SILVA FENNICA

Silva Fennica 46(2) research articles

www.metla.fi/silvafennica · ISSN 0037-5330
The Finnish Society of Forest Science · The Finnish Forest Research Institute

Performance Modelling in Forest Operations Through Partial Least Square Regression

Corrado Costa, Paolo Menesatti and Raffaele Spinelli

Costa, C., Menesatti, P. & Spinelli, R. 2012. Performance modelling in forest operations through partial least square regression. Silva Fennica 46(2): 241–252.

Partial Least Square (PLS) regression is a recent soft-modelling technique that generalizes and combines features from principal component analysis (PCA) and multiple regression. It is particularly useful when predicting one or more dependent variables from a large set of independent variables, often collinear. The authors compared the potential of PLS regression and ordinary linear regression for accurate modelling of forest work, with special reference to wood chipping, wood extraction and the continuous harvesting of short rotation coppice. Compared to linear regression, PLS regression allowed producing models that better fit the original data. What is more, it allowed handling collinear variables, facilitating the extraction of sound models from large amounts of field data obtained from commercial forest operations. On the other hand, PLS regression analysis is not as easy to conduct, and produces models that are less user-friendly. By producing alternative models, PLS regression may provide additional – and not alternative – ways of reading the data. Ideally, a comprehensive data analysis could include both ordinary and PLS regression and proceed from their results in order to get a better understanding of the phenomenon under examination. Furthermore, the computational complexity of PLS regression may stimulate interdisciplinary team-building, to the greater benefit of scientific research within the field of forest operations.

Keywords chipping, harvesting, productivity, skidding
Addresses Costa & Menesatti: CRA ING, Monterotondo Scalo (Roma), Italy; Spinelli: CNR IVALSA, Sesto Fiorentino (FI), Italy
E-mail spinelli@ivalsa.cnr.it, corrado.costa@entecra.it, paolo.menesatti@entecra.it
Received 16 January 2012 **Revised** 19 March 2012 **Accepted** 21 March 2012
Available at http://www.metla.fi/silvafennica/full/sf46/sf462241.pdf

1 Introduction

Performance studies in forest operations often produce empirical models used for many purposes, including wood-flow planning, harvesting cost calculation and work rate setting (Björheden 1988). At a more fundamental level, performance studies also allow understanding the behaviour of harvesting machines and systems under varying stand and terrain conditions (Visser and Spinelli 2011). That is particularly important when deploying specialised industrial technology (Chiorescu and Grönlund 2001), which is more expensive and less flexible than traditional general-purpose equipment (Spinelli and Magagnotti 2011a).

Empirical performance models are generally developed by collecting field data and testing the statistical significance of any relationships with regression analysis (Samset 1990). The most commonly used regression type is ordinary least square linear regression (OLS). This technique is used to "calculate" an equation capable of representing the relationship between a dependent variable (typically time consumption or productivity) and one or more independent variables (Bergstrand 1987). Indicator (Dummy) variables are often used to include influencing factors that assume discrete rather than continuous values (Olsen et al. 1998).

A fundamental assumption of ordinary least square linear regression is that variables are independent, and not collinear (Freedman et al. 2007). That can be obtained through a strict experimental design, carefully planned before data collection and eventually integrated as work proceeds (Howard 1989). However, a large number of variables can impact the performance of forest machines, including piece size (Nakagawa et al. 2010), stocking density and thinning intensity (Eliasson 1999), type of cut and total volume (Suadicani and Fjeld 2001) and terrain characteristics (Visser and Stampfer 1998). Further variation is introduced by the widely varying skills of both machine operators (Ovaskainen et al. 2004) and researchers (Nuutinen et al. 2008). To overcome such variation, productivity models should be based on large samples (Nurminen et al. 2006). Bergstrand (1987) estimates that about 400 operators should be included in each performance study, in order to detect the existence of differences between groups at a 95% confidence level.

That explains why it is so difficult and expensive to implement a strict experimental design when developing an empirical performance model (Spinelli et al. 2011). The large samples needed to obtain a reliable general model are often assembled by studying commercial operations, which makes it difficult to implement a controlled study design (Spinelli and Magagnotti 2009). As a result, variables are often collinear and most such studies can estimate primary effects only, while missing secondary effects (Spinelli et al. 2010).

