DTD 5
ARTICLE IN PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Ecological Modelling xxx (2005) xxx–xxx

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees

Mika Sulkava*, Jarkko Tikka, Jaakko Hollmén

Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 HUT, Finland

Abstract

Analyzing and predicting the development of foliar nutrient concentrations are important and challenging tasks in environmental monitoring. This article presents how linear sparse regression models can be used to represent the relations between different foliar nutrient concentration measurements of coniferous trees in consecutive years. In the experiments the models proved to be capable of providing relatively good and reliable predictions of the development of foliage with a considerably small number of regressors. Two methods for estimating sparse models were compared to more conventional linear regression models. Differences in the prediction accuracies between the sparse and full models were minor, but the sparse models were found to highlight important dependencies between the nutrient measurements better than the other regression models. The use of sparse models is, therefore, advantageous in the analysis and interpretation of the development of foliar nutrient concentrations.

Keywords: Linear sparse regression; Prediction; Foliar nutrition; Coniferous tree

1. Introduction

Analyzing the condition and development of forests is challenging due to the complex nature of forest ecosystems. Many different environmental conditions affect the growth and well-being of trees. There are several large-scale forest monitoring programs worldwide trying to assess the present and future condition of forests. An example is the International Co-operative Programme on the Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests). Because environment and foliar mineral composition are related, the programs monitor, among other things, the mineral composition of tree foliage.

Analysis of foliar nutrient concentrations is an important part of forestry and environmental monitoring. Understanding and predicting the development of nutrient concentrations based on measurement data of the forest are challenging tasks, due to the diversity of environmental conditions that affect the development.

* Corresponding author. Tel.: +358 9 451 3647; fax: +358 9 451 3277.
E-mail addresses: Mika.Sulkava@hut.fi (M. Sulkava), tikka@mail.cis.hut.fi (J. Tikka), Jaakko.Hollmen@hut.fi (J. Hollmén).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.08.016

We take a systematic view in evaluating the value of parsimonious models in an ecological domain, specifically, estimating the dependencies between foliar nutrients. We approach the problem by employing sparse regression techniques. In the experiments two linear sparse regression models are compared to simple one-parameter linear regression models and full linear regression models in a foliar nutrient prediction task. One-parameter regression tries to explain the data with only one regressor, whereas full regression uses all available regressors for the task. Sparse regression models only use a few regressors, that are capable of explaining the data. The regressors are selected using two different algorithms: conventional forward selection (Hastie et al., 2001) and least angle regression (LARS), a more sophisticated algorithm recently published by Efron et al. (2004).

The different models are tested with foliar nutrient measurement data acquired from the coniferous forests of Finland. It is studied, how well linear models can predict the nutrition status of 1-year-old tree needles given the nutrition status of new needles in the same location 1 year before and some information about e.g. the weather and deposition in the forest. We are trying to model this way the combined effect of aging and environment on needles.

The results of the experiments revealed that usually a simple one-parameter model is not capable of providing comparable results to the other three models. The prediction accuracy of the sparse and full regression models was found to be rather similar to each other. However, the sparse models provide the predictions with a much smaller number of parameters, which makes the models more interpretable.

2. Foliar nutrient data

The foliar nutrient data was measured from needle samples collected from plots dominated by coniferous trees. The plots were located in different parts of Finland in the so-called background areas, where local sources of air pollution were considered to be absent. The nutrient data used in the analysis consist of needle mass (NM) and 12 element concentrations: Al, B, Ca, Cu, Fe, K, Mg, Mn, N, P, S and Zn. Needle mass was measured as the mass of 1000 needles and the concentrations either as mg/g or μg/g. These 13 measurements were made to needles of foliar age classes C and $C + 1$, i.e. the needles that were grown in the year the measurements were done and in the previous year, respectively.

The measurements were made annually for 14 years between 1987 and 2000 in October or November in 36 stands of the Finnish Level I network of ICP forests (Luyssaert et al., 2003). Sixteen of the stands were dominated by Norway spruce and 20 by Scots pine. Thus, the nutrient data consist of $2 \times 16 \times 14 \times 13$ measurements of Norway spruce and $2 \times 20 \times 14 \times 13$ measurements of Scots pine. However, 33.1% of the nutrient data of spruce and 23.9% of pine were unavailable. For details concerning the sampling procedure, see Stefan et al. (1997).

In addition, there were nine additional measurements available for the stands, namely the geographic coordinates (X and Y), the total N and S deposition (NT and ST), the average temperature TA and total precipitation PT, the deviations of average temperature and precipitation from their long term averages (TD and PD) and the age A of the forest. The depositions were available for years 1987–1996. All other additional measurements were known for years 1986–1999. Except the coordinates, also the additional measurements were done once per year.

Luyssaert et al. (2004) built nutrition profiles for the nutrient data described above, but only using the measurements of C needles. It was also pointed out that the use of $C + 1$ needles could help to show the dynamics of the elements. In this work the data of $C + 1$ needles is used, which allows analysis of the aging of the needles.

The problem in this work is to predict the 12 different nutrient concentrations and needle mass of $C + 1$ needles in year t using the measurements of C needles in year $t - 1$ and the additional measurements in year t, altogether 22 measurements. That is, we want to model the effect of the environment and nutrients on the aging of the needles. The aim is to use only a few significant regressors of total 22 for each response and to highlight the importance of these variables. The most significant regressors are selected separately for each response, so that differences in dependency relations between the response and regressors in different models can be observed more easily. Separate models are also generated for spruce and

pine in order to compare the differences between the species.

3. Methods

Let us assume that the form of the available data is $(Y, X_1, \ldots, X_K)$, where Y and X_i, $i = 1, \ldots, K$ are $(N \times 1)$-vectors. The problem is to predict the values of Y using the variables X_i. Dependencies between the variables can be analyzed using the multiple linear regression model

$$Y = \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon, \tag{1}$$

where Y is a dependent variable or response. X_i and β_i, $i = 1, \ldots, K$ are regressors and corresponding regression coefficients, respectively, and ϵ is normally distributed random noise with zero mean and unknown variance. Eq. (1) can equivalently be represented in matrix form as $Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]^{\mathrm{T}}$ and the columns of $(N \times K)$–matrix $\boldsymbol{X}$ are regressors X_i, $i = 1, \ldots, K$. Here we assume that the data are normalized to zero mean and thus, there is no need for a constant term in the model. The coefficients β_i are usually estimated by minimizing the residual sum of squares (RSS) between the target value and the estimated value. The ordinary least squares (OLS) solution is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}Y, \tag{2}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = [\hat{\beta}_1, \ldots, \hat{\beta}_K]^{\mathrm{T}}$ are estimates of the regression coefficients.

