Procedure for Managing Large-scale Progeny Test Data: a Case Study of Scots Pine in Finland

Martti Venäläinen and Seppo Ruotsalainen

Venäläinen, M. & Ruotsalainen, S. 2002. Procedure for managing large-scale progeny test data: a case study of Scots pine in Finland. Silva Fennica 36(2): 475–487.

Large progeny test networks are typical for conventional forest tree breeding programmes. The individual progeny tests differ with respect to age, composition and ability to screen the breeding values of the parent trees. Several approaches have been introduced to manage the unbalanced and diverse nature of the data generated by progeny tests. This report presents a procedure for ranking breeding material on the basis of ”messy” data. Plot means were used as input values and missing plots were estimated from least squares. The differences between test means and variances were standardised by the performance level method. The different precision of the tests was quantified through the reliability coefficient. In order to facilitate the selection of plus trees for different purposes, all the available test results were combined into a single variable that was used for ranking. Three different kinds of ranking variable were calculated and each of them proved to be more useful for the selection of plus trees than an arithmetic or weighted mean. One of them, *WMEAN*, relied on the reliability and number of the progeny tests, while the others, $WCONF_{0.50}$ and $WCONF_{0.10}$, relied on the standard error of the plus tree mean, thus emphasising the precision of the values obtained. The analyses were carried out with SAS® procedures, which require only moderate skills in statistics, programming and data processing technology. The procedure has functioned well throughout an eight-year development phase. Nearly three thousand Scots pine (*Pinus sylvestris*) plus trees have been ranked for various characters, and the results have been used for roguing the seed orchards, to establish new ones, and to select plus trees for breeding populations.

Keywords *Pinus sylvestris*, progeny testing, performance level, reliability coefficient, selection

Authors' address Finnish Forest Research Institute, Punkaharju Research Station, FIN-58450 Punkaharju, Finland **Fax** + 358 15 644333 **E-mail** martti.venalainen@metla.fi

Received 11 December 2000 **Accepted** 4 February 2002

List of abbreviations

PL	=	performance level
$\overline{PL}$	=	plus tree-wise arithmetic mean of performance level values
w_j	=	a character-wise reliability coefficient for test j
PL'	=	performance level corrected with w_j
$\overline{PL'_w}$	=	weighted mean of corrected performance level values
$h^2{}_f$	=	family heritability
$WMEAN$	=	$\overline{PL'_w}$ corrected by the number of (reliable) tests
$WCONF_{0.5}$	=	the lower confidence limit of the $\overline{PL'_w}$ at risk level 0.5
$WCONF_{0.1}$	=	the lower confidence limit of the $\overline{PL'_w}$ at risk level 0.1
N	=	number of reliable tests

1 Introduction

Progeny testing is one of the key phases in each cycle of a multiple-generation tree breeding programme. The main goal of progeny testing is to rank parents based on the performance of their offspring, i.e. their breeding values. Estimates of breeding values are used to select trees for further breeding and to establish seed orchards. In most tree breeding programmes the progeny test data are generated from large-scale progeny test networks. Data from progeny tests are often unbalanced and the tests differ in age and precision. Data generated from the Finnish Scots pine (*Pinus sylvestris* L.) progeny testing programme shares these problems with other large breeding programmes.

Several approaches have been introduced to manage large amounts of progeny test data. White and Hodge (1989) have summarised the methods commonly used for calculating a family mean in a single test and, further, the methods for combining data from multiple tests: arithmetic average, least squares estimates, weighted least squares and averaged standard scores. The performance level method introduced by Hatcher et al. (1981) relies on standardized scores. Cotterill et al. (1983) compared six mathematical procedures for estimating the average performance of families: rank-score, site-adjustment, standard site-adjustment, least squares, weighted least squares, and shrunken least squares. They preferred the weighted least squares procedure "due to its providing minimum variance unbiased estimates of the effects of families under the additive model", although they concluded that the six methods ranked the material in approximately the same order. Kurinobu et al. (1985) compared last four procedures of Cotterill et al. (1983) and concluded that "the standard site adjustment is recommended in most practical situations because it is easier to compute than the other three, and it probably ranks families with sufficient accuracy". White and Hodge (1989) suggested applying the best linear (unbiased) predictions, (BLP and BLUP), for the breeding values because these methods "account for test data of different quality and quantity". Jansson (1998) agreed that " the method recommended today is Henderson's mixed-model equation (MME). Solutions to MME yield BLUP breeding values".

The aim of this report is to introduce a procedure which is being used to rank breeding values of plus trees from "messy" progeny test data and to evaluate its efficacy.

2 Material

The phenotypic selection of Scots pine plus trees began in Finland in 1947. In the 1960's there was a selection surge motivated by the decision to establish 3400 ha of first generation seed orchards (Oskarsson 1995). By 1992 a total of nearly 6800 plus trees had been registered, 66% of them in North Finland (Rusanen 1992). The first full-scale progeny test was planted in 1960, but a large series of tests was not initiated until 1968. The test establishment was completed in South Finland in 1990 while in North Finland, where the progeny testing began later, not until 2001. However, there is already a considerable number of progeny test measurements available for ranking the plus trees even in the north. In 1999 the total number of Scots pine progeny tests was 1394. They covered an area of 2111 ha (Yrjänä et al. 2000) and included the progenies of 4700 plus trees.