Hence the interest in exploring alternatives to ordinary linear regression (OLS), such as multivariate predictive modelling based on the recombination of principal components (Principal Component Regression – PCR) or latent variables (Partial Least Square – PLS). Different authors (Nsofor, 2006) observed that in many cases the PLS approach returns better results than PCR, including a better goodness-of-fit and a more robust model. PCR is a multivariate method where a multiple linear regression is performed on the Principal Component Analysis scores. In contrast, PLS is a soft-modelling technique, i.e. it has "soft" distributional assumptions (Pulos and Rogness 1995) and can be used when distributions are highly skewed (Bagozzi and Yi 1994). PLS finds a linear regression model by projecting the predicted variables and the observed variables to a new space (projection to latent structures) that is component-based rather than covariance-based. PLS is particularly useful when predicting one or more dependent variables from a large set of independent variables, often collinear. This technique originated within the field of economics (Wold 1966) but became popular first in computational chemistry (Geladi and Kowalski 1986) and then in sensory evaluation (Martens and Naes 1989). Today PLS regression is becoming a tool of choice in the social sciences, as a multivariate technique for non-experimental and experimental data alike (Mcintosh et al. 1996, Costa et al. 2010, Capoccioni et al. 2011). PLS regression was first presented as an algorithm akin to the power method used for computing eigenvectors and was rapidly interpreted in a statistical frame-

work (Frank and Friedman 1993, Helland 1990, Hoskuldsson 1988, Abdi 2003).

The goal of this study was to explore the potential of multivariate approaches different from OLS (i.e. PCR and PLS), focusing mainly on PLS regression when developing forest operation performance models. In particular, the study aimed at comparing the main statistical significance indicators associated to models calculated with OLS, PCR and PLS regression from the same original datasets, for the purpose of quantifying the eventual improvements obtained with the new techniques.

2 Materials and Methods

Three datasets were selected for comparing the three regression techniques: OLS, PCR and PLS. These datasets represented a wide variety of forest operations, with clearly different characteristics and influencing factors. The same datasets had already been used for published modelling studies. Dataset 1 concerned chipping whole trees, logs and forest residues with mobile chippers, and was used to estimate chipping time in min ton^{-1} as a function of eleven independent variables (Spinelli and Hartsough 2001). Dataset 2 concerned the skidding of whole trees, delimbed stems and logs with forestry-fitted farm tractors, and was used to estimate productivity in m^3 $hour^{-1}$ as a function of ten independent variables (Spinelli and Magagnotti 2011b). Dataset 3 concerned harvesting short rotation coppice with modified foragers, and was used to estimate the forager harvesting rate in min km^{-1} as a function of five independent variables (Spinelli et al. 2009). The complete list of dependent and independent variables is shown in Table 1.

Table 1. Variables used for the regression analysis.

Dataset 1
Y-block variables:
– Chipping time (min t^{-1})
X-block variables:
– Species (Austrian pine, Beech, Chestnut, Douglas, Eucalyptus, Hardwood, Maritime pine, Pine, Poplar, Radiata, Robinia, Spruce, Umbrella pine)
– Material (Tops, Logs, Slash, Whole, Complete)
– Wood (Fresh, Semi-dry, Dry)
– Piece size (ton $piece^{-1}$)
– Lay-out (Loads, Bunched, Stacked, Aligned)
– Type (Disc, Drum)
– Power (kW)
– Feeding (Crane, Hand-fed)
– Chipping (Landing, Terrain)
– Operator (Top prof, Prof, Full time prof, Part time prof, Beginners)

Dataset 2
Y-block variables:
– Skidding productivity (m^3 $hour^{-1}$)
X-block variables
– Power (kW)
– Operators (n)
– Chokerman (With, Wthout)
– Distance (m)
– Winching Distance (m)
– Pieces (n $load^{-1}$)
– Load size (m^3)
– Piece Size (m^3)
– Treatment (Maturity, Thinning)
– Suspension (Full, Half)

Dataset 3
Y-block variables:
– Harvest rate: (min km^{-1})
X-block variables:
– Stocking (t ha^{-1})
– Forager Power (kW)
– Header (HS2, GBE)
– Row System (twin-, single-row)
– Stocking (t km^{-1})

Note: underlined variables are also significant to the model obtained through ordinary regression

In order to determine the most robust PCR and PLS models in terms of reducing the overfitting in prediction, each dataset was partitioned into 80% to estimate the model, and 20% for the independent validation tests. The partitioning strategy is one of the most reliable and advanced approaches to validate models and correct overfitting, and is directly linked with model robusteness. The partitioning algorithm used was SPXY (Harrop Galvao et al. 2005, Antonucci et al. 2011). This algorithm accounts for the variability of both the dependent and independent variables, constituting the Y-block and the X-block, respectively. This procedure was not used for the ordinary models, which had been previously published as calculated from the whole data set without any partitioning. Hence, calculating them again after partitioning would have generated a result inconsistent with the published formulations.

For the purpose of both PCR and PLS regression analyses, the X-blocks from Datasets 1 and 2 were transformed by column centering ('mean center') procedure, while Dataset 3 was transformed by column normalization ('autoscale' equal to mean centering * stand dev^{-1}). All the Y-blocks were transformed using the 'autoscale' procedure.

PCR is a three-step multivariate method: in the first step, a Principal Component Analysis (PCA) of the data matrix is performed and measured variables are converted into new ones (scores on latent variables). In the second step the Principal Components relevant in the prediction model are selected on the base of the highest goodness of fit in the validation phase. This is followed by a multiple linear regression (3rd step) between the scores obtained in the PCA (1st step) and the characteristic response variable to be modelled (De Maesschalck et al. 1999), Because it directly addresses the collinearity problem, PCR can be said to be less susceptible to overfitting than Multiple Linear Regression (MLR).