The full regression model estimated using OLS is not always the best model for prediction or interpretation of the dependencies. Using a sparse regression model instead of the full model is often more sensible. In a linear sparse regression model there are $k < K$ nonzero regression coefficients, where K is the number of regressors in a full regression model. The ratio of observations to the regressors should never be under 5, but the desired level is between 15 and 20 (Hair et al., 1995). The desired ratio is not achieved by the full regression model with the data used in this study. Dropping out non-informative regressors increases the ratio of observations to regressors. The most significant regressors are also more clearly seen from a sparse model. At the same time the sparse regression model should be less prone to overfitting than the full model.

A well-known algorithm for improving the OLS estimates is the forward selection algorithm (Hastie et al., 2001). In this study it is used as a baseline method for sparse regression. Forward selection starts by finding the most correlated regressor with the response. After that, the regressors, which improve the fit the most, are sequentially added to the model. The improvement in the fit can be calculated according to the F-statistics based on the reduction in the residual sum of squares

$$F = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}}_k) - \mathrm{RSS}(\hat{\boldsymbol{\beta}}_{k+1})}{\mathrm{RSS}(\hat{\boldsymbol{\beta}}_{k+1})/(N-k-2)}. \tag{3}$$

Above $\hat{\boldsymbol{\beta}}_k$ represents the current model with k regressors. To obtain $\hat{\boldsymbol{\beta}}_{k+1}$ the regressor that maximizes F is added to the model. The addition of regressors is stopped when no regressor produces an F-ratio greater than the 95th percentile of the $F_{1,N-k-2}$ distribution. Forward selection usually takes too long steps toward the final model and it might ignore useful regressors which are correlated with the already added regressors. Small perturbations in the data may cause drastic changes in the model estimated using forward selection (Breiman, 1995).

Ridge regression (Hoerl and Kennard, 1970) and lasso (Tibshirani, 1996) algorithms also improve the OLS estimates and produce a sparse solution or at least shrink estimates of the regression coefficients toward zero. A penalized sum of squares is minimized in both algorithms

$$\min_{\boldsymbol{\beta}} \left\{ \|Y - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{K} |\beta_i|^{\gamma} \right\}, \tag{4}$$

where $\gamma = 2$ in ridge regression and $\gamma = 1$ in lasso and λ is a tuning parameter. The tuning parameter controls the amount of shrinkage that is applied to the coefficients. The problem in Eq. (4) can be represented equivalently as a constrained optimization problem

$$\min_{\boldsymbol{\beta}} ||Y - \boldsymbol{X}\boldsymbol{\beta}||^2 \quad \text{s.t.} \sum_{i=1}^{K} |\beta_i|^{\gamma} \leq \tau. \tag{5}$$

In the previous equation τ controls the amount of shrinkage. The parameters λ in Eq. (4) and τ in Eq.

(5) are related to each other by a one-to-one mapping (Hastie et al., 2001). A large value of λ corresponds to a small value of τ. The benefit of applying shrinkage to the regression coefficients can be achieved in the improved prediction accuracy (Copas, 1983).

The ridge regression solution is easy to calculate, because the penalty term is continuously differentiable. The solution can be written as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}Y, \tag{6}$$

where $\boldsymbol{I}$ is an identity matrix. This solution does not necessarily set any coefficients exactly to zero. Thus, the solution may still be hard to interpret if the number of regressors K is large. The lasso penalty sets some coefficients to zero with a proper τ, but finding the lasso solution is more complicated due to absolute values in the penalty term. A quadratic programming algorithm has to be used to compute the lasso solution. Also, the value of τ or λ which controls the shrinkage is strongly dependent on data. Therefore, seeking such a value might be difficult in many cases. Data-based techniques for estimation of the parameter τ are presented by Tibshirani (1996).

3.1. Model selection

In this study forward selection and least angle regression (LARS) (Efron et al., 2004) model selection algorithms are used to select the most informative regressors. LARS produces nearly the same solutions which can be achieved by lasso using different values of tuning parameter τ. LARS is completely parameter free and it is computationally more efficient than lasso. LARS is also closely related to the forward selection algorithm, but LARS can be regarded as a less greedy version of it (i.e., LARS does not take as long steps toward the final model as forward selection).

In the LARS algorithm K steps are needed for the full set of solutions. One regressor is added to the model in each step. First, all regression coefficients are set to zero. Then, the most correlated regressor X_{i_1} with Y is found. The largest possible step is taken in the direction of X_{i_1} until some other regressor X_{i_2} is as correlated with current residuals as X_{i_1}. The next step is taken in a direction equiangular between X_{i_1} and X_{i_2}. LARS proceeds in this direction until a third regressor X_{i_3} is as correlated with current residuals as X_{i_1} and X_{i_2}. The following step is taken in a direction equiangular between X_{i_1}, X_{i_2} and X_{i_3} until a fourth regressor can be added to the model. This procedure is continued as long as there are regressors left. The LARS algorithm is presented in more detail by Efron et al. (2004).

A problem is to find the best solution of the K solutions returned by LARS. An initial value k_i is selected based on the minimum description length (MDL) information criterion (Hansen and Yu, 2001). The MDL criterion can be written in the context of linear regression as follows:

$$\mathrm{MDL}(k) = \frac{N}{2}\log\|Y - \hat{Y}\|^2 + \frac{k}{2}\log\, N, \tag{7}$$

where Y is the dependent variable, $\hat{Y}$ the estimate of the dependent variable, N the sample size and k is the number of added regressors. The value of Eq. (7) is calculated for each K solution. The initial value k_i has the smallest value.

The Mallows (1973) C_{p} criterion is a common criterion in subset selection. However, C_{p} is not used in this study, because it might select submodels of too high dimensionality (Breiman, 1992). Akaike's information criterion (AIC) is also presented in the context of linear regression by Hansen and Yu (2001). The first term in the AIC is the same as in Eq. (7), but the second term is k. Thus, in the MDL criterion larger penalty is applied to the number of regressors than in AIC if $N > 7$, and if the natural logarithm is used. In this study this condition is fulfilled so the MDL criterion likely produces sparser models than AIC.

Subsequently, the k_i regressors that are selected based on the LARS algorithm and the MDL criterion are taken to further analysis and the rest of the regressors are discarded. The OLS solution is calculated using those selected k_i regressors. The final k regressors are obtained by setting statistically insignificant coefficients to zero. The t-test is used in estimation of the confidence intervals of the coefficients. Coefficients are regarded insignificant if they have value zero in the confidence interval $1 - \alpha$. Only the significant regressors are taken to the final linear sparse regression model.