The first obstacle to accurately rank parent trees based on data from several tests is the unbalanced nature of the data. The plan was to include each progeny in at least two medium-term test orchard tests and two long-term field tests (Mikola 1984). However, the progenies were not included in the same number of tests. The most common field test design used in Finland was complete randomised blocks with 25-seedling square plots with a 2×2 m spacing. Smaller plots would have been statistically more efficient (Haapanen 1992) and, because the area of homogeneous test sites has often been limited, large plot size led to a low number of entries per test, thus increasing number of tests. Later on, some of the tests were so seriously damaged by a number of agents that they were abandoned.

Open pollinated seed harvested from young, no male strobili producing, seed orchards and clonal archives was used to raise the seedlings for progeny testing. Thus, the plus tree composition of the tests reflects the clonal composition of the seed orchards. The Scots pine breeding zones were delineated after the seed orchards were established and it was discovered that most seed orchards included plus trees from more than one breeding zone. The composition of the progeny tests established before the middle of 1980's was therefore seldom optimal because the plus trees are ranked within breeding zones. Plus trees originating from distant breeding zones were not included in the same analysis even if they were present in the same test.

The second main obstacle to accurate ranking was the heterogeneity of the data that arises from the diverse testing environments and the age distribution of the tests. The tests were established on a range of site types and soils and subsequent management practices varied. Furthermore, nursery practices changed during the test establishment process. At the outset tests were established with 2-year-old bare-rooted seedlings. Later several types of 1-year-old containerised seedlings were used. In addition, soil preparation practices changed considerably during the course of the programme. All these factors led to varying means and variances in the characters even if the test ages were the same, and tests of different ages had different means and variances. Data from older tests were considered more reliable, at least with respect to volume and quality properties.

A third main problem was encountered when testing the northern material. Seed orchards with northern plus trees were established in South Finland in order to guarantee seed maturation (Sarvas 1970). As a result of background pollination, the seed harvested for testing has been provenance hybrid seed. Such hybrid progenies were considered to be suitable for ranking parents but were not hardy enough to survive well in the original breeding zone of the plus trees (Mikola 1993). Differences in survival can be tested there, but the growth chracters must be tested in less severe conditions. Thus, data for different characters may accumulate from different tests.

3 Methods

3.1 Calculation Methods

The test-wise mean for the progeny of plus tree *i* (referred to hereafter as plus tree *i*) was calculated as a least squares mean from the plot means (e.g. the survival percentage of the plot or the plot mean height) to adjust for missing plots. The progeny means were standardised for each test using the performance level method (Hatcher et al. 1981). The performance level (*PL*) for plus tree *i* in the progeny test *j* is

$$PL_{ij} = 50 + 25\frac{x_{ij} - \bar{x}_j}{s_j} \qquad (1)$$

where x_{ij} is the mean of plus tree *i* on the original scale in the progeny test *j*, $\bar{x}_j$ is the mean of all plus trees in the progeny test *j*, and s_j is the standard deviation of the plus tree means in the progeny test *j*. The standardised performance level values have a mean of 50 and standard deviation of 25 in each progeny test. Thus, the character-wise performance level values make it possible to meaningfully compare and average the plus tree values (e.g. calculate the arithmetic mean $\overline{PL}$) across progeny tests with different means and variances.

The progeny tests presumably differed in their ability to reveal the breeding value of plus trees, i.e. the information obtained from one test was more valuable for the correct ranking of the material than the information derived from another. These differences were taken into account by using correction factors to adjust the performance level values. The correction factors form a character-wise reliability coefficient w_j, with values between 0 and 1.

The reliability coefficient w_j was calculated from three factors two of which were empirical and one that was derived from the measured data. The age of the test was included as one of the empirical factors. Data measured at ages less than 7, from 7–12, and older than 12 were weighted 0, 0.9, and 1, respectively. The second empirical factor was the location of the test. Data from tests situated in non-optimal climatic conditions were weighted less than 1 or even equal to zero. Different weights were assigned to characters when the climate was thought to have different effects.

Family heritability played the most important role in constructing the reliability coefficient. It was estimated from the test data using the F- test value for the ANOVA of randomized blocks.

$$h_f^2 = \frac{\left(\frac{MS_{plustree} - MS_{err}}{r}\right)}{\left(\frac{MS_{plustree} - MS_{err}}{r}\right) + \frac{MS_{err}}{r}} = 1 - \frac{1}{F} \qquad (2)$$

where r is the number of blocks, and F is the test value for the F statistics.

In theory, the family heritability lies between 0 and 1. In practice, however, values near 1 were rare and thus the family heritability values were arbitrarily rescaled for the reliability coefficient by dividing them by 0.8. If the rescaled sub-coefficient exceeded 1 it was truncated to 1. Negative values were set to zero. The final reliability coefficient w_j for test j was obtained by multiplying the three sub-coefficients together.

Corrected performance level values (PL'_{ij}) were calculated from

$$PL'_{ij} = (PL_{ij} - 50)w_j + 50 \qquad (3)$$

This correction moved the PL'_{ij} values of a less-informative progeny test closer to the mean 50 than the PL'_{ij} values of a progeny test with a high reliability coefficient w_j. Thus, selection aimed at the tails of the pooled distribution was based more on the informative progeny test.