PLS is used to find the fundamental relations between two matrix (X and Y) and represents a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.

The general underlying model of multivariate PLS is:

$$X = TP^T + E \quad (1)$$

$$Y = UQ^T + F \quad (2)$$

where X is a $n \times$ m matrix of predictors, Y is a $n \times p$ matrix of responses; T and U are $n \times l$ matrices that are, respectively, projections of X (the *X score*, component or factor matrix) and projections of Y (the *Y scores*); P and Q are, respectively, m x l and $p \times l$ orthogonal loading matrices; and matrices E and F are the error terms, assumed to be i.i.d. normal. The decompositions of X and Y are made so as to maximize the covariance of T and U.

A number of variants of PLS exist for estimating the factor and loading matrices T, P and Q. In this study we used the SIMPLS (De Jong 1993) algorithm (equal to PLS1 for univariate y) that constructs estimates of the linear regression between X and Y as (B and B_0 are parameters):

$$Y = XB + B_0 \quad (3)$$

Both PCR and PLS were computed using PLS toolbox 6.2 (Eigenvector research) for Matlab 7.1. The programme also calculated residual error indicators, such as the root mean square errors in calibration (RMSEC) and in validation (RMSECV). The predictive ability of the model was partially dependent on the number of the latent vectors used and was assessed through the following statistical indicators: root mean square error (RMSE), standard error of prevision (SEP), correlation coefficient (r) and bias. Finally, the programme calculated the ratio of percentage deviation (RPD), which is the ratio of the standard deviation of the measured data to the RMSE (Williams 1987). This represents the factor by which the prediction accuracy has been increased compared with using the mean of the original data. Generally, a good predictive model should have high values for r and low values for RMSE, SEP and bias. The model chosen was for the number of LV (Latent Vector) that yielded the highest r, minimum SEP for predicted and known Y-block and maximum RPD.

RPD values were classified as follows: RPD < 1.0 for a very poor model, whose use is not recommended; RPD between 1.0 and 1.4 for a poor model where only high and low values can be separated; RPD between 1.4 and 1.8 for a fair model that may be used for assessment and correlation; RPD values between 1.8 and 2.0 for a good model able to produce quantitative predictions; RPD between 2.0 and 2.5 for a very good quantitative model, and RPD > 2.5 for an excellent model, highly accurate and reliable (Viscarra Rossel et al. 2007).The main differences between OLS, PCR and PLS are summarized in Table 2.

3 Results

The models generated through ordinary regression analysis are available on the quoted original publications, and namely: Spinelli and Hartsough (2001), Spinelli and Magagnotti (2011b)

Table 2. Principal features of OLS, PCR and PLS modelling techniques (modified from Nsfor 2006).

OLS	PCR	PLS
No standardization or scaling required	Standardization or scaling needed	Standardization or scaling needed
Gives good predictions when the inputs variables are truly independent	Predicts better when input variables are not independent of each other	Predicts better when input variables are not independent of each other
Good when the input variables are all useful in predicting the response	Works well when there is a need for variable reduction	Works well when there is a need for variable reduction
Simpler to understand and interpret	More complex in its solutions	More complex in its solutions
Does not handle well ill-conditioned or collinear data	Works well with ill-conditioned or collinear data	Works well with ill-conditioned or collinear data
Does not handle well redundant input variablest	Removes collinearity	Removes collinearity
Maximizes the squared correlation between projected inputs and output	Better for dimensionality reduction or feature selection	Better for dimensionality reduction or feature selection
	Maximizes variance of the projected inputs	Maximizes the covariance between projected inputs and output
	Considers only input variables in their transformations	Considers both input and output variables in their transformations

and Spinelli et al. (2009), for Datasets 1, 2 and 3, respectively. On the other hand, the models generated through PCR and PLS regression analyses are rather complex to write and will not be reported.

Table 3 shows the main indicators for the OLS, PCR and PLS regression models, for the three datasets tested. The number of independent variables used by PCR and PLS regressions are from 2 to 5 times higher than used in OLS. Model error indicators (SEP and RMSE) are 20 to 40% lower for the PCR and PLS regression models, compared to the OLS ones. Moreover, *r* values are always higher for both PCR and PLS models, with an increment between 5 and 20% over OLS models. RPD is always higher for the PLS regression models. Based on the previously mentioned RPD classification, PLS regression allows the systematic upgrading of ordinary regression models: Model 1 goes from "very good" to "excellent", Model 2 from "fair" to "very good" and Model 3 from "poor" to "fair". The indicators for the validation tests are also encouraging, with the predicted values following quite closely the actual values in the subset reserved for independent validation.

PLS is the best performing model for Dataset 1 while for Datasets 2 and 3, with a reduced number of X variables, PCR and PLS converged to the same results.

The observed *vs* predicted independent Y variables for the OLS and PLS models for the three datasets were reported in Figs. 1, 2 and 3, respectively.

Table 4 shows the relative contribution (loadings) of individual X-variables to each of the first three latent vectors of each PLS model.

Regarding Dataset 1, the variables with the highest contribution to the first LV (*x*-block 99.97%; *y*-block 0.08%) are the indicators for

Table 3. Main goodness-of-fit indicators for the regression models.

Regression analysis	Dataset 1			Dataset 2			Dataset 3		
	OLS	PCR	PLS	OLS	PCR	PLS	OLS	PCR	PLS
Observations (n)	99	99	99	324	324	324	480	480	480
X Variables (n)	2	38	38	5	13	13	2	7	7
Latent Vectors / PC axes (n)	–	11	4	–	9	9	–	5	5
% Cumulated Variance X-block	–	99.99	31.40	–	100	100	–	100	100
% Cumulated Variance Y-block	–	72.67	86.17	–	78.28	78.27	–	63.90	63.90
RMSEC	–	0.52	0.33	–	0.47	0.47	–	0.60	0.60
RMSECV	–	0.65	0.55	–	0.48	0.48	–	0.61	0.61
r model (80% of observations)	0.889	0.870	0.942	0.828	0.880	0.884	0.655	0.800	0.799
SEP model (80% of observations)	5.462	5.024	4.327	3.773	0.820	0.820	17.743	3.143	3.143
RMSE model (80% of observations)	5.462	5.000	4.300	1.224	0.818	0.818	5.240	3.140	3.139
RPD model (80% of observations)	2.177	1.982	2.961	1.776	2.141	2.141	1.270	1.662	1.662
r test (20% of observations)	–	0.859	0.917	–	0.822	0.823	–	0.656	0.656
SEP test (20% of observations)	–	9.452	5.252	–	1.755	1.755	–	6.535	6.535
RMSE test (20% of observations)	–	11.408	7.811	–	2.330	2.331	–	9.582	9.582
RPD test (20% of observations)	–	1.680	0.613	–	1.745	1.745	–	1.299	1.299

Feeding (crane or hand-fed) and Material (tops), and Power. Lay out (aligned) and Operator (part-time professional) give the highest contribution to the second LV (*x*-block 0.01%; *y*-block 70.45%); Lay-out (loads), and Species (Poplar) has the strongest effect on the third LV (*x*-block <0.01%; *y*-block 15.75%). Of these variables, only one is included in the model obtained with OLS. On the other hand, the PLS regression model is not particularly sensitive to piece size, which is a key independent variable for the ordinary regression model. In the PLS regression model, piece characteristics are reflected by attributes other than size (Lay-out and Species).

As to Dataset 2, the variable with the highest contribution on the first LV is distance (*x*-block 58.16%; *y*-block 3.17%). Power and Distance have the strongest effect on the second LV (*x*-block 41.08%; *y*-block 4.11%). Winching Distance and Pieces give the highest contribution to the third LV (*x*-block 0.71%; *y*-block 18.57%). These are the same variables included in the model calculated with ordinary regression techniques.

Finally, the first LV (*x*-block 76.01%; *y*-block 26.22%) of Dataset 3 receives the highest contribution from Stocking in t ha^{-1} and Power in kW. Power is also the main contributor to the second LV (*x*-block 19.63%; *y*-block 1.96%). Stocking in t ha^{-1} and t km^{-1} give the highest contribution to the third LV (*x*-block 4.00%; *y*-block 3.44%). Power and Stocking are the two main variables included in the model calculated with ordinary regression techniques. PCR loadings were not reported, having similar values than PLS.

4 Discussion

Although variable selection may reduce the error of a prediction model, it may also inadvertently discard useful redundancy. Using fewer variables to make a prediction means that each variable has a larger influence on the final result. Hence, one should carefully consider the requirements of the final model before variable selection. For this reason, we decided to use full-spectrum PCR and PLS models.

As observed by Nsofor (2006) PLS gives better results than PCR for latent vectors that maximize the correlation between LV's and the Y var (Table 2). Moreover, we observed that with fewer variables (Datasets 2 and 3) PCR and PLS models tend to offer the same results.

PLS regression analysis does offer some benefits over ordinary regression analysis (Lipp 1996). The substantial improvement of all goodness-of-fit indicators is probably the most visible benefit. Moreover, other benefits of the PLS regression technique are not merely the increase of a coef-

Table 4. PLS Models: X variable loadings for each of the first 3 LVs (Latent Vectors).

Variable	Dummy opt.	LV1	LV2	LV3
Dataset 1				
Av.size	ton piece^{-1}	–0.243	–0.044	0.092
Power	kW	–0.336	0.142	–0.145
Species	Austrian pine	0.146	–0.143	0.118
Species	Beech	0.000	0.000	0.000
Species	Chestnut	–0.028	–0.052	–0.052
Species	Douglas	–0.018	0.060	0.076
Species	Eucalyptus	0.052	–0.227	–0.057
Species	Hardwood	0.128	0.287	–0.164
Species	Maritime pine	0.072	0.206	–0.005
Species	Pine	0.034	–0.039	–0.194
Species	Poplar	–0.095	–0.255	0.333
Species	Radiata	–0.066	0.002	0.111
Species	Robinia	0.027	–0.116	–0.144
Species	Spruce	–0.042	–0.024	–0.147
Species	Umbrella pine	–0.086	0.159	–0.201
Material	Complete	–0.053	0.051	–0.023
Material	Logs	0.069	–0.139	–0.134
Material	Slash	0.130	0.221	–0.077
Material	Tops	–0.249	0.040	0.329
Material	Whole	0.158	–0.164	–0.224
Wood	Dry	–0.038	–0.120	0.038
Wood	Fresh	0.040	0.054	0.056
Wood	Semi-dry	–0.023	0.004	-0.081
Lay-out	aligned	0.067	-0.335	0.234
Lay-out	bunched	-0.094	0.134	0.167
Lay-out	loads	0.079	0.095	-0.367
Lay-out	stacked	–0.022	0.089	–0.147
Type	Disc	0.111	–0.056	0.235
Type	Drum	–0.111	0.056	–0.235
Feeding	Crane	–0.428	0.098	–0.044
Feeding	Hand-fed	0.428	–0.098	0.044
Chipping	Landing	–0.064	0.267	–0.222
Chipping	Terrain	0.064	–0.267	0.222
Operator	Beginners	0.320	–0.220	0.069
Operator	Full-time prof	0.009	–0.237	–0.002
Operator	Part-time prof	0.240	0.309	0.111
Operator	Prof	–0.005	0.124	–0.107
Operator	Top Prof	–0.249	0.148	0.008
Dataset 2				
Power	kW	0.028	0.781	–0.206
Operators	n	0.000	–0.006	–0.005
Distance	m	1.000	–0.575	–0.001
WinchDist	m	–0.006	–0.201	–0.890
Pieces	n	0.003	–0.136	0.406
Load size	m^3	0.001	0.013	0.009
Piece size	m^3	0.000	0.008	–0.011
Chokerman	With	0.000	–0.006	–0.005
Chokerman	Without	0.000	0.006	0.005
Treatment	Maturity	0.000	–0.006	0.001
Treatment	Thinning	0.000	0.006	–0.001
Suspension	Full	0.001	0.005	–0.006
Suspension	Half	–0.001	–0.005	0.006
Dataset 3				
Stocking	t ha^{-1}	0.823	–0.038	–0.294
Power	kW	0.510	–0.999	0.044
Stocking	t km^{-1}	0.251	0.003	0.882
Header	HS2	0.007	–0.014	–0.009
Header	GBE	–0.007	0.014	0.009
Row System	Twin	–0.004	0.013	0.258
Row System	Single	0.004	–0.013	–0.258

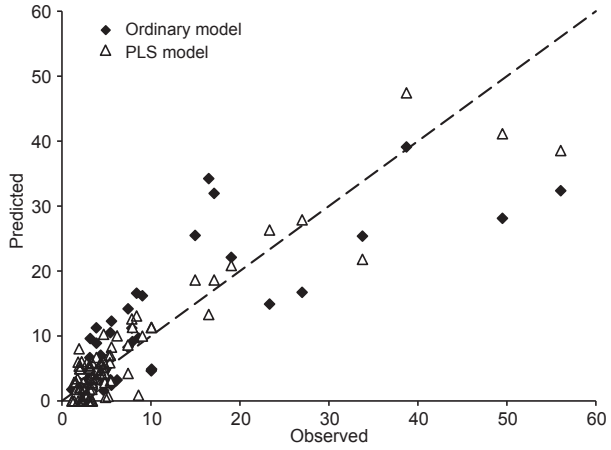

Fig. 1. Dataset 1: observed *vs* predicted Y for the ordinary and PLS model.

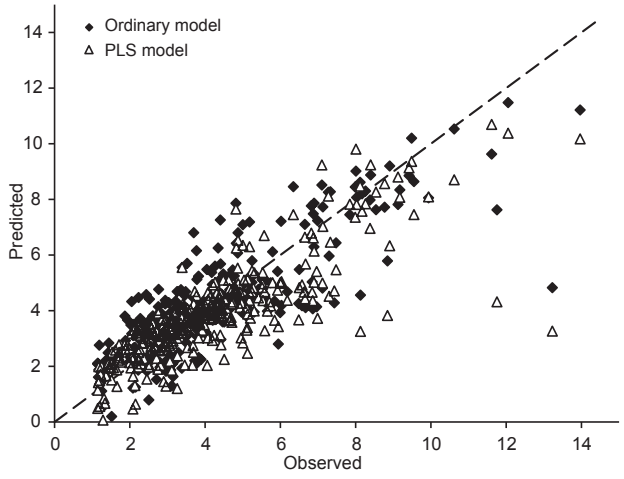

Fig. 2. Dataset 2: observed *vs* predicted Y for the ordinary and PLS model.

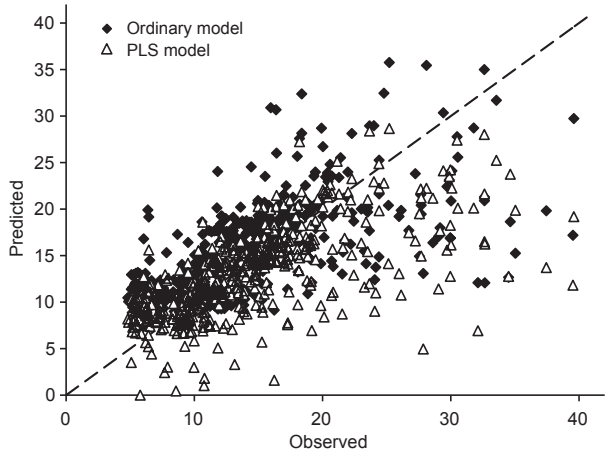

Fig. 3. Dataset 3: observed *vs* predicted Y for the ordinary and PLS model.

ficient, but the capacity of detecting significant variables otherwise missed with ordinary regression techniques (Costa et al. 2009). This is the advantage of latent vectors, which are capable of integrating the effect of more independent variables. A further advantage of PLS regression over multiple linear regression is in the definition of the new variables, which takes into account not only the values assumed by the X but also their correlation with the dependent variables (Kresta 1992).

In this respect, it is most interesting to compare the *X*-variables included in the ordinary and PLS regression models obtained from the same datasets. In most cases (Datasets 2 and 3) the balance remains the same: the strongest variables in the ordinary regression model are also the strongest in the PLS regression model. Hence, PLS regression may have the capacity of drawing additional variables into the models, without radically changing its conceptual structure. That is most logical, because both model types still describe one single real-life phenomenon, and the phenomenon is bound to be the driver, not the model. The model describes the phenomenon, and regardless of how it does that, skidding still involves a machine dragging a load over a certain distance. Hence, machine pulling power, load size and distance are bound to have the strongest effect on skidding performance.

On the other hand, the same event can be seen from different angles, and different observers can choose different attributes to describe the same quality. That may explain why the PCR and PLS regression models underestimated the effect of piece size, which the ordinary regression model picked as one of its strongest independent variables. In contrast, PLS regression analysis selected other piece attributes than size. Hence, the new technique still detected the strong effect of piece characteristics, but chose different specific attributes for inclusion into the model. That is likely dependent on the capacity of PLS regression to handle collinear variables. Ordinary regression would pick one or the other, but the use of latent vectors in PLS regression make it possible to select more than one attribute for the same characteristic, after weighing their contribution through pre-treatment.

When different variables are picked by different models, it is difficult to decide which model best represents the real phenomenon. Direct experience with the phenomenon and convenience should be the best guides, but they are highly subjective. In the specific case of Dataset 1, the choice would be between Size (ordinary regression model) and Species combined with Layout (PLS regression model). There are good reasons for defending both models. The effect of piece size on productivity is generalized and well known (Visser and Spinelli 2011). On the other hand, operator experience often hints at raw material lay-out as a main driver of chipping productivity. The distinctive effect of a given tree species can be related to different wood characteristics. In our case, poplar wood is indeed the softest wood type among those represented in the dataset. It can be debated that a model electing size over lay-out and species is somewhat more flexible, as it may adapt to a wider number of situations. On the other hand, flexibility may tempt users into extrapolation, whereas a model is properly used only within the range set by the original data pool.

The larger number of *X*-variables included in the PCR and PLS regression models also warrants some comments. While this larger number guarantees a more accurate description of the phenomenon, it also requires a larger effort when gathering input data. Hence, PCR and PLS regression models may be less convenient to use than similar models calculated through ordinary regression. Furthermore, users may be somewhat less careful when collecting many input variables, than when they need to collect fewer. Pressed by time constraints, they may settle for approximate values, rather than going all the way and get accurate representative figures. In that case, the alternative is between using fewer better figures or more approximate figures. Therefore, the larger modelling effort required by PLS regression analysis may be frustrated.

PCR and PLS regression analyses are not as easy to perform as OLS. The latter is easily available within any mainstream software package, including the basic Excel. More sophisticated users may scorn the base Excel package and turn to R, or to any commercial statistical softwares – all of which rightly include comprehensive linear regression programmes. All researchers are familiar with ordinary least square regression analysis, and can quickly adopt the results pub-

lished by their colleagues. In contrast, PCR and PLS regression analyses require specific packages, algorithms and skills that are not as readily available. The models themselves are somewhat less handy than standard regression equations. Nevertheless, PLS modelling and more in general the advanced multivariate approach, are getting increasingly popular, because they are very robust and are particularly suitable for modelling complex systems.

This very same reasons may justify the introduction of multivariate regression to forest work science. If its merits turned out to be so valuable, PLS regression would spread rapidly, and the sector would evolve from an older established technique to a new one – as it has already happened before, when regression analysis was first introduced. At the moment, the most practical thing to do for accessing PLS regression is probably to team with researchers who already use it, building multidisciplinary work groups. This way, one may multiply the comparisons, and decide if, when and how PLS regression should replace ordinary least square regression.

5 Conclusions

Compared to OLS analysis, PCR and PLS regression analyses allow producing models that better fit the original data. What is more, they allow handling collinear variables, facilitating the extraction of sound models from large amounts of field data obtained from commercial forest operations. This could lead to more robust models in terms of both variable oscillations and higher repeatability.

On the other hand, PCR and PLS regression analyses are not as easy to conduct, and produce models that are less user-friendly.

In fact, we believe that PCR and PLS regression analyses offer significant benefits in terms of theory-building, and that these benefits may far outweigh the strictly practical ones. By producing alternative models, PCR and PLS regression may provide additional – and not alternative – ways of reading the data. Ideally, the analysis could include ordinary, PCR and PLS regression and proceed from their results in order to get a better understanding of the phenomenon under examination. By comparing the ways and the variables used by both analyses to mirror the actual phenomenon, researcher could get a better understanding of it, which is the ultimate goal of any field study.

Furthermore, the computational complexity of PCR and PLS regression may stimulate interdisciplinary team-building, to the greater benefit of scientific research within the field of forest operations. Cross-pollination could generate new ideas, improve study methods and eventually accelerate scientific progress in this field.

References

Abdi, H. 2003. Partial Least Squares (PLS) regression. In: Lewis-Beck M., Bryman, A. & Futing T. (eds.). Encyclopedia of social sciences research methods. Thousand Oaks (CA – USA).

Antonucci, F., Menesatti, P., Holdenb, N., Canali, E., Giorgi, S., Maienza, A. & Stazi, S. 2011. Hyperspectral visible and near infrared determination of copper concentration in agricultural polluted soils. Communications in Soil Science and Plant Analysis. (In press).

Bagozzi, R. & Yi, Y. 1994. Advanced topics in structural equation models. In: Bagozzi, R. (ed.). Advanced methods of marketing research. Blackwell, Oxford. 51 p.

Bergstrand, K. 1987. Planning and analysis of time studies on forest technology. The Forest Operations Institute of Sweden. Report 17. 58 p.

Björheden, R. 1988. New work-study methods help to decide processing technique in logging. Scandanavian Journal of Forest Research 3: 569–574.

Capoccioni, F., Costa, C., Aguzzi, J., Menesatti, P., Lombarte, A. & Ciccotti, E. 2011. Ontogenetic and environmental effects on otolith shape variability in three Mediterranean European eel (Anguilla anguilla, L.) populations. Journal of Experimental Marine Biology and Ecology 397: 1–7.

Chiorescu, S. & Grönlund, A. 2001. Assessing the role of the harvester within the forestry-wood chain. Forest Products Journal 51: 77–84.

Costa, C., Angelini, C., Scardi, M., Menesatti, P. & Utzeri, C. 2009. Using image analysis on the ventral colour pattern in Salamandrina perspicillata (Savi, 1821) (Amphibia, Salamandridae) to dis-

criminate among populations. Biological Journal of the Linnean Society 96: 35–43.

— , Vandeputte, M., Antonucci, F., Boglione, C., Menesatti, P., Cenadelli, S., Parati, K., Chavanne, H. & Chatain, B. 2010. Genetic and environmental influences on shape variation in the European sea bass (Dicentrarchus labrax). Biological Journal of the Linnean Society 101: 427–436.

De Jong, S. 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18: 251–263.

De Maesschalck, R., Estienne, F., Verdú-Andrés, J., Candolfi, A., Centner, V., Despagne, F., Jouan-Rimbaud, D., Walczak, B., Massart, D.L., de Jong, S., de Noord, O.E., Puel, C. & Vandeginste, B.M.G. 1999. The development of calibration models for spectroscopic data using principal component regression. Internet Journal of Chemistry 2. p. 19.

Eliasson, L. 1999. Simulation of thinning with a single-grip harvester. Forest Science 45: 26–34.

Frank, I. & Friedman J. 1993. A statistical view of chemometrics regression tools. Technometrics 35: 109–148.

Freedman, D., Pisani, R. & Purves, R. 2007. Statistics. 4th edition. W.W. Norton, Inc. New York. 697 p.

Geladi, P. & Kowalski, B. 1986. Partial least square regression: a tutorial. Analytica Chemica Acta 35: 1–17.

Harrop-Galvao, R., Ugulino-Araujo, M., Emıdio-Jose, M., Coelho-Pontes, M, Cirino-Silva, E. & Bezerra-Saldanha, T. 2005. A method for calibration and validation subset partitioning. Talanta 67: 736–740.

Helland, I. 1990. Pls regression and statistical models. Scandivian Journal of Statistics, 17: 97–114.

Hoskuldson, A. 1988. Pls regression methods. Journal of Chemometrics 2: 211–228.

Howard, A. 1989. A sequential approach to sampling design for time studies of cable yarding operations. Canadian Journal of Forest Research 19: 973–980.

Kresta, J.1992 The application of Partial Least Squares to problems in chemical engineering. Ph. D. thesis at McMaster University, Canada. 215 p.

Lipp, M. 1996. Comparison of PLS, PCR and MLR for the quantitative determination of foreign oils and fats in butter fats of several European countries by their triglyceride composition. Zeitschrift für Lebensmittel Untersuchung und Forschung A 202(3): 193–198.

Martens, H. & Naes, T. 1989. Multivariate calibration. London, Wiley.

McIntosh, A., Bookstein, F., Haxby, J. & Grady, C.L. 1996. Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3: 143–157.

Nakagawa, M., Hayashi, N. & Narushima, T. 2010. Effect of tree size on time of each work element and processing productivity using an excavator-based single-grip harvester or processor at a landing. Journal of Forest Research 15: 226–233.

Nsofor, G.C. 2006. Comparative analysis of predictive data-mining techniques. MSc thesis, University of Tennessee, Knoxville, USA.

Nurminen, T., Korpunen, H. & Uusitalo, J. 2006. Time consumption analysis of mechanized cut-to-length harvesting systems. Silva Fennica 40: 335–363.

Nuutinen, Y., Väätäinen, K., Heinonen, J., Asikainen, A. & Röser, D. 2008. The accuracy of manually recorded time study data for harvester operation shown via simulator screen. Silva Fennica 42: 63–72.

Olsen, E., Hossain, M. & Miller, M. 1998. Statistical comparison of methods used in harvesting work studies. Oregon State University, Forest Research Laboratory, Corvallis, OR. Research Contribution 23. 31 p.

Ovaskainen, H., Uusitalo, J. & Väätäinen, K. 2004. Characteristics and significance of a harvester operator's working technique in thinnings. International Journal of Forest Engineering 15: 67–77.

Pulos, S. & Rogness, N. 1995. Soft modeling and special education. Remedial and Special Education 16: 184–192.

Samset, I. 1990. Some observations on time and performance studies in forestry. Communications of the Norwegian Forest Research Institute 43(5). 80 p.

Spinelli, R. & Hartsough, B. 2001. A survey of Italian chipping operations. Biomass and Bioenergy 21: 433–444.

— & Magagnotti, N. 2009. A tool for productivity and cost forecasting of decentralised wood chipping. Forest Policy and Economics 12(3): 194–198. doi:10.1016/j.forpol.2009.10.002.

— & Magagnotti, N. 2011a. The effects of introducing modern technology on the financial, labour and energy performance of forest operations in the Italian Alps. Forest Policy and Economics 13: 520–524.

— & Magagnotti, N. 2011b. Wood extraction with farm tractor and sulky: estimating productivity, cost and energy consumption. Small Scale Forestry. DOI 10.1007/s11842-011-9169-8.
— , Nati, C. & Magagnotti, N. 2009. Using modified foragers to harvest short-rotation poplar plantations. Biomass and Bioenergy 33: 817–821.
— , Hartsough, B. & Magagnotti, N. 2010. Productivity standards for harvesters and processors in Italy. Forest Products Journal 60: 226–235.
— , Magagnotti, N., Paletto, G. & Preti, C. 2011 Determining the impact of some wood characteristics on the performance of a mobile chipper. Silva Fennica 45(1): 85–95.
Suadicani, K. & Fjeld, D. 2001. Single-tree and group selection in montane Norway spruce stands: factors influencing operational efficiency. Scandinavian Journal of Forest Research 16: 79–87.
Viscarra-Rossel, R., Taylor, H. & McBratney, A. 2007. Multivariate calibration of hyperspectral g-ray energy spectra for proximal soil sensing. European Journal of Soil Science 58(1): 343–353.
Visser, R. & Spinelli, R. 2011. Determining the shape of the productivity function for mechanised felling and felling-processing. Journal of Forest Research. DOI 10.1007/s10310-011-0313-2.
— & Stampfer, K. 1998. Cable extraction of harvester felled thinnings: an Austrian case study. Journal of Forest Engineering 9: 39–46.
Williams, P. 1987. Variables affecting near-infrared reflectance spectroscopic analysis. In: Williams, P. & Norris, K. (eds.). Near-infrared technology in the agricultural and food industries. American Association of Cereal Chemists, St Paul, Minnesota, USA. p. 143–166.
Wold, H. 1966. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P.R. (ed.). Multivariate analysis. New York: Academic Press. p. 391–420.

Total of 45 references