3.2. Other related methods

The use of linear models is justified by their interpretability and the fact that over short ranges, any process can be well approximated by a linear model (Guthrie et al., 2005). An advantage of linear sparse models over more complicated models, e.g. neural networks, is that the number of user defined parameters that affect the result and make the modeling more difficult is small or in some cases zero. For example, in a multilayer perceptron (Haykin, 1999) one has to define the whole structure of the network, which has a great impact on the quality of the predictions and generalization capability of the network. If the structure is not optimal, there is a risk of obtaining very poor predictions.

In this study the objective is to predict and explain multivariate responses using the same set of regressors. Linear sparse regression models are constructed separately for each response. The Curds & Whey (C&W) procedure (Breiman and Friedman, 1997) could also be used in prediction of multivariate responses. First, in the C&W procedure the OLS estimates $\hat{Y}_i$ are calculated for each response. After that, the final estimates of the responses $\tilde{Y}_i$ are constructed as linear combinations of the OLS estimates $\hat{Y}_i$. In the C&W procedure $LK + L^2$ parameters have to be estimated, where L is the number of responses. The number of estimated parameters in the L linear sparse models is much lower. Therefore, the relations of dependencies in the linear sparse models are easier to interpret than in the C&W model. In the C&W procedure it is also assumed that the responses are correlated. The data sets analyzed in this study do not fulfill this assumption. In the set of responses of spruce the correlation is on average 0.21 and in the set of responses of pine the average is 0.25. The C&W procedure has some similarities with nonnegative (nn) garrote (Breiman, 1995). The regression coefficients are shrunk by some positive values in nn-garrote, but it is a single response method when the C&W procedure is a multivariate response method.

If the goal was only the prediction accuracy, it might be beneficial to use some piecewise linear procedure, e.g. multivariate adaptive regression splines (MARS) (Friedman, 1991). In the MARS procedure several piecewise linear basis functions are constructed from the original regressors. The final model is generated using weighted sums and products of the basis functions. The number of estimated parameters can be very large compared to linear sparse regression. The advantage of basis functions is that they explain the local dependencies better than a global linear model. On the other hand, the use of basis functions decimates interpretability of the dependencies between the original variables in the model, which is an important criterion for the estimated models in this study. The MARS procedure is also computationally more intensive than linear sparse regression algorithms.

3.3. Model validation

The quality of LARS model is compared to forward selection, full linear regression and simple one-parameter models. The quality of predictions of the different models is analyzed using cross-validation (Bishop, 1995). Let us assume that M training and validation data sets are used. The models are estimated using the training data sets and the validation sets are used to measure the prediction accuracy. The coefficient of determination R^2 is used as a measure of prediction accuracy.

Cross-validation is also used for computing the relative importance of the regressors in LARS models. The value of relative importance describes the strength of belief that the corresponding regressor belongs to the linear sparse regression model. If the regressors are scaled to zero mean and unit variance, the relative importance of the regressors is computed as follows:

$$\boldsymbol{w} = \frac{1}{M} \sum_{j=1}^{M} \frac{|\hat{\boldsymbol{\beta}}_j^*|}{\mathbf{1}^{\mathrm{T}} |\hat{\boldsymbol{\beta}}_j^*|}, \tag{8}$$

where M is the number of training data sets used in cross-validation, $\hat{\boldsymbol{\beta}}_j^*$ are estimates of the regression coefficients from jth training data set, $j = 1, \ldots, M$, and $\mathbf{1}$ is a vector of ones. The absolute values are taken over all components of the parameter vector $\hat{\boldsymbol{\beta}}_j^*$. $\boldsymbol{w}$ is a K-vector including a value of relative importance for each regressor. The value of each $w_i, i = 1, \ldots, K$ is within range $w_i \in [0, 1]$ and $\sum_i w_i = 1$. A large value of relative importance indicates that the corresponding regressor is important in the estimated model.

The values of relative importance show which regressors are likely to be included in the final linear sparse regression model. In addition, the values of

relative importance are estimates of the explanation power of the regressors. Frequencies of occurrences of the regressors in the models computed using cross-validation could also be used as estimates of probabilities that the corresponding regressors should be included in the final model. However, there is a disadvantage in the frequencies compared to the values of relative importance.

Let us assume that regressors X_{j_1} and X_{j_2} are selected in every model estimated using the M training data sets and the regression coefficient of X_{j_1} is always clearly inequal to zero and the regression coefficient of X_{j_2} is always very close to zero. According to the frequencies of occurrences both regressors would be highly important, although in practice regressor X_{j_2} does not have remarkable explanation and prediction power. In this case the value of relative importance of X_{j_1} would be high compared to the value of relative importance of X_{j_2}, which characterizes better the importance of the regressors in the final model. In this kind of case regressor X_{j_2} can be rejected from the final model. This way, a sparser model could be obtained without a notable loss in the prediction accuracy. However, the frequencies should not be discarded completely, because they can give useful additional information about the robustness of the model selection alongside with the values of relative importance.

The reliability of LARS models is also studied with permutation tests (Good, 2000). Multiple sparse models are generated by randomly selecting the regressors in the model. The number of randomly selected regressors equals the number of regressors in the LARS model. Comparing the random models to the estimated model allows analysis of the quality of the model selection procedure.

4. Experiments

The models used in the prediction are different multiple linear regression models

$$X_{i,t,C+1} = \sum_{j=1}^{13} \beta_{i,j} X_{j,t-1,C} + \sum_{j=14}^{22} \beta_{i,j} Z_{j,t} + \epsilon_i. \quad (9)$$

Above, $X_{j,t,C}$ denotes the concentration of the jth nutrient (or needle mass) in C needles in year t. $Z_{j,t}$ is the value of the jth additional measurement in year t, and ϵ_i is normally distributed noise with zero mean and unknown variance. The same model is assumed to hold in all the measurement stands.

In the experiments the data were normalized to zero mean and unit variance in order to give all the measurements equal weights in the construction of the models and to make the values of the regression coefficients more comprehensible.

Four different models were compared: the conventional full linear regression, forward selection and LARS linear sparse regression explained in Section 3 and one-parameter regression, that tries to predict the value of a $C + 1$ measurement in year t by only using its C-value in year $t - 1$, i.e. $\beta_{i,j} = 0 \, \forall i \neq j$.

The quality of the predictions of the different models, one-parameter, forward selection, LARS and full regression models, was studied using 10-fold cross-validation, that was repeated 20 times. Thus, the total number of training and validation sets was $M = 200$ for each model class. The data were randomly permuted before each cross-validation round. The random permutation was also done separately for different models. The accuracy of prediction of the different models was measured with the coefficient of determination R^2.

Cross-validation was used in order to determine the actual performance of the models and to ensure that instead of overfitting the training data, the models also work with other data from the same source. It is insufficient only to study the R^2-values of the whole data set, because it can be improved e.g. simply by using random noise as additional regressors.

The results of cross-validation for spruce and pine are shown in Fig. 1 for both the training and the validation sets. In Fig. 1b and d it can be seen that for most measurements the sparse models outperform the simple one-parameter model, and their prediction accuracy is mainly comparable to the full model. Even though the full model gets the highest average scores with the training set, for majority of the measurements the difference in the average score with the validation set, when compared to the sparse models, is rather small.

However, the number of parameters in the sparse models is much lower. In LARS models there are on average $k = 3.6$ coefficients for spruce and $k = 5.2$ for pine ($K = 22$). In forward selection models the average number of coefficients is $k = 5.3$ for spruce and $k = 6.8$ for pine. The sparse models fit rather well to the data without any noticeable signs of overfitting.

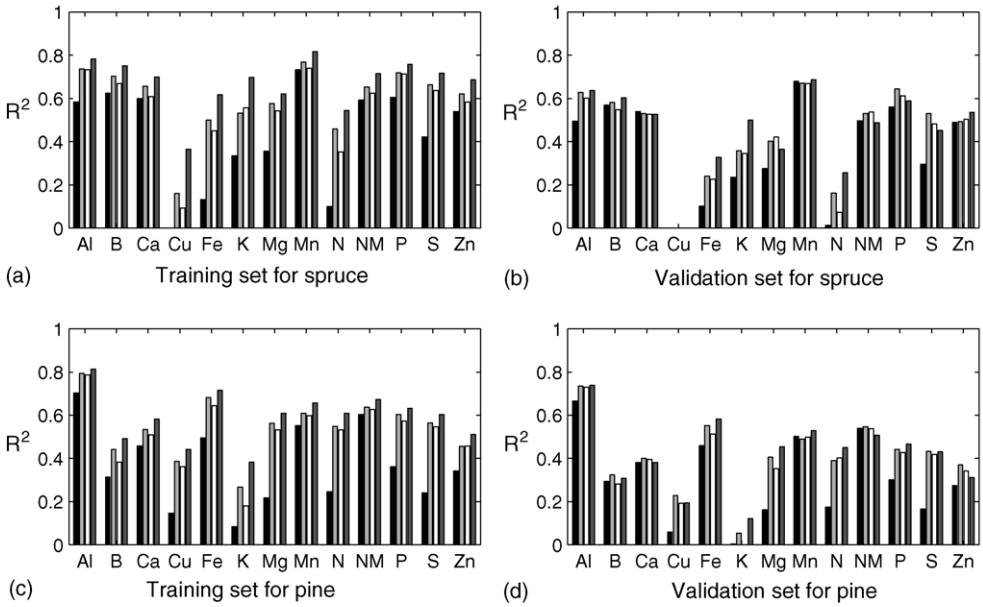

Fig. 1. Average R^2-values of the measurements of spruce and pine for one-parameter regression (black bars), forward selection (light gray bars), LARS (white bars) and full regression (dark gray bars) obtained using cross-validation. Results are for both training (a: spruce and c: pine) and validation sets (b: spruce and d: pine).

Leaving the less important regressors out of the model decreases the difference between the average R^2 score of the training and validation sets. This also facilitates the study of the dependencies between different measurements. Therefore, the sparse models can be regarded to be more suitable for the problem than the two other models.

The effect of leaving the statistically insignificant regressors out of the LARS model was rather small. For the spruce data the average decline of the R^2-value per regressor that was left out was 0.013 and for the pine data 0.006.

For some models the average R^2-value of the validation sets is ≤ 0. These include Cu for spruce and K for pine. It can be concluded that using the given regressors, the sample mean is the best linear predictor for Cu concentration in spruce needles, and for concentration of K in pine needles the sample mean provides comparable prediction accuracy to the other linear models.

The differences of the average R^2-values were studied with two-tailed *t*-tests. The three null hypotheses were the following. H_{0a}: the average R^2-values of one-parameter model and LARS model are equal in the validation sets, H_{0b}: the average R^2-values of forward selection model and LARS model are equal in the validation sets and H_{0c}: the average R^2-values of full regression model and LARS model are equal in the validation sets. The alternative hypotheses were that there is a difference in the average R^2-values in one direction or the other.

Due to the multiple random permutations during cross-validations it is very unlikely that there would have been similar validation sets for different models. Therefore, the R^2-values of different models in the validation sets can be regarded independent of each other and the *t*-test is appropriate in this case. If the null hypothesis is rejected, it is concluded that there is a difference in the average prediction accuracy of the models with significance level $1 - \alpha$.

For spruce significant differences ($p < 0.05$, $1 - \alpha = 0.95$) between the average R^2 score of the one-parameter model and LARS model were found with Al, Cu, Fe, K, Mg, N, P and S and for pine with Al, Cu, Fe, Mg, N, P, S and Zn. Significant differences between forward selection model and LARS model were found with Cu and N for spruce and with K and Mg

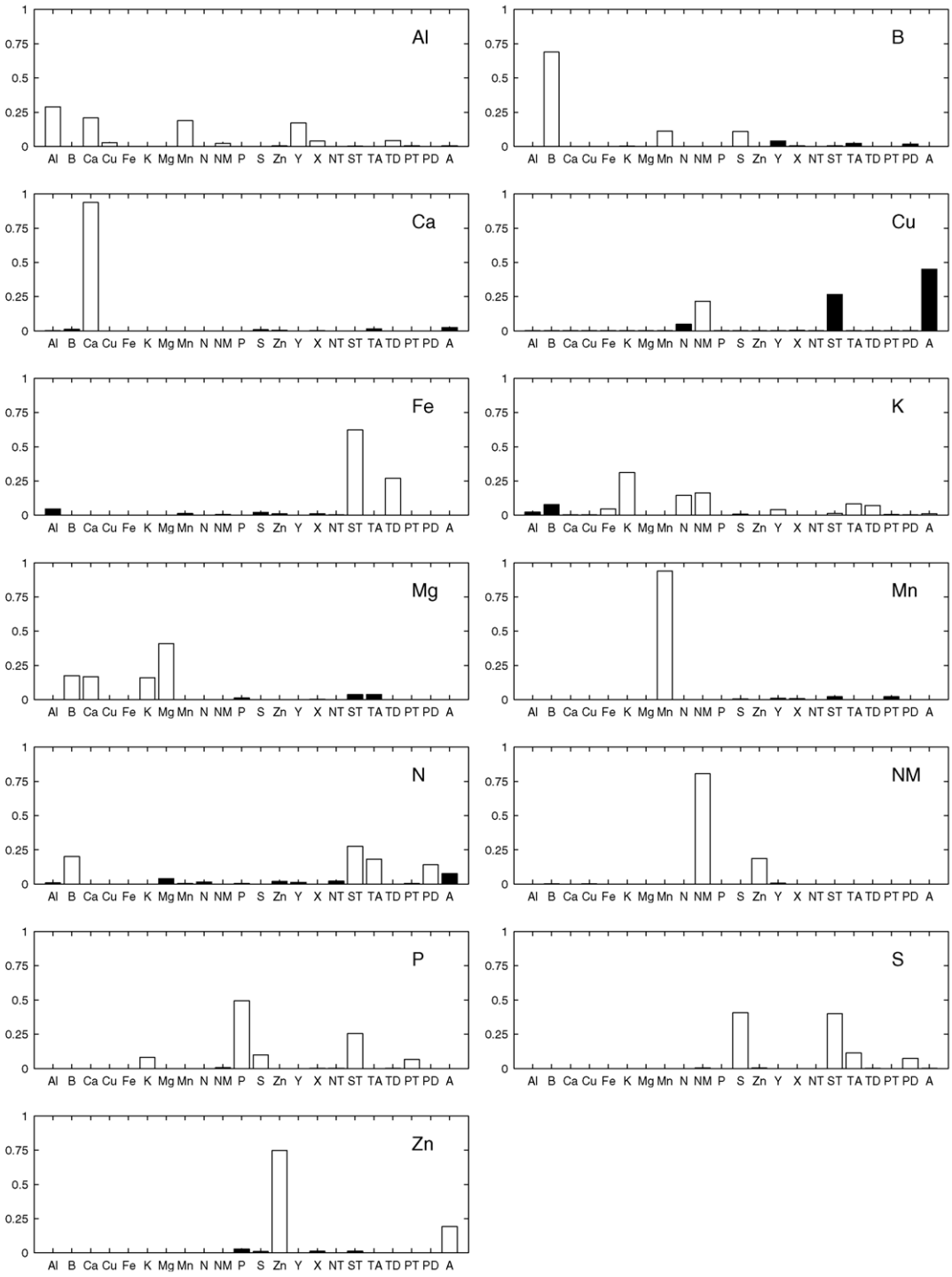

Fig. 2. Relative importance of the regressors in LARS models for spruce (see Eq. (8)). White and black bars denote the regressors that were selected and not selected in the final sparse model, respectively.

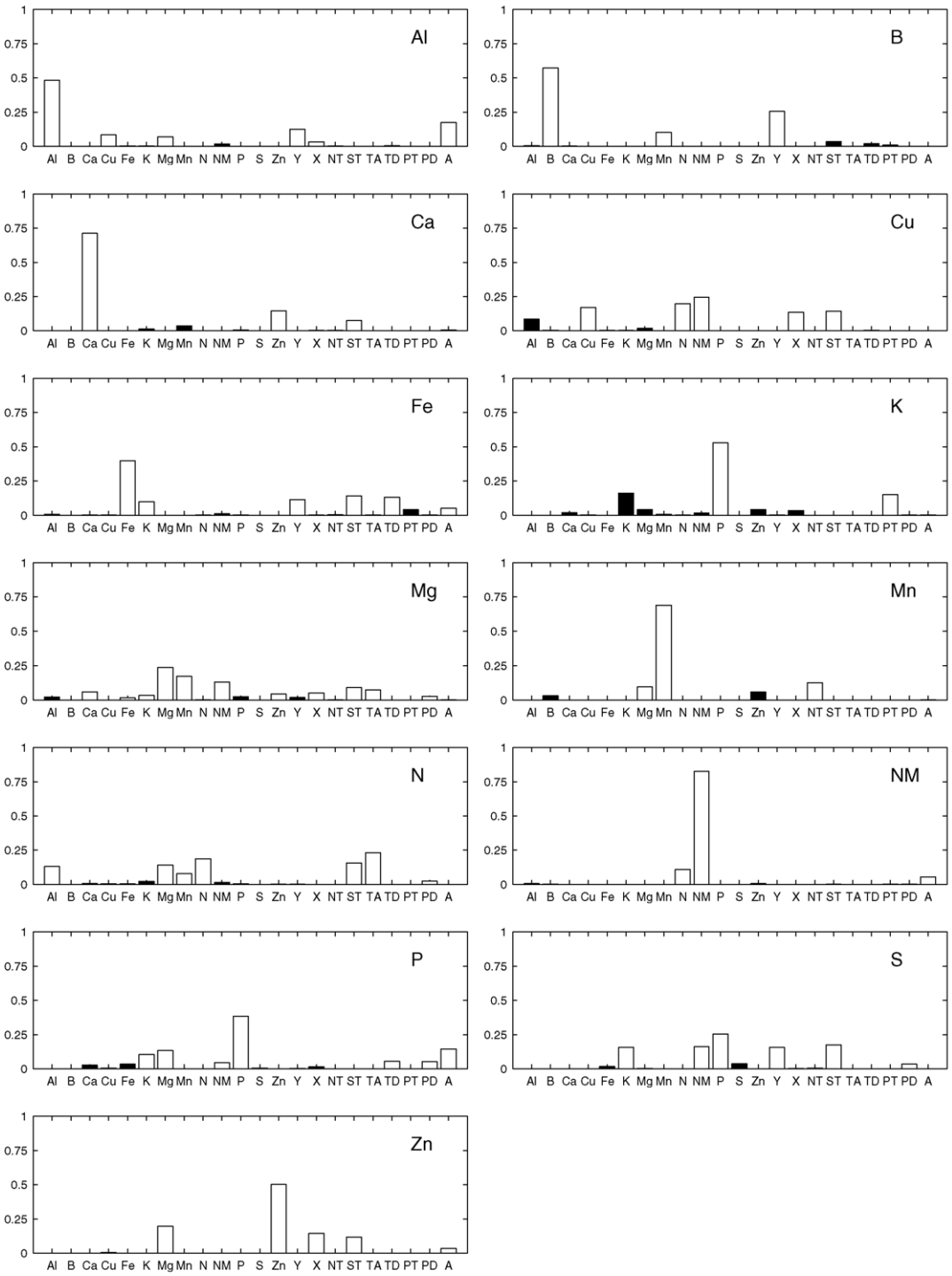

Fig. 3. Relative importance of the regressors in LARS models for pine (see Eq. (8)). White and black bars denote the regressors that were selected and not selected in the final sparse model, respectively.

for pine. The differences between the full model and LARS model for spruce are significant with B, Cu, Fe, K, Mg, N and NM and for pine with Fe, K, Mg, and N.

In addition, using the 20 repetitive 10-fold cross-validations, relative importance of the regressors in LARS model was computed, that reveals the strength of the connections between different measurements. The value of relative importance was computed according to Eq. (8) as the average weight of the regressors obtained using each of the $M = 200$ training sets. The results for both spruce and pine are shown in Figs. 2 and 3. Also, it is shown, which regressors were selected in LARS model, when the model selection procedure presented in Section 3.1 was applied to the whole data. The selected regressors have almost always the largest relative importance, i.e. they have the strongest explanation and prediction power. Two responses, Cu for spruce and K for pine, are exceptions. Using cross-validation it was found that those responses cannot be explained by the given regressors and thus, they are not of special interest in this case.

The frequencies corresponding to how often the regressors were selected in LARS models during cross-validation are shown in Appendix B (Tables B.1 and B.2). For each response the regressors in the pine data that were selected in the final model have the highest frequencies. The response Cu in spruce data is again an exception. In addition, the regressor B is rejected from the final model when K is the response in the spruce data, although its frequency is 0.62 and some selected regressors have lower frequencies. Also, its relative importance is greater than of some selected regressors. Otherwise, linear sparse regression model selection using LARS works very well according to the available data based on the prediction and explanation power of the regressors. Forward selection tends to choose the regressors more variably. The frequencies corresponding to how often the regressors were selected to the models during cross-validation are more evenly distributed than with LARS.

Usually the strongest connection for a variable is naturally found to its previous year value. Also other, more interesting dependencies were found between the variables. A typical example of a sparse model is the LARS model of needle mass for pine:

$$\mathrm{NM}_{t,C+1} = 0.14 N_{t-1,C} + 0.76 \mathrm{NM}_{t-1,C} - 0.10 A_t. \quad (10)$$

Above, the needle mass of $C + 1$ needles in year t is most strongly dependent on the needle mass of C needles in year $t - 1$. In addition, the previous year nitrogen concentration $N_{t-1,C}$ has a small positive effect on the needle mass, whereas the age of the forest A_t has a slightly smaller negative effect. All the LARS models for spruce and pine are shown in Appendix A.

We studied how well the regressors were selected in the LARS model using a permutation test. For each nutrient numerous models with randomly selected regressors were generated, where the number of regressors equaled the number of regressors in the LARS model. Their performance was compared and it was found that virtually always the best possible regressors were selected in the sparse model. Out of 1000 random regressor permutations on average 0.08% explained the spruce data better than the LARS model. For pine the percentage was 0.02%. Thus, given the number of regressor, it is extremely difficult to construct a linear model that would better characterize the relations between the measurements. In all cases if the model with randomly selected regressors was better than LARS model, the difference between the R^2-values was extremely small.

5. Conclusions

The applicability of different linear regression models for predicting the foliar nutrient concentrations in coniferous trees was studied in this work. It was found that linear sparse regression models are well suitable models for the task. The main advantage of the models is their sparsity, that makes interpretation of the model easy, without significantly reducing the prediction accuracy compared to the full model.

A simple one-parameter model is, in many cases, not sufficient to characterize the development of nutrients in the needles. Thus, the concentration of an element in needles cannot be assumed only to depend on its concentration in the previous year. The sparse model selection procedures automatically find the strongest and most informative dependencies between the measurements. This makes the sparse models easy to understand and at the same time more useful than a conventional linear regression model, because the most important regressors are clearly highlighted.

The results of the full linear model were of similar quality as the results of the sparse models. Using cross-validation it was found that the reliability of the estimated sparse models was very good, because the responses could be explained well compared to the full model, but only using a small number of regressors. In addition, using permutation tests the LARS model selection was shown to be of outstanding quality. Therefore, the use of sparse models can be recommended in foliar nutrient prediction.

LARS models are slightly sparser than forward selection models, and the difference between the prediction accuracy is mostly insignificant. Also, LARS seems to be somewhat more robust in the selection of the regressors. Therefore, use of LARS instead of forward selection can be encouraged in sparse model selection.

The resulting sparse models can help in the interpretation of the domain. One example is guiding in the laboratory work after the sample collection. By analyzing the dependencies between the measurements, the choice of measured elements can be guided, if the amount of laboratory work is a limiting factor. Further interpretation from an ecological point of view is a subject of future research.

Acknowledgments

The authors would like to express their gratitude to the Finnish Forest Research Institute, Parkano Research Station for providing the data and especially Dr. Sebastiaan Luyssaert and Dr. Hannu Raitio for fruitful collaboration. We would also like to thank the reviewers for their valuable comments.

Appendix A. Constructed LARS models

The LARS models for spruce obtained using the model selection procedure described in Section 3.1 are as follows:

$$\begin{aligned}\mathrm{Al}_{t,C+1} &= 0.48\mathrm{Al}_{t-1,C} - 0.35\mathrm{Ca}_{t-1,C} - 0.14\mathrm{Cu}_{t-1,C} \\ &\quad + 0.30\mathrm{Mn}_{t-1,C} + 0.13\mathrm{NM}_{t-1,C} \\ &\quad - 0.31Y - 0.11X - 0.14\mathrm{TD}_t,\end{aligned}$$

$$\mathrm{B}_{t,C+1} = 0.80\mathrm{B}_{t-1,C} + 0.19\mathrm{Mn}_{t-1,C} - 0.14\mathrm{S}_{t-1,C},$$

$$\mathrm{Ca}_{t,C+1} = 0.86\mathrm{Ca}_{t-1,C},$$

$$\mathrm{Cu}_{t,C+1} = 0.22\mathrm{NM}_{t-1,C},$$

$$\mathrm{Fe}_{t,C+1} = 0.68\mathrm{ST}_t - 0.28\mathrm{TD}_t,$$

$$\begin{aligned}\mathrm{K}_{t,C+1} &= 0.22\mathrm{Fe}_{t-1,C} + 0.57\mathrm{K}_{t-1,C} \\ &\quad + 0.28N_{t-1,C} + 0.28\mathrm{NM}_{t-1,C} \\ &\quad - 0.24\mathrm{S}_{t-1,C} - 0.75Y + 0.21\mathrm{ST}_t \\ &\quad - 1.36\mathrm{TA}_t + 1.16\mathrm{TD}_t + 0.22A_t,\end{aligned}$$

$$\begin{aligned}\mathrm{Mg}_{t,C+1} &= -0.30\mathrm{B}_{t-1,C} - 0.29\mathrm{Ca}_{t-1,C} \\ &\quad - 0.26\mathrm{K}_{t-1,C} + 0.69\mathrm{Mg}_{t-1,C},\end{aligned}$$

$$\mathrm{Mn}_{t,C+1} = 0.95\mathrm{Mn}_{t-1,C},$$

$$\begin{aligned}N_{t,C+1} &= -0.31\mathrm{B}_{t-1,C} + 0.37\mathrm{ST}_t \\ &\quad + 0.30\mathrm{TA}_t + 0.18\mathrm{PD}_t,\end{aligned}$$

$$\mathrm{NM}_{t,C+1} = 0.83\mathrm{NM}_{t-1,C} + 0.20\mathrm{Zn}_{t-1,C},$$

$$\begin{aligned}\mathrm{P}_{t,C+1} &= -0.13\mathrm{K}_{t-1,C} + 0.80\mathrm{P}_{t-1,C} - 0.20\mathrm{S}_{t-1,C} \\ &\quad + 0.45\mathrm{ST}_t - 0.12\mathrm{PT}_t,\end{aligned}$$

$$\mathrm{S}_{t,C+1} = 0.55\mathrm{S}_{t-1,C} + 0.49\mathrm{ST}_t - 0.25\mathrm{TA}_t + 0.16\mathrm{PD}_t,$$

$$\mathrm{Zn}_{t,C+1} = 0.78\mathrm{Zn}_{t-1,C} + 0.19A_t.$$

For pine the obtained LARS models are:

$$\begin{aligned}\mathrm{Al}_{t,C+1} &= 0.67\mathrm{Al}_{t-1,C} + 0.12\mathrm{Cu}_{t-1,C} - 0.11\mathrm{Mg}_{t-1,C} \\ &\quad + 0.17Y + 0.08X - 0.25A_t,\end{aligned}$$

$$\mathrm{B}_{t,C+1} = 0.49\mathrm{B}_{t-1,C} - 0.19\mathrm{Mn}_{t-1,C} - 0.27Y,$$

$$\mathrm{Ca}_{t,C+1} = 0.79\mathrm{Ca}_{t-1,C} - 0.24\mathrm{Zn}_{t-1,C} + 0.17\mathrm{ST}_t,$$

$$\begin{aligned}\mathrm{Cu}_{t,C+1} &= 0.29\mathrm{Cu}_{t-1,C} + 0.33N_{t-1,C} \\ &\quad - 0.34\mathrm{NM}_{t-1,C} - 0.18X + 0.21\mathrm{ST}_t,\end{aligned}$$

$$\begin{aligned}\mathrm{Fe}_{t,C+1} &= 0.61\mathrm{Fe}_{t-1,C} - 0.15\mathrm{K}_{t-1,C} - 0.19Y \\ &\quad + 0.22\mathrm{ST}_t - 0.21\mathrm{TD}_t - 0.11A_t,\end{aligned}$$

$$\mathrm{K}_{t,C+1} = 0.40\mathrm{P}_{t-1,C} + 0.22\mathrm{PT}_t,$$

$$\begin{aligned}\mathrm{Mg}_{t,C+1} &= 0.20\mathrm{Ca}_{t-1,C} + 0.07\mathrm{Fe}_{t-1,C} - 0.17\mathrm{K}_{t-1,C} \\ &\quad + 0.53\mathrm{Mg}_{t-1,C} - 0.30\mathrm{Mn}_{t-1,C} \\ &\quad - 0.28\mathrm{NM}_{t-1,C} - 0.24\mathrm{Zn}_{t-1,C} - 0.25X \\ &\quad + 0.17\mathrm{ST}_t - 0.37\mathrm{TA}_t + 0.13\mathrm{PD}_t,\end{aligned}$$

$$\mathrm{Mn}_{t,C+1} = -0.18\mathrm{Mg}_{t-1,C} + 0.77\mathrm{Mn}_{t-1,C} + 0.14\mathrm{NT}_t,$$

$$N_{t,C+1} = 0.26\mathrm{Al}_{t-1,C} - 0.27\mathrm{Mg}_{t-1,C} - 0.17\mathrm{Mn}_{t-1,C} + 0.36N_{t-1,C} + 0.29\mathrm{ST}_t + 0.45\mathrm{TA}_t + 0.10\mathrm{PD}_t,$$

$$\mathrm{NM}_{t,C+1} = 0.14N_{t-1,C} + 0.76\mathrm{NM}_{t-1,C} - 0.10A_t,$$

$$\mathrm{P}_{t,C+1} = -0.24\mathrm{K}_{t-1,C} - 0.22\mathrm{Mg}_{t-1,C} - 0.13\mathrm{NM}_{t-1,C} + 0.78\mathrm{P}_{t-1,C} - 0.13\mathrm{TD}_t + 0.13\mathrm{PD}_t - 0.28A_t,$$

$$\mathrm{S}_{t,C+1} = -0.38\mathrm{K}_{t-1,C} - 0.37\mathrm{NM}_{t-1,C} + 0.67\mathrm{P}_{t-1,C} - 0.41Y + 0.41\mathrm{ST}_t + 0.11\mathrm{PD}_t,$$

$$\mathrm{Zn}_{t,C+1} = -0.27\mathrm{Mg}_{t-1,C} + 0.69\mathrm{Zn}_{t-1,C} - 0.20X + 0.18\mathrm{ST}_t + 0.09A_t$$

Appendix B. Tables

Proportions of the LARS models obtained using cross-validation for spruce and pine are shown in Tables B.1 and B.2, respectively.

Table B.1
Proportions of the LARS models obtained using cross-validation for spruce that contained the different regressors

Y_i	Al	B	Ca	Cu	Fe	K	Mg	Mn	N	NM	P	S	Zn
Al	**1.00**	–	0.01	–	0.26	0.15	0.01	–	0.07	–	–	–	–
B	–	**1.00**	0.09	–	–	0.62	**1.00**	–	**0.92**	0.01	–	–	–
Ca	**0.97**	–	**1.00**	–	–	0.09	**1.00**	–	–	–	–	–	–
Cu	**0.38**	–	–	–	–	0.07	–	–	–	0.01	–	–	–
Fe	–	–	–	0.01	–	**0.40**	–	–	–	–	–	–	–
K	–	0.01	–	–	–	**1.00**	**1.00**	–	–	–	**0.91**	–	–
Mg	–	–	–	–	0.01	0.01	**1.00**	–	0.30	–	–	–	–
Mn	**0.95**	**0.77**	–	–	0.10	0.02	–	**1.00**	0.03	–	–	–	–
N	–	–	–	0.07	–	**0.97**	–	–	0.14	–	–	–	–
NM	**0.30**	–	–	**0.29**	0.05	**1.00**	–	–	–	**1.00**	0.08	0.04	–
P	–	–	–	–	–	0.01	0.12	–	0.03	–	**1.00**	–	0.23
S	–	**0.83**	0.07	–	0.14	**0.17**	–	0.04	–	–	**0.81**	**1.00**	0.07
Zn	0.03	–	0.04	–	0.07	–	–	–	0.14	**0.99**	–	0.04	**1.00**
Y	**1.00**	0.32	–	–	0.01	**0.28**	–	0.10	0.09	0.06	–	–	–
X	**0.62**	0.05	0.01	0.01	0.10	0.01	0.03	0.09	–	–	0.02	–	0.12
NT	–	–	–	–	0.01	–	–	–	0.09	–	0.01	–	–
ST	0.01	0.04	–	0.40	**1.00**	**0.26**	0.31	0.16	**0.97**	–	**0.98**	**1.00**	0.08
TA	–	0.17	0.10	–	–	**0.31**	0.41	0.01	**0.88**	–	–	**0.65**	–
TD	**0.56**	–	–	–	**0.99**	**0.31**	–	–	–	–	0.01	0.01	–
PT	0.09	–	–	–	0.01	0.04	–	0.17	0.06	–	**0.82**	–	–
PD	–	0.20	–	–	0.01	0.02	–	–	**0.92**	–	–	**0.62**	–
A	0.06	–	0.18	0.60	–	**0.20**	–	0.01	0.32	–	–	0.01	**1.00**

"–" denotes that the regressor was never selected in the model during cross-validation. Regressors that were selected in the final model are indicated by bold face.

Table B.2
Proportions of the LARS models obtained using cross-validation for pine that contained the different regressors

Y_i	Al	B	Ca	Cu	Fe	K	Mg	Mn	N	NM	P	S	Zn
Al	**1.00**	0.04	–	0.59	0.06	–	0.19	–	**0.99**	0.04	–	–	–
B	–	**1.00**	–	0.06	–	–	–	0.27	–	0.01	–	–	–
Ca	0.01	0.02	**1.00**	–	0.03	0.16	**0.80**	–	0.07	–	0.35	–	–
Cu	**0.99**	–	–	**1.00**	0.01	0.04	–	–	0.04	–	0.06	–	0.07
Fe	0.06	–	–	0.02	**1.00**	–	**0.40**	–	0.04	–	0.29	0.14	–
K	0.05	–	0.12	0.03	**0.98**	0.46	**0.39**	–	0.32	–	**0.83**	**0.98**	–
Mg	**0.78**	–	–	0.20	–	0.29	**1.00**	**0.60**	**1.00**	–	**0.99**	0.04	**0.99**
Mn	–	**0.54**	0.28	–	–	0.04	**1.00**	**1.00**	**0.92**	–	–	0.03	–
N	–	–	–	**1.00**	0.01	0.01	0.01	–	**1.00**	**0.66**	–	0.01	–
NM	0.26	0.01	–	**1.00**	0.17	0.12	**0.99**	–	0.20	**1.00**	**0.70**	**1.00**	–
P	–	–	0.04	–	0.01	**1.00**	0.20	–	0.06	–	**1.00**	**1.00**	–
S	–	–	–	–	–	–	–	–	–	–	0.10	0.29	–
Zn	–	–	**0.80**	–	–	0.19	**0.42**	0.39	0.03	0.04	–	0.01	**1.00**
Y	**0.95**	**0.99**	–	–	**0.76**	0.01	0.16	–	0.01	–	0.03	**0.94**	–
X	**0.47**	–	0.03	**0.96**	0.02	0.27	**0.60**	–	–	–	0.21	0.03	**1.00**
NT	0.01	–	0.03	–	0.04	–	0.01	**0.91**	–	–	–	0.02	–
ST	–	0.20	**0.55**	**0.97**	**0.97**	–	**0.84**	–	**1.00**	0.01	–	**1.00**	**0.92**
TA	–	–	0.01	–	0.01	–	**0.52**	–	**1.00**	–	–	–	–
TD	0.10	0.17	0.01	0.04	**1.00**	–	–	–	–	–	**0.80**	0.01	0.01
PT	–	0.06	–	–	0.46	**0.64**	–	–	–	0.01	–	0.01	–
PD	0.01	0.01	0.01	–	0.03	0.01	**0.43**	–	**0.41**	0.01	**0.79**	**0.64**	–
A	**1.00**	–	0.07	–	**0.71**	0.01	0.01	0.01	–	**0.43**	**1.00**	–	**0.46**

"–" denotes that the regressor was never selected in the model during cross-validation. Regressors that were selected in the final model are indicated by bold face.

References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, New York.

Breiman, L., 1992. The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. J. Am. Statist. Assoc. 87 (419), 738–754.

Breiman, L., 1995. Better subset regression using the nonnegative garrotte. Technometrics 37 (4), 373–384.

Breiman, L., Friedman, J.H., 1997. Predicting multivariate responses in multivariate regression. J. R. Statist. Soc. B 59 (1), 3–54.

Copas, J.B., 1983. Regression, prediction and shrinkage. J. R. Statist. Soc. B 45 (3), 311–354.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 19 (1), 1–67.

Good, P., 2000. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer Series in Statistics, 2nd ed. Springer Verlag.

Guthrie, W., Filliben, J., Heckert, A. 2005. Process modeling. NIST/SEMATECH e-Handbook of Statistical Methods. National Institute of Standards and Technology (Chapter 4). http://www.itl.nist.gov/div898/handbook/.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1995. Multivariate Data Analysis, 5th ed. Prentice-Hall.

Hansen, M.H., Yu, B., 2001. Model selection and the principle of minimum description length. J. Am. Statist. Assoc. 96 (454), 746–774.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics. Springer Verlag.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Luyssaert, S., Raitio, H., Fürst, A., 2003. Forest nutrition on the Finnish and Austrian Level I plots in 1987–2000. Forest Condition in Europe—Results of the 2002 Large-scale Survey. UN-ECE, EC, Geneva, Brussels pp. 72–83 (Chapter 5).

Luyssaert, S., Sulkava, M., Raitio, H., Hollmén, J., 2004. Evaluation of forest nutrition based on large-scale foliar surveys: are nutrition profiles the way of the future?. J. Environ. Monit. 6 (2), 160–167.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15 (4), 661–675.

Stefan, K., Fürst, A., Hacker, R., Bartels, U., 1997. Forest foliar condition in Europe—results of large-scale foliar chemistry surveys 1995. Technical Report. EC, UN/ECE, Brussels, Geneva.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58 (1), 267–288.