The reliability coefficients w_j were also used to calculate the plus tree-wise weighted means ($\overline{PL'_{Wi}}$) over the n progeny tests in which plus tree i was included:

$$\overline{PL'}_{Wi} = \frac{\sum_{j=1}^{n} w_j PL'_{ij}}{\sum_{j=1}^{n} w_j} \qquad (4)$$

Different plus trees were tested in a widely differing number of progeny tests. A plus tree included in several tests will have a more precisely estimated progeny mean and should thus be favoured in selection over a plus tree that has the same progeny mean, but which is included in fewer progeny tests. However, when the mean is calculated from a higher number of tests, the error, plot and progeny × block variances of the estimate decrease, and the means of well-tested plus trees therefore tend to concentrate closer to the population mean. This would lead to an undesired selection result because the imprecisely tested trees would be more likely selected than the more precisely tested trees (White and Hodge 1989, Danell 1991).

Two options were used to counteract the effect of different testing frequency. In the first option, the progeny means were further corrected by the number of tests applying the same equation as Nikkanen and Pukkala (1987). This correction spreads the progeny means further from the population mean as a function of the number of tests. The final mean, $WMEAN_i$, of the corrected performance level values for plus tree i, weighted by the test-wise reliability coefficient and corrected by the number of tests, was obtained from the formula

$$\begin{aligned} & WMEAN_i \\ & = \left(\frac{\sum_{j=1}^{n} w_j((PL_{ij} - 50)w_j + 50)}{\sum_{j=1}^{n} w_j} - 50 \right) N^q + 50 \qquad (5) \\ & = (\overline{PL'}_{Wi} - 50)N^q + 50 \end{aligned}$$

where N is the number of (reliable) progeny tests and q is an empirical parameter which determines how strongly the number of tests spreads the plus tree means. The other option was to rank the plus trees according to the lower confidence limit of the weighted performance level mean, referred to later as $WCONF_{\alpha i}$.

$$WCONF_{\alpha i} = \overline{PL'}_{Wi} - t_{\alpha\left(\sum w_{ij}\right)} \left(\frac{s_{Wi}}{\sqrt{\sum_{j=1}^{n_i} w_{ij}}} \right) \qquad (6)$$

where $t_{\alpha(\sum w_{ij})}$ is the critical value of the Students t-distribution at risk level α with $\sum w_{ij}$ degrees of freedom, and s_{Wi} the weighted standard deviation associated with the tests in which plus tree i is included.

The calculations and database management were performed using SAS® statistical software (Fig. 1). The final product of the procedure was a database with three ranking variables, *WMEAN*, $WCONF_{0.50}$ and $WCONF_{0.10}$ for each character (Appendix 1).

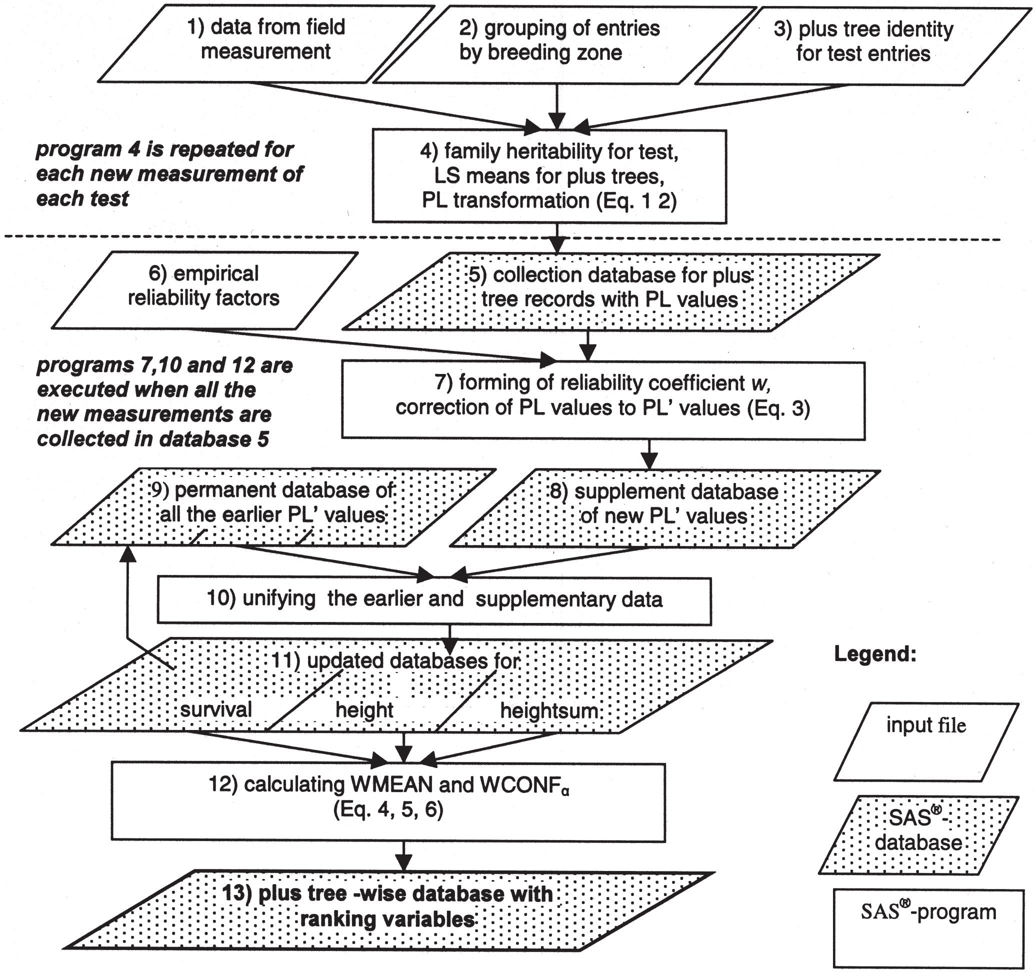
1) data from field measurement
2) grouping of entries by breeding zone
3) plus tree identity for test entries
program 4 is repeated for each new measurement of each test
4) family heritability for test, LS means for plus trees, PL transformation (Eq. 1 2)
6) empirical reliability factors
5) collection database for plus tree records with PL values
programs 7,10 and 12 are executed when all the new measurements are collected in database 5
7) forming of reliability coefficient w, correction of PL values to PL' values (Eq. 3)
9) permanent database of all the earlier PL' values
8) supplement database of new PL' values
10) unifying the earlier and supplementary data
11) updated databases for
survival
height
heightsum
12) calculating WMEAN and WCONFα (Eq. 4, 5, 6)
13) plus tree -wise database with ranking variables
Legend:
input file
SAS®-database
SAS®-program

Fig. 1. Flow-chart of the procedure for managing very large amounts of progeny test data.

Table 1. Number of Scots pine plus trees for which progeny test results were processed.

	Survival		Height sum		Height		Quality	
	$\overline{PL}$	*WMEAN*	$\overline{PL}$	*WMEAN*	$\overline{PL}$	*WMEAN*	$\overline{PL}$	*WMEAN*
South Finland	–[1]	–	1699	1505	1696	1600	996	847
North Finland	1279	1151	1199	1092	1177	1041	–[2]	–

[1] not a breeding goal in South Finland
[2] not measured in North Finland

3.2 Methods Used to Evaluate the Procedure

After processing the available data, retrospective studies were carried out in order to evaluate the procedure. Two methods were used to evaluate how well the reliability coefficients classified the tests according to the precision of the information. The first evaluation method relied on plus trees that were represented in 30 or more tests. These tests were ranked for each plus tree according to the reliability coefficient in descending order. The standard deviation of the *PL* values was calculated over the first nine tests. Then the calculation was repeated for the 2nd–10th tests and so on until the last test in the ranking list was reached. This gave a series of running standard deviations for the plus tree-wise *PL* values. Averaging across the plus trees sequence by sequence provided data that connected the standard deviation and the reliability. The hypothesis was that the standard deviation should increase along with a decrease in the reliability.

The other means of evaluating the reliability coefficients was to study the correlation between plus tree-wise *PL* values in those tests that had varying reliability coefficients. In order to do this the data were divided into six classes according to the reliability coefficients. The first class included observations obtained from tests with a reliability coefficient of from 1.0 to 0.81, the following classes from 0.80 to 0.61, 0.6 to 0.41, 0.4 to 0.21, 0.20 to 0.01 and the sixth class of 0. Because in a number of cases there were several observations from one plus tree in the reliability class, the plus tree-wise means were calculated by classes. The Pearson's correlation coefficient for the plus tree-wise *PL* value means was calculated between the reliability classes. The hypothesis was that the correlation of the *PL* values between the groups with high coefficients was high because they better estimated the true breeding values.

In order to study the effect of testing frequency on the final survival ranking of the plus trees, they were grouped into ten test frequency classes according to the number of tests in which each plus tree was included. Two selection simulations were carried out: an intensive selection of 6% of the material simulated the selection for a multiplication population (e.g. seed orchard), and an extensive selection of 44% simulated the selection for a breeding population. The set of all the tested plus trees from zones 5–11 was used as the base population (Table 1). Five selection criteria were used: $\overline{PL}$, $\overline{PL}'_w$, *WMEAN*, $WCONF_{0.10}$ and $WCONF_{0.50}$. The proportion of selected plus trees in each test frequency class was then calculated. The five different survival rankings obtained were also compared in order to determine the extent to which the same plus trees were selected by the different selection criteria.

4 Results

4.1 Result of Combining the Progeny Test Data

The procedure described above appeared to function well in combining the measurement data of 620 progeny tests. The analysis produced results from about 3000 Scots pine plus trees, grouped according to breeding zones. The survival and height data of the plus trees originating from breeding zones 1–4 were analysed according to the algorithm tailor-made for South Finland, and the rest (zones 5–11) were analysed using the algorithm for North Finland (Table 1). Different algorithms were applied because the breeding goals varied in different parts of the country. In South Finland the age and the phase of development of the progeny tests made it possible to

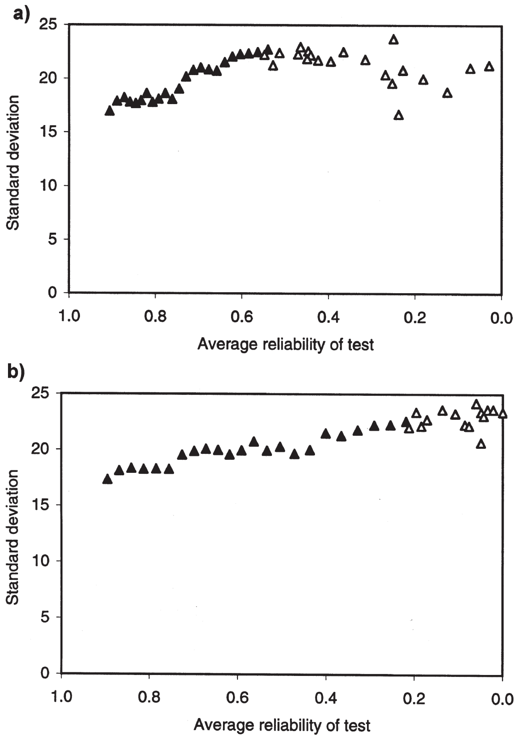

Fig. 2. Average running standard deviation of *PL* values between progeny tests arranged in decreasing order according to their reliability coefficient *w*. In survival (a) the average of the standard deviation is based on a maximum of 24 and in height (b) on a maximum of 19 plus trees. The observations in which the maximum number of plus trees was included are marked with the symbol ▲. When the number of plus trees included in the observation was less than the maximum, Δ is used as the symbol.

measure external quality chracters, for which a specific algorithm was used.

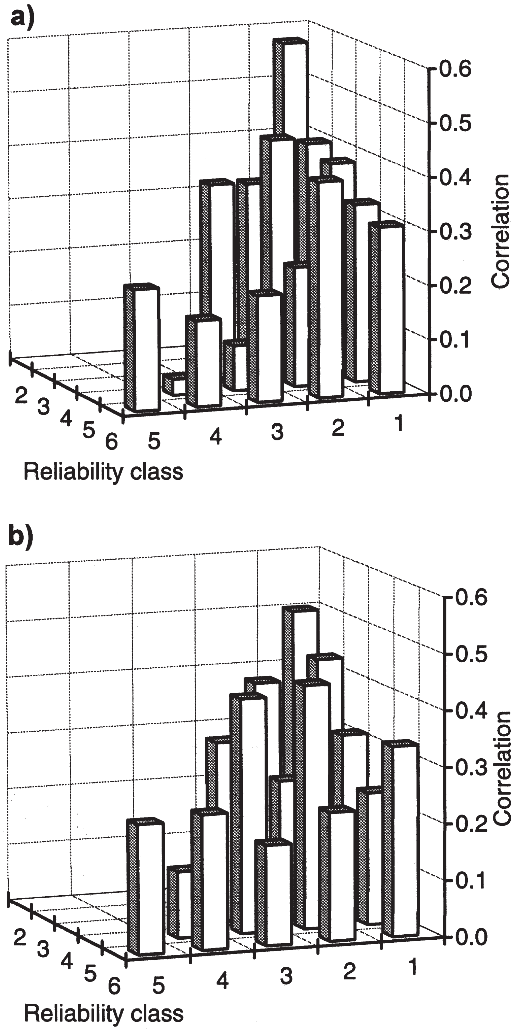

Fig. 3. Correlation between plus tree-wise *PL* means based on different reliability classes (class 1: 0.8 < w 1, class 2: 0.6 < w 0.8 … class 6: w = 0) for survival (a) and for height (b).

4.2 Functioning of the Reliability Coefficient

In South Finland about 13% and in North Finland 2% of the progeny test measurements did not contribute at all to the ranking of plus trees because the value of the reliability coefficient was zero. The variation between the results of different progeny tests increased with decreasing reliability coefficient of the test (Fig. 2). Correlations between plus tree means calculated from the progeny tests of different reliability classes were positive, the highest correlation being obtained between the two highest reliability classes (Fig. 3). The corresponding correlation for survival was 0.57 (p = 0.000) and for height 0.50 (p = 0.000). The correlation coefficients for both characters decreased irregularly among the lower reliability classes.

4.3 Number of Tests and Variance of the Plus Tree Means

When an intensive selection was based on $\overline{PL}$, those plus trees that had been tested in only one or two progeny tests were over-represented (Fig. 4a). Among the selection criteria, *WMEAN* was the least sensitive to the testing frequency (sd = 1.2) although it did show a slight tendency to favour the most frequently tested trees (r = 0.28) (Table 2). The most distinctive feature for the confidence limit-based selection criteria was that they excluded plus trees tested in only one test. In a less intensive selection all the criteria behaved in a similar way and gave a rather equal representation for different test frequency classes. The exception was $WCONF_{0.10}$, which favoured the most frequently tested plus trees (Fig. 4b).

The variables $\overline{PL}$ and *WMEAN* gave a similar ranking for the plus trees (Fig 5a). However, a closer study revealed the correction property of *WMEAN*: the extreme plus trees moved towards the population mean 50 if their value was based only on a few reliable tests (symbol "–"). When the number of reliable tests increased (symbol "x") the other property of *WMEAN* was revealed: values which are based on several tests move away from the mean. The comparison between $\overline{PL}$ and $WCONF_{0.10}$ shows that a plus tree tested in none or only a few reliable progeny tests can have a much lower ranking with the $WCONF_{0.10}$ variable than would be expected on the basis of its $\overline{PL}$ value (Fig. 5b).

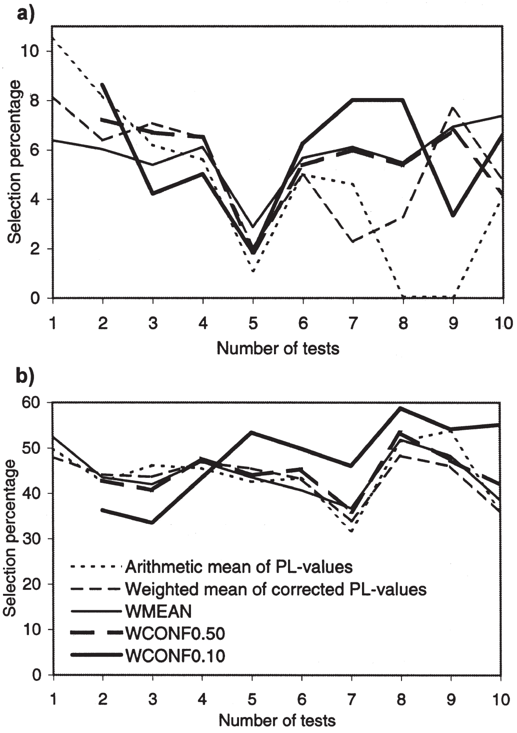

Fig. 4. Comparison of different selection criteria for survival as a function of the number of progeny tests when selecting for a multiplication population (selected proportion 6%) a) and when selecting for a breeding population (selected proportion 44%) b).

Table 2. Standard deviation of the selection percentages in different test frequency classes and the correlation coefficient (with statistical significance in parentheses) between test frequency class and selection percentage using different selection criteria (n=10). Result of two selection simulations: for a multiplication (6% selected) and for a breeding population (44% selected).

		Selection criteria				
		$\overline{PL}$	$\overline{PL'_w}$	*WMEAN*	$WCONF_{0.5}$[1)]	$WCONF_{0.1}$[1)]
6%	Sd	3.4	2.2	1.2	1.7	2.3
selected	r	–0.775	–0.418	0.280	–0.270	0.028
	(p)	*(0.009)*	*(0.229)*	*(0.433)*	*(0.482)*	*(0.944)*
44%	Sd	6.8	4.9	5.3	4.9	8.7
selected	r	–0.167	–0.427	–0.259	0.184	0.841
	(p)	*(0.644)*	*(0.219)*	*(0.470)*	*(0.636)*	*(0.004)*

[1)] excludes plus trees tested in only one test

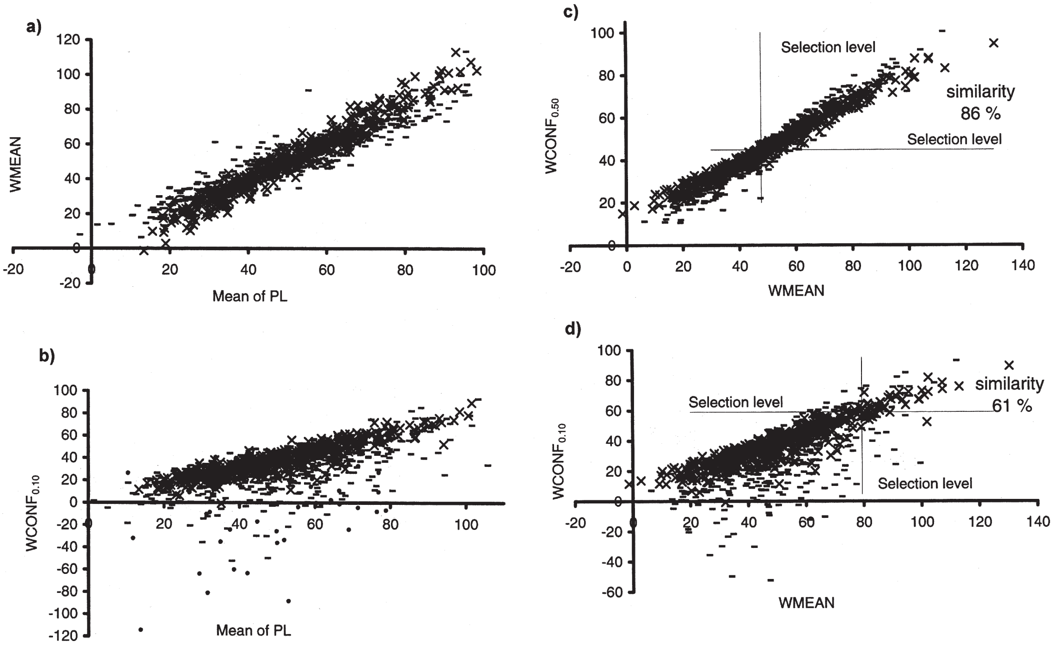

Fig. 5. Comparison of the different selection criteria for survival: *WMEAN* (a) and *WCONF*$_{0.50}$ (b) versus the arithmetic mean of *PL* values, *WCONF*$_{0.50}$ versus *WMEAN* (the selection level corresponds to selection for a breeding population) (c) and *WCONF*$_{0.10}$ versus *WMEAN* (the selection level corresponds to selection for a multiplication population) (d). Number of reliable progeny tests on which the plus tree value is based: × = 3 or more, ▬ = 1 or 2, • = no reliable tests. The total number of tests included in the observations is higher because only 50% of the tests were considered to be "reliable".

Table 3. Pair-wise comparisons of different selection criteria in two selection simulations for survival. The values presented are proportions (%) of plus trees selected by both methods. Selection for a multiplication population (selection ratio 6%) is above the diagonal and for a breeding population (selection ratio 44%) below the diagonal

	Selection criteria				
	$\overline{PL}$	$\overline{PL'_w}$	WMEAN %	*WCONF*$_{0.50}$	*WCONF*$_{0.10}$
$\overline{PL}$		81	72	69	55
$\overline{PL'_w}$	90		84	79	52
WMEAN	85	91		79	61
WCONF$_{0.50}$	80	86	86		73
WCONF$_{0.10}$	68	71	75	84	

Different selection criteria resulted in selecting different groups of plus trees. The correspondence between the groups is shown in Table 3. The difference between the criteria was greater when selecting the small top fraction for a multiplication population than when selecting a larger breeding population with a lower selection intensity. *WMEAN* and *WCONF*$_{0.50}$ were in good agreement (Fig. 5c), giving 79% and 86% of the same plus trees for multiplication and breeding populations, respectively, whereas *WCONF*$_{0.10}$ differed more from the others. This was due to its great emphasis on the standard error of the mean.

A comparison of $WCONF_{0.10}$ and *WMEAN* results revealed that the plus trees with the greatest change in ranking were those tested in a low number of reliable progeny tests (Fig. 5d). More than half of the plus trees that were selected for the multiplication population with *WMEAN*, but not with $WCONF_{0.10}$, were tested in only one or two reliable progeny tests. The rest of the trees in this group usually had a large standard error of the mean due to the fact that the progeny tests yielded contradictory results for their performance. Both of these factors contributed to a wide confidence limit of the mean, and thus a low value for the selection criteria $WCONF_{0.10}$. The opposite group, trees selected according to $WCONF_{0.10}$, but not according to *WMEAN*, also consisted mainly of trees with a low number of reliable progeny tests. However, these trees typically had an exceptionally low standard error of the mean.

5 Discussion

The procedure has functioned well throughout the eight-year development phase. Nearly three thousand plus trees have been ranked for various characters, and the results have been used in roguing the seed orchards, in establishing new ones, and in selecting plus trees for breeding populations. The procedure has practical importance for managing of the huge amount of progeny test data. However, it is difficult to say how well the three ranking variables correlate with the real breeding values of the plus trees, because we neither knew the values nor had any other independent estimates of them. Thus we could only evaluate the procedure by studying how it handled the problems that commonly occur in the processing of progeny test results.

Problems addressed by the procedure were unequal number of trees per plot, missing plots and different means and variances of numerous tests. The unequal number of trees per plot was counteracted by using plot means as input values. Missing plots were estimated from least squares. The differences between test means and variances were standardised by using the performance level method (Hatcher et al. 1981). The performance level method has also been applied recently by Bridgwater and McKeand (1997). The different precision of the tests was quantified through the reliability coefficient. The selection of plus trees for different purposes was facilitated by combining all the available test results in slightly different ways into three ranking variables.

The role of the character-wise reliability coefficient was fundamental in the procedure because it determined the effect of a specific progeny test measurement in four different ways. It was initially used to correct the original *PL* values to *PL'* values (Eq. 3). This correction seems to be similar to the calculation of a 'shrunken estimate' proposed by Burdon (1998) in which the entry-mean departures from the overall site mean were multiplied by the heritability of the means. Its second use was as a weighting factor in calculating plus tree means (Eq. 4). Since family heritability was the main component of the reliability coefficient, weighting the test results by the reliability coefficient had essentially the same effect as weighting by the family heritability (Burdon 1998) or weighting by the inverse of the error variance (Cotterill et al. 1983, Kurinobu et al. 1985). Equations 3 and 4 acted in same direction: less resolution power was given tests with low reliability. However, there was one important difference between these two equations. Weighting alone cannot influence the plus tree mean if there is only one test or even several tests with equally low reliability. This is why the correction performed by Equation 3 was also necessary. The reliability coefficient was taken into account for the third time when half of the tests were labelled as relatively reliable. When *WMEAN* (Eq. 5) was calculated, the number of labelled tests was used as the parameter *N*. If there was no reliable test for a plus tree, *WMEAN* could not be calculated even though the plus tree was included in several tests. And finally, if a plus tree had been tested in a large number of tests, it was reasonable to limit the number of tests included in the calculation of the plus tree-wise averages. The best ten tests, according to the reliability coefficient, were included.

The results indicate that the reliability coefficient described the progeny test precision. The variation between the progeny test results was the greater, the lower were the reliability coefficients of the progeny tests (Fig. 2). The decreasing correlation coefficients between the results for plus trees in progeny tests with lower reliability coefficients (Fig. 3) reflected the same phenomenon.

The development of various kinds of ranking variable proved to be necessary. The intensive selection of plus trees on the basis of the arithmetic mean of the *PL* values favoured plus trees which were included in only a few progeny tests (Fig 4a, Table 2) and were therefore tested the least precisely. This observation fits well with the theoretical expectation (White and Hodge 1989). The ranking variable *WMEAN*, one parameter of which was the number of reliable tests, was less sensitive to the testing intensity. It appeared (Fig 5a) that *WMEAN* treated the plus trees tested in only a few tests in the same way as the method of 'shrunken least squares' used by Cotterill et al. (1983) and Kurinobu et al. (1985). The variables based on the lower confidence limit, $WCONF_{0.50}$ and $WCONF_{0.10}$, also considerably improved the equality of selection in class two and upwards. Selection on the basis of a single trial was totally omitted when these variables were used.

The lack of true breeding values for the plus trees hampered the real evaluation of the different ranking variables. However, they were studied by comparing the selection results obtained by each of the ranking variables. The ranking variables mainly differed in the way they treated plus trees tested in a small number of tests (Fig. 5c,d). Trees that were selected according to *WMEAN* but not according to $WCONF_{0.10}$ were trees with a large standard error. It is more usual for a plus tree to have a large standard error on the basis of a few than on several progeny tests. However, even the opposite group, i.e. trees selected by $WCONF_{0.10}$ but rejected by *WMEAN*, was dominated by trees included in only one or two reliable progeny tests. This may be a weakness of this ranking variable: a low standard error can be obtained accidentally if there are only a few progeny tests. This demonstrates that $WCONF_{0.10}$ is essentially a ranking variable that emphasises the standard error and only indirectly takes the number of tests into account.

The purpose of a selection task determines which ranking variable is the most appropriate in each situation. If the aim is to select a limited number of plus trees for a multiplication population, the infallibility of the selection can be so important that selection should be based on a criterion like $WCONF_{0.10}$ to weight the precision of the test results. However, the infallibility is paid for by a decrease in genetic gain, since promising plus trees are omitted because of a large or non-estimable standard error. When selecting for a long-term breeding population, the average genetic gain being of major importance and failed selections being revealed in further testing, selection could be based on a criterion like *WMEAN*.

More studies are needed to determine the sound threshold for a "reliable" test. At present a test is regarded as reliable for the character in question if its reliability coefficient exceeds the median of all the tests. However, it might be better to determine the limit according to an absolute value. Such a determination would change the final result of the selection to some extent, because the number of reliable tests is an important factor in calculating *WMEAN*.

The procedure will not solve the dilemma of simultaneous selection for growth and quality, which is a current task in South Finland. Neither will the procedure yield any meaningful estimates of genetic gain achieved by selection. This is due to the difficulty of reverse-transformation that is common to all methods based on transformed values (White and Hodge 1989). Bridgwater and McKeand (1997) proposed transforming the gain calculated in *PL* units back to standard deviation units. However, their proposal does not solve the problem satisfactorily in large-scale and "messy" data like ours.

Although the procedure was developed to manage the results from a large number of progeny tests, it had the advantage that each progeny test was analysed separately, and the results were collated into intermediate databases for further processing. This means that even extremely unbalanced and large databases like the Scots pine progeny tests in Finland can be managed and analysed. The analyses were carried out with SAS® procedures that required only moderate skills in statistics, programming or data technology. This procedure has yielded valuable information about the plus trees for the needs of tree breeders. Thus, although it is not theoretically the optimal solution, it will be used in practice until a more advanced procedure can be introduced.

Acknowledgements

A manual database for the Scots pine progeny test results in Finland was initiated by Jouni Mikola in 1976. Construction of the present, SAS® statistical program-based, semi-automatic procedure was started in 1991. Pekka Vakkari guided the use of SAS® in the early stages. The first version of the procedure was documented in a publication in Finnish in 1994. The actual databases have been continuously managed by Marja-Leena Annala and Erkki Kosonen. The authors are grateful to all colleagues for their help during this long-lasting undertaking, and especially to Matti Haapanen for his support and criticism. John Derome has patiently revised the language of the manuscript. The contribution of the referees, prof. Csaba Matyas and the anonymous one, improved the manuscript considerably.

References

Bridgwater, F.E. & McKeand, S.E. 1997. Early family evaluation for growth of Loblolly pine. Forest Genetics 4(1): 51–58.

Burdon, R.D. 1998. Relative performance values in genetic tests: Alternatives and their properties. Silvae Genetica 47(1): 1–5.

Cotterill, P.P., Correll, L.L. & Boardman, R. 1983. Methods of estimating the average performance of families across incomplete open-pollinated progeny tests. Silvae Genetica 32(1–2): 28–32.

Danell, Ö. 1991. Prediction of genetic values – basic concepts and methods. Institutet för skogsförbättring. Arbetsrapport 250. The Institute for Forest Improvement. Work report. Uppsala. 30 p.

Haapanen, M. 1992. Effect of plot size and shape on the efficiency of progeny tests. Tiivistelmä: Koeruudun koon ja muodon vaikutus jälkeläiskokeen tehokkuuteen. Silva Fennica 26(4): 201–209.

Hatcher, A.V., Bridgwater, F.E. & Weir, R.J. 1981. Performance level – standardized score for progeny test performance. Silvae Genetica 30(6): 184–187.

Jansson, G. 1998. Approaches to genetic testing and evaluation in forest tree breeding. Acta Universitatis Agriculturae Sueciae. Silvestria 56. Uppsala. 24 p.

Kurinobu, S., Shingai, Y. & Ohba, K. 1985. Comparing the plus-tree evaluation methods with unbalanced data of progeny-testing plantations. Journal of Japanese Forestry Society 67(8): 322–326.

Mikola, J. 1984. Methods used for the genetic evaluation of tree breeding material in Finland. In: Tigerstedt, P.M.A., Puttonen, P. & Koski, V. (eds.), Crop physiology of forest trees. University press. Helsinki. p. 225–231.

— 1993. Provenance and individual variation in climatic hardiness of Scots pine in northern Finland. In: Alden, J., Mastrantonio, J.L. & Ødum, S. (eds.), Forest development in cold climates. Proceedings of a NATO Advanced Research Workshop. Laugarvatn, Iceland, June 18–23.1991. Plenum Press. New York and London. p. 333–342.

Nikkanen, T. & Pukkala, T. 1987. Siemenviljelysten harvennussuunnitelman laatