aspartate aminotransferase is fixed for alternate alleles in the island and mainland populations (Fig. 2e, solid red bars). Subsequent splits of the genetic dataset are based on quantitative traits and approximately conform to the partitioning described in the previous example. Growth and adaptive traits primarily separate northern and southern sources, with southern sources generally showing better performance, later budbreak, and delayed leaf abscission. However, since this analysis was performed on a reduced set of 19 instead of 55 provenances (Fig. 2c, red dots), the results are not identical.

Adaptation to climatic source environments

Instead of geographic variables in the form of ecosystem delineations or geographic coordinates, we repeat the multivariate regression tree analysis with climatic variables at provenance sample locations as criteria for partitioning genetic variation. Aspen provenances are split in a very similar way: Minnesota sources (n=5) are characterized by the highest mean annual temperature (Fig. 3). Because of multicollinearity among climate variables, we also report alternative predictors that could be used for the same partitioning. For this split, one alternative variable, mean warmest month temperature, was identified by the regression tree analysis (Table 1). The second node of the regression tree separates the Boreal Plains provenances (n=32) from more northern sources. This partitioning can be based on mean annual temperature or chilling degree days. The last division separates the two northern groups

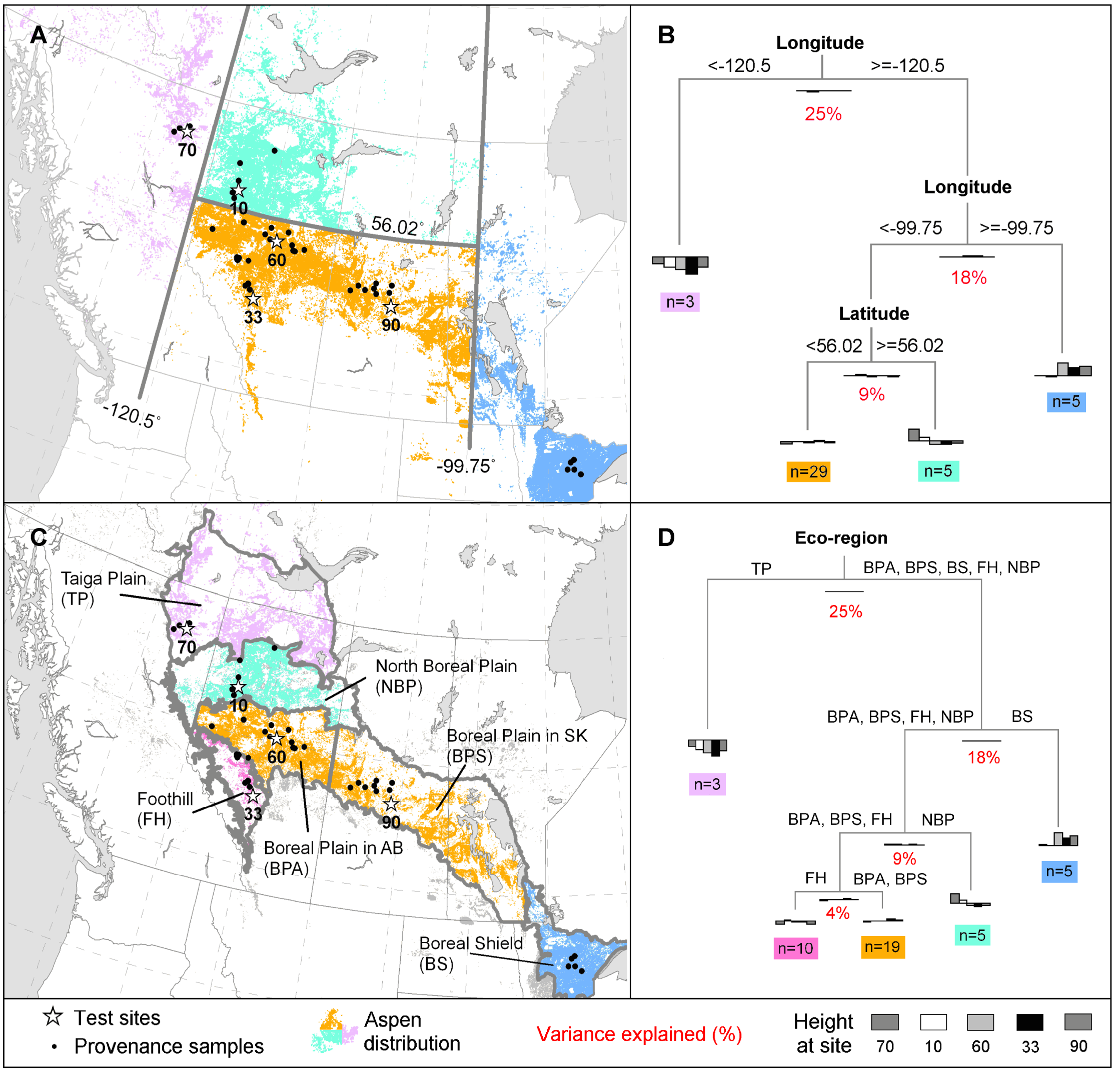

Fig. 1 Multivariate regression tree analysis of 8-year height of aspen provenances at five test sites. Candidate predictor variables are latitude, longitude, and elevation (**a**, **b**) and ecological region and elevation (**c**, **d**)

from Alberta and British Columbia based on differences in summer precipitation or winter temperature.

The regression tree for alder first roughly separates the southern provenances, which are associated with a long growing season (Fig. 4). Alternative variables that could explain this division are mean warmest month temperature and the date where the growing degree days reach 100, an indicator for the start of the growing season (Table 2). The southern group is then subdivided by separating three sources from the Vancouver Island west coast based on a low annual heat moisture index (indicating moist conditions). These provenances are characterized by a high wilting score at the relatively dry test location. Subsequently, another four provenances from the driest areas of the Georgia Depression are separated, which are characterized by the lowest wilting score. Notably, climate variables were not able to separate provenances from the Queen Charlotte Island, Mid Coast, and North Coast regions, which potentially explains 5% to 6% of the total variation in the genetic dataset (Fig. 2b, d).

Discussion

Interpreting dendrograms

First, we should point out that regression tree partitioning should not be taken as optimal seed-zone delineations at face value. The groupings not only reflect genetic differentiation but they are also influenced by the provenance sampling design and the nature of geographic variables. For example, longitudinal divisions in Fig. 1a do not necessar-

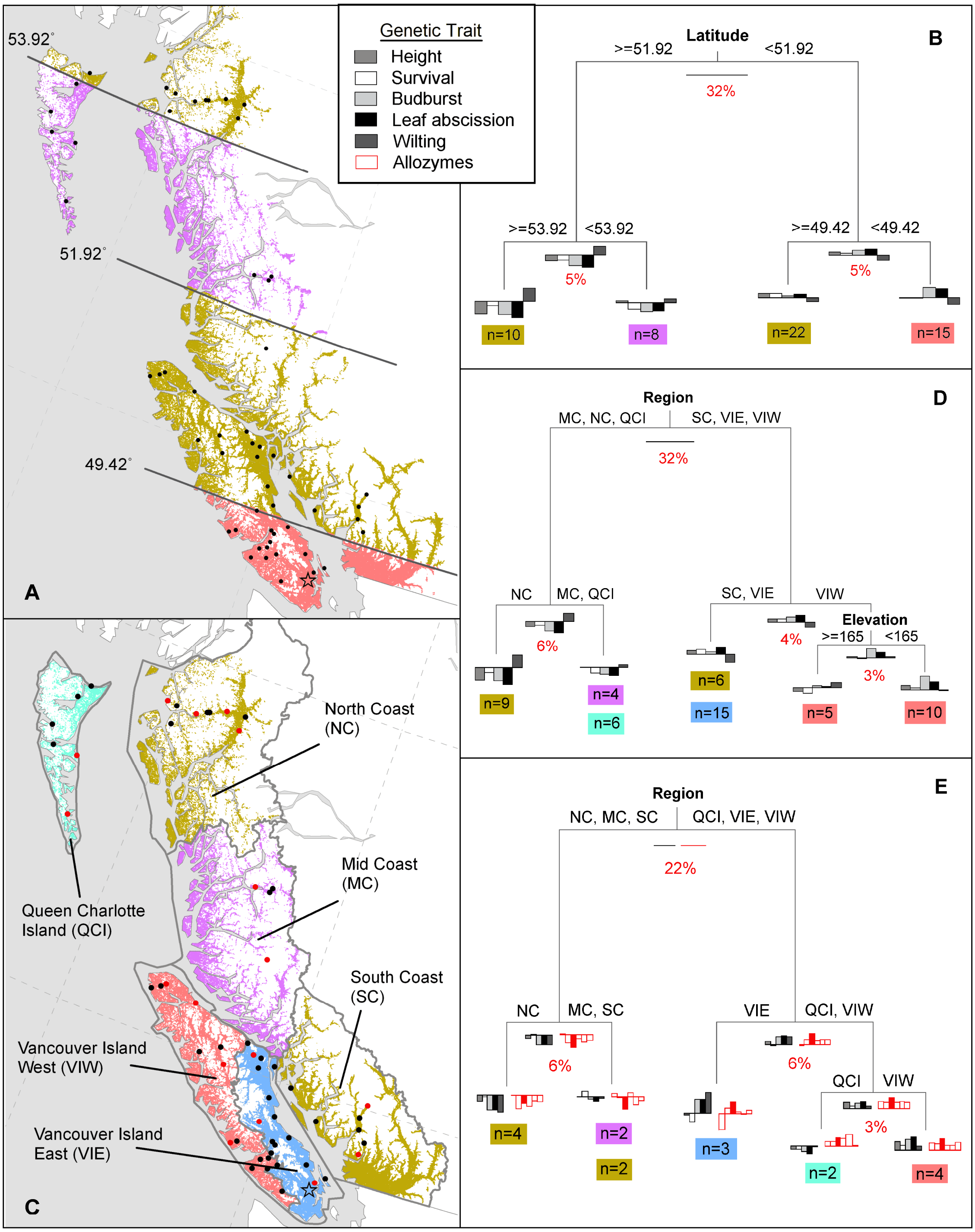

Fig. 2 Multivariate regression tree analysis of quantitative traits of red alder observed at a single test site (*star*). Candidate predictor variables are latitude, longitude, and elevation (**a**, **b**) and ecological region plus elevation (**c**–**e**). Allozyme frequencies for 19 sample locations (**c**, *red dots*) are included in the last dendrogram (**e**, *red bar charts*), where a fixed allele is highlighted in *solid red*

ily imply a longitudinal genetic cline. Since the sampling design is roughly diagonal from northwest to southeast, a latitudinal division is equally possible. As an example of how the nature of geographic variables influences partitioning, note that the Foothills group of the aspen provenance trial (Fig. 1c) could not be separated based on latitude and longitude because it would have required a diagonally oriented (southwest to northeast) predictor variable.

Another obvious limitation is that genetic differentiation cannot be detected in regions that have never been sampled.

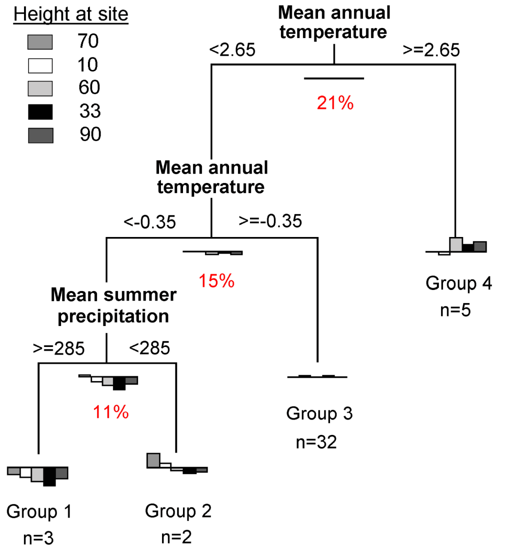

Fig. 3 Multivariate regression tree analysis of 8-year height of aspen provenances at five test sites. Candidate predictor variables are 15 climate variables and the amount of variance explained by the groupings of genotypes are indicated in *red*. Alternative climate variables are listed in Table 1

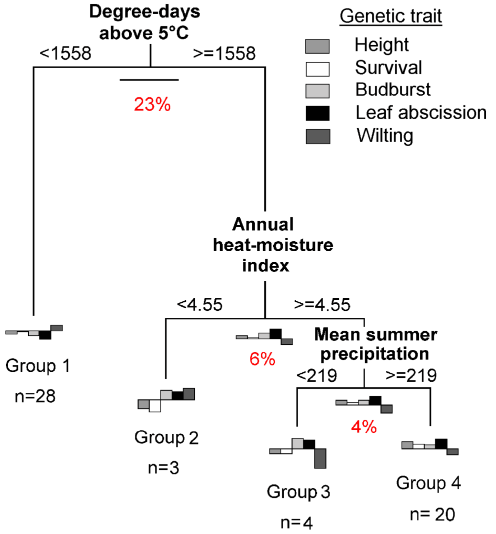

Fig. 4 Multivariate regression tree analysis of growth and adaptive traits of 55 red alder provenances observed at a single test site. Candidate predictor variables are 15 climate variables and the amount of variance explained by the groupings of genotypes are indicated in *red*. Alternative climate variables are listed in Table 2

In Fig. 1a, Minnesota sources are separated by the latitudinal mid-point between Saskachewan and Minnesota samples, an arbitrary division in a region that lacks sample coverage. However, the two case studies we present here are not unusual for the level of information that researchers and resource managers have available from provenance tests. A widely used means to overcome the limitation of very restricted geographic sampling is to partition species habitat into ecological zones that represent reasonably homogeneous environments as the basis of seed zones (Campbell 1991; Ying and Yanchuk 2006). Essentially, ecological proxy data (plant community composition and physiogeographic variables) are used to extrapolate where genotypes similar to a few tested samples may be found.

These limitations are not unique to multivariate regression tree analysis and need to be considered in developing seed zones and seed transfer guidelines using any technique. Ultimately, the regression tree clusters are driven by the degree of genetic differentiation observed in the available samples. If there is none, provenances will not be separated by any candidate partitioning criteria: geographic, ecosystem, or climate variables (e.g., Saskachewan vs. Alberta provenances from the boreal plains region). We find that groupings of genotypes are very robust, which can be explored by randomly or regionally removing provenance samples and repeating the analysis with a reduced dataset. Another indication of robustness is that alternative predictor variables such as geographic coordinates, climate variables, or various ecosystem delineations yield similar if not identical results.

Table 1 Climate variables that could alternatively be used as criteria in the multivariate regression tree analysis for aspen shown in Fig. 3

Climate variable	Threshold (right side)
Groups 1, 2, 3 vs. 4	
Mean annual temperature (°C)	≥2.65
Mean warmest month temperature (°C)	≥17.8
Groups 1, 2 vs. 3	
Mean annual temperature (°C)	≥ −0.35
Degree days <0°C	<2,248
Groups 1 vs. 2	
Mean summer precipitation (mm)	<285
Mean coldest month temperature (°C)	< −20.95

Table 2 Climate variables that could alternatively be used as criteria in the multivariate regression tree analysis for red alder shown in Fig. 4

Climate variable	Threshold (right side)
Groups 1 vs. 2, 3, 4	
Degree days >5°C	≥1,558
Mean warmest month temperature (°C)	≥15.65
Julian day where degree days reach 100	<117.5
Groups 2 vs. 3, 4	
Annual heat/moisture index (°C/m)	≥4.55
Mean coldest month temperature (°C)	<2.25
Groups 3 vs. 4	
Mean summer precipitation (mm)	≥219
Mean annual precipitation (mm)	≥1,206
Annual heat/moisture index (°C/m)	<77.1

Interpreting dendrogram bar charts

There is substantial explanatory potential in cluster averages of genetic traits that are conveniently provided by the *MVpart* software implementation as bar charts (note that if the dependent variable is univariate categorical, these charts can also represent a histogram of class frequencies). The regression tree technique maximizes multivariate homogeneity within groups. For genetic information from a single test environment, as in the red alder example, bar charts therefore represent groups of genotypes that are similar in multiple genetic traits, each bar representing the group average for a different trait. If a single trait is analyzed over multiple test environments, as in the aspen example, each bar of the chart represents the trait average in a different environment. The groups could therefore be interpreted as provenances with similar response functions to multiple environments, although a better term would be similar "genetic reaction norms", since no mathematical function is fitted to a quantitative trait in response to an environmental predictor.

The lack of curve fitting could be viewed as an advantage or disadvantage. A fitted mathematical function allows general inferences of how to maximize growth throughout a study area through seed movement. But unless there is a good mathematical fit for response functions, such general inferences may not be appropriate for multifaceted genetic variation across complex landscapes. Multivariate regression tree analysis simply clusters genotypes with similar genetic reaction norms that can take any shape. The trade-off is that the environmental factors that drive the observed reaction norms remain undefined (for example, unknown soil conditions at the planting sites could very well contribute to the observed reaction norms). Instead of relying on general response functions, we now have to examine reaction norms for each group of genotypes to determine sensible restrictions to seed movement.

In the case of aspen, the bar charts are not difficult to interpret, suggesting asymmetric seed transfer guidelines. North or northwest movement maximizes height growth or results in a neutral effect compared to growth performance of local sources (Fig. 1b, d, bar charts for the Boreal Shield and Boreal Plains genotypes). In contrast, transfer to the south or southeast has increasingly negative effects relative to the local sources (Fig. 1b, d, bar charts for the Taiga Plains and Northern Boreal Plains genotypes). In a separate study, we make the case that the observed sub-optimality of local sources should be interpreted as adaptational lag due to recent climate trends toward warmer and drier conditions throughout the study area and that seed sources should therefore be moved north or northwest by 1° to 2° latitude (Gray et al., submitted).

For the red alder case study, there is an obvious cline in growth and adaptive traits (Fig. 2a, b) that could be accounted for by zones in north–south direction along the coast. In addition, it would be prudent to delineate a seed zone corresponding to the Georgia Depression seed zone (eastern Vancouver Island and the southern mainland). Populations of the dry Georgia Depression and wet west coast of Vancouver Island are genetically differentiated, which is revealed both by geographic criteria (Fig. 2d) and by climatic predictor variables (Fig. 4). Although only a small amount of genetic variation is explained, the genetic differences appear related to fitness: On the west coast, we find very late budbreak indicating high heat-sum or chilling requirements. This prevents populations from premature leaf flush due to forcing temperatures in mid-winter under the very maritime conditions of the west coast. In contrast, provenances from the comparatively dry southern Mainland and east Vancouver Island regions lack those high heat-sum requirements, resulting in an early growing season start that takes advantage of moist growing conditions in spring. They further show higher resilience to summer droughts.

While the example with neutral genetic markers that reflect refugial history (Fig. 2e) is meant as a proof of concept, such applications are likely to become increasingly valuable as we learn how variation in molecular genetic markers shape quantitative trait variation in trees. For example, Eckert et al. (2009) identified a number of candidate genes that appear to be the target of selection for cold hardiness in Douglas-fir. In a similar range-wide study for Sitka spruce, Holliday et al. (2010) identified genes that cumulatively explained one third of the phenotypic variance in cold hardiness and bud set. Multivariate regression tree analysis could powerfully link such data to geographic or environmental variables, and surveys of nucleotide diversity linked to adaptive traits could in turn become valuable information to develop seed transfer guidelines.

Although some of the data in this paper has been analyzed before (Hamann et al. 1998, 2000, 2002), we find that none of the previously employed multivariate analysis, spatial statistics, or curve fitting techniques is as powerful in reducing complexity in the genetic dataset and in revealing reaction norms, geographic structure, and local adaptation of populations. Although we did not use soil and topoedaphic predictor variables, we think it would be promising to employ the technique to study patterns of adaptive genetic variation at finer scales. However, this will likely require reliable soil and topoedaphic data and a higher density of provenance samples.

Developing seed zones and seed transfer guidelines

Partitioning genetic variation with multivariate regression trees primarily helps us to understand how populations are adapted to the environments in which they occur. Alternatively, we can emphasize practical applications rather than scientific insight by replacing environmental predictor variables with geographic variables or ecosystem delineations. Multivariate regression trees then provide a "first draft" of seed transfer guidelines or seed-zone delineations. The groupings of genotypes seem generally very robust with respect to the choice of predictor variables, and using ecosystem delineations or geographic variables as alternative predictors can meet different objectives.

Ecosystem delineations that are based on forest types can be very good integrators of biologically relevant environmental factors for which we often have only limited data. This may allow us to make reasonable extrapolations of seed zones beyond the data coverage of provenance samples. For example, we think that extrapolation of seed zones beyond coverage of provenance samples for aspen is more realistic for ecosystems (Fig. 1c) than for geographic variables (Fig. 1a). On the other hand, geographic variables as predictors can be useful if the objective is to arrive at simple and flexible transfer guidelines. Drawing a fixed seed-zone boundary across a continuous cline of genetic variation is not always a practical solution. In the case study for red alder, one could infer from Fig. 2a that a north or south movement of any seed sources up to 2° latitude is a sensible seed transfer rule.

In this context, we should point out that multivariate regression tree analysis does not support null hypothesis testing. For example, the technique does not test whether there is a significant difference among genotypes from region A versus B, or whether transfer distance X is associated with a significant change in trait Y. We do not think that this is a drawback because such null hypotheses are a priori known to be false (e.g., Cohen 1994; Johnson 1999). However, we find it useful to carry out an analysis of effect-size statistics after groups have been determined, in order to test whether a genetic difference among regions is "significant enough" (in a non-statistical sense) to warrant separate seed zones. This requires determining thresholds of acceptable performance following seed movement and the use of confidence intervals. For example, it is straightforward to determine if seed movement from one region to another results in performance that is above an acceptable threshold (e.g., at least 90% of local sources), given a predetermined confidence level. Such thresholds could also be set in terms of absolute values for growth traits if they determine the economic viability of a reforestation program.

Unfortunately, this approach will not work for genetic markers or adaptive traits because we are usually unable to determine a meaningful threshold to apply effect-size statistics. For example, we do not know if an advance in the date of budbreak by, say, 1, 5, or 10 days relative to local sources is problematic or perhaps even desirable. We further caution against using the amount of explained variance as a criteria how many seed zones are needed. Technically, optimal delineations for a predetermined number of breeding zones are possible with multivariate regression tree analysis, but it should only be carried out if all genetic traits can be objectively weighted. For example, economic weights of growth and wood quality traits may be applied by multiplying the economic weights with the normalized dataset of genetic traits. However, determining such weights is usually not possible, and even a small amount of genetic variation explained may indicate an important adaptation of a local population.

In conclusion, we concur with Ying and Yanchuk (2006) that the development of seed transfer guidelines is as much an art as a science. Multivariate regression tree analysis simply enables us to make better-informed subjective decisions on how seed transfer should be regulated, as illustrated in the case study examples.

Acknowledgments Funding was provided by an NSERC/Industry Collaborative Development Grant CRDPJ 349100-06 and Alberta Ingenuity Grant #200500661. We thank Alberta-Pacific Forest Industries, Ainsworth Engineered Canada LP, Daishowa-Marubeni International Ltd., Western Boreal Aspen Corporation, and Weyerhaeuser Company Ltd. for their financial and in-kind support. Dave Roberts contributed a useful missing value imputation algorithm based on regression tree analysis.

References

Breiman L (2001) Random forests. Mach Learn 45:5–32

Campbell RK (1979) Genecology of Douglas-fir in a watershed in the Oregon Cascades. Ecology 60:1036–1050

Campbell RK (1986) Mapped genetic variation of Douglas-fir to guide seed transfer in southwest Oregon. Silvae Genetica 35:85–96

Campbell RK (1991) Soils, seed-zone maps, and physiography—guidelines for seed transfer of Douglas-fir in southwestern Oregon. Forest Sci 37:973–986

Cohen J (1994) The earth is round ($p<0.05$). Am Psychol 49:997–1003

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int J Climatol 28:2031–2064

De'Ath G (2002) Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83:1105–1117

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB (2009) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*). Genetics 183:289–298

EPA (2007) U.S. ecoregion maps and GIS resources. Available online: http://www.epa.gov/wed. Environmental Protection Agency, Western Ecology Division, Corvallis

Hamann A, El-Kassaby YA, Koshy MP, Namkoong G (1998) Multivariate analysis of allozymic and quantitative trait variation in *Alnus rubra*: geographic patterns and evolutionary implications. Can J For Res 28:1557–1565

Hamann A, Koshy MP, Namkoong G, Ying CC (2000) Genotype × environment interactions in *Alnus rubra*: developing seed zones and seed transfer guidelines with spatial statistics and GIS. For Ecol Manage 136:107–119

Hamann A, Namkoong G, Koshy MP (2002) Improving precision of breeding values by removing spatially autocorrelated variation in forestry field experiments. Silvae Genetica 51:210–215

Hamann A, Smets P, Yanchuk AD, Aitken SN (2005) An ecogeographic framework for in situ conservation of forest trees in British Columbia. Can J For Res 35:2553–2561

Hamann A, Wang TL (2005) Models of climatic normals for genecology and climate change studies in British Columbia. Agric For Meteorol 128:211–221

Holliday JA, Ritland K, Aitken SN (2010) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytologist 188:501–514

Johnson DH (1999) The insignificance of statistical significance testing. J Wildl Manage 63:763–772

Lesser MR, Parker WH (2006) Comparison of canonical correlation and regression based focal point seed zones of white spruce. Can J For Res 36:1572–1586

Lindgren D, Ying CC (2000) A model integrating seed source adaptation and seed use. New For 20:87–104

Little EL (1971) Atlas of United States trees. Miscellaneous publication no. 1146. United States Department of Agriculture, Washington, DC

Mangold RD, Libby WJ (1978) Model for reforestation with optimal and suboptimal tree populations. Silvae Genetica 27:66–68

Marris E (2009) Planting the forest of the future. Nature 459:906–908

Mátyás C (1990) Adaptational lag: a general feature of natural populations. Joint meeting of Western Forest Genetics Association and IUFRO Working Parties, Douglas-fir, Contorta Pine, Sitka Spruce, and Abies Breeding and Genetic Resources 20–24 August, Olympia, WA, pp 1–11

Mbogga MS, Hamann A, Wang T (2009) Historical and projected climate data for natural resource management in western Canada. Agric For Meteorol 149:881–890

McKenney D, Pedlar J, O'Neill G (2009) Climate change and forest seed zones: past trends, future prospects and challenges to ponder. Forest Chron 85:258–266

Meidinger DV, Pojar J (1991) Ecosystems of British Columbia special report series, no. 6. Research Branch, Ministry of Forests, Victoria

Morgenstern EK (1996) Geographic variation in forest trees—genetic basis and application of knowledge in silviculture. University of British Columbia Press, Vancouver

Namkoong G (1969) Nonoptimality of local races. Proceedings of the 10th Southern Conference on Forest Tree Improvement. Texas A&M University Press, College Station, pp 149–153

NRC (2006) Natural regions and subregions of Alberta. Natural Regions Committee, publication no. I/005 Alberta Environment. Natural Regions Committee, Edmonton

O'Neill GA, Aitken SN (2004) Area-based breeding zones to minimize maladaptation. Can J For Res 34:695–704

O'Neill GA, Nigh G, Wang T, Ott PK (2007) Growth response functions improved by accounting for nonclimatic site effects. Can J For Res 37:2724–2730

Parker WH (1992) Focal point seed zones—site-specific seed zone delineation using geographic information systems. Can J For Res 22:267–271

R Development Core Team (2008) R: a language and environment for statistical computing. http://www.R-project.org. R Foundation for Statistical Computing, Vienna. ISBN: 3-900051-07-0

Raymond CA, Lindgren D (1990) Genetic flexibility—a model for determining the range of suitable environments for a seed source. Silvae Genetica 39:112–120

Rehfeldt GE (1988) Ecological genetics of *Pinus contorta* from the Rocky Mountains (USA)—a synthesis. Silvae Genetica 37:131–135

Rehfeldt GE (1989) Ecological adaptations in Douglas-fir (*Pseudotsuga menziesii* var *glauca*)—a synthesis. For Ecol Manage 28:203–215

Roberds JH, Hyun JO, Namkoong G, Rink G (1990) Height response functions for white ash provenances grown at different latitudes. Silvae Genetica 39:121–129

Selby CJ, Santry MJ (1996) A national ecological framework for Canada: data model, database and programs. State of the Environment Directorate, Environment Canada, Ottawa

Wang T, Hamann A, Spittlehouse DL, Aitken SN (2006a) Development of scale-free climate data for western Canada for use in resource management. Int J Climatol 26:383–397

Wang T, Hamann A, Yanchuk A, O'Neill GA, Aitken SN (2006b) Use of response functions in selecting lodgepole pine populations for future climates. Global Change Biol 12:2404–2416

Wulder MA, White JC, Cranny M, Hall RJ, Luther JE, Beaudoin A, Goodenough DG, Dechka JA (2008) Monitoring Canada's forests. Part 1: completion of the EOSD land cover project. Can J Rem Sens 34:549–562

Ying CC, Yanchuk AD (2006) The development of British Columbia's tree seed transfer guidelines: purpose, concept, methodology, and implementation. For Ecol Manage 227:1–13

Electronic Supplement 1. Sample dataset of the aspen provenance trial with normalized height measurements from five test sites in comma seperated values (.csv) format. To obtain corresponding climatic data for the last analysis of the aspen case study, the first four columns need to be processed with the software package ClimatePP (Mbogga et al. 2009).

```
ECO,  LAT,   LONG, ELE, HT70, HT10, HT60, HT33, HT90
TP ,58.20,-123.33,1177,-0.69,-1.05,-1.99,-1.85,-2.22
TP ,58.40,-123.00, 511,-1.54,-2.00,-2.15,-2.95,-1.23
TP ,58.60,-122.33, 335,-0.77,-1.46,-1.68,-2.98,-1.27
NBP,56.61,-117.98, 709, 1.97, 1.02,-0.40,-0.01,-0.50
NBP,57.78,-117.96, 459, 2.10, 0.33,-0.13,-0.08, 0.29
NBP,56.41,-117.76, 739, 0.78, 0.28,-0.84,-0.60, 0.20
NBP,57.12,-117.74, 606, 2.22, 0.59,-0.49,-0.34,-0.76
NBP,58.56,-115.64, 343, 1.43, 0.52,-0.51,-1.40,-1.40
BPA,55.60,-116.67, 632, 0.53,-0.01,-0.16,-0.60, 0.37
BPA,55.60,-116.67, 632, 0.70,-0.62, 0.14,-0.59, 0.74
BPA,55.60,-116.67, 632, 0.06,-1.05, 0.51,-0.15, 0.87
FH ,55.03,-118.73, 649,-1.24, 0.56, 0.43,-0.57,-0.16
FH ,54.21,-116.59, 914,-1.17, 1.52, 0.20, 0.13, 1.00
FH ,54.14,-116.58, 868, 1.13,-0.48,-0.50,-0.38,-0.55
FH ,54.10,-116.50,1018, 0.98,-2.67,-1.33,-0.63,-1.25
FH ,54.21,-116.44, 803, 0.55, 0.06,-0.17,-0.51,-0.27
FH ,54.21,-116.44, 803,-1.22,-0.72,-0.03,-0.45,-1.02
FH ,54.18,-115.78, 731, 0.55, 1.61, 0.16, 1.25,-0.55
FH ,53.20,-115.60, 939,-1.34, 0.35,-0.27, 0.10,-0.97
FH ,53.31,-115.46, 939,-1.94, 0.59,-0.62,-0.05,-1.10
FH ,53.30,-115.43, 927,-1.14, 0.35,-0.54,-0.45,-0.35
BPA,55.34,-115.01, 646,-0.32,-0.22, 0.03, 0.53, 0.16
BPA,55.64,-114.69, 709, 0.58, 0.30,-0.40, 0.42, 0.55
BPA,55.19,-114.61, 726, 1.13, 0.83, 0.43, 0.76, 1.25
BPA,55.60,-113.41, 762, 0.35,-0.10, 0.08, 1.37, 0.91
BPA,55.14,-113.02, 601,-0.62, 0.18,-1.17, 0.17,-0.42
BPA,54.94,-112.86, 546,-0.52, 0.56, 0.33, 0.29,-0.93
BPA,54.93,-112.74, 545,-1.02, 1.90,-0.66, 0.20,-0.40
BPA,55.06,-112.11, 624,-1.07,-0.86,-1.13, 0.24,-0.87
BPS,53.80,-108.50, 710,-0.37,-1.24, 0.52, 0.34, 0.46
BPS,54.03,-108.00, 530,-0.57,-1.08, 0.66, 0.84, 0.50
BPS,53.90,-107.50, 570, 0.18, 1.09,-0.11, 0.43,-0.05
BPS,54.00,-106.90, 519,-1.12, 1.40, 0.40, 0.89, 1.55
BPS,54.20,-106.80, 513, 0.11,-0.67, 0.28, 0.65,-0.37
BPS,53.80,-106.70, 583,-0.04, 0.30, 0.20,-0.19,-0.82
BPS,53.90,-105.80, 517, 0.33,-1.22,-0.09, 0.04, 0.59
BPS,54.20,-105.70, 490,-0.09, 0.30, 0.37, 0.18,-0.89
BS ,47.20, -93.80, 405, 0.83, 0.42, 2.65, 2.23, 1.19
BS ,47.50, -93.60, 433,-0.02, 0.33, 2.49, 1.34, 1.79
BS ,47.60, -93.40, 424,-0.37,-0.60, 1.87, 1.17, 1.77
BS ,47.20, -93.40, 395,-0.17,-0.89, 1.26,-0.02, 1.38
BS ,47.00, -93.00, 384, 0.13,-0.10, 1.37, 1.04, 1.38
```

Electronic Supplement 2. R code for multivariate regression-tree analysis, which generates the dendrograms and bar charts equivalent to Figures 1B and 1D. Copy and paste the dataset from Appendix A into a plain text file for import to R. Running the code requires the installation of the *mvpart* package for R.

```
library(mvpart)
data=read.csv("C:\Your Path\Dataset in Supplement 1.csv")

# Partitioning by latitude, longitude, and elevation. Note that
# the option xv="p" allows the user to control the tree size by
# clicking on the desired y-axis position in the scree plot:
out=mvpart(data.matrix(data[,5:9])~LAT+LONG+ELE, data, xv="p")
summary(out)

# Partitioning by ecosystem and elevation:
out=mvpart(data.matrix(data[,5:9])~ECO+ELE, data, xv="p")
summary(out)
```

Electronic Supplement 3. R code for imputation of missing values. For illustration, we first generate a number of random missing values in the dataset, then impute the missing values, and subsequently repeat the analysis of Appendix B. Running the code requires the installation of the *randomForest* package for R.

```
library(randomForest)
library(mvpart)
data=read.csv("C:\Your Path\Dataset in Supplement 1.csv")

# Randomly generate ~10% missing values and view the result:
# Increase "sample(5)" for a larger proportion of missing values.
for (i in 5:9) data[sample(42, sample(5)), i]<-NA; data

# Impute missing values and view results. "ECO" can be any ID.
filled=rfImpute(ECO~HT70+HT10+HT60+HT33+HT90, iter=3,
ntrees=10000, data); filled

# Repeat analysis and compare to results from Supplement 2 code:
out=mvpart(data.matrix(filled[,2:6])~LAT+LONG+ELE, data, xv="p")
summary(out)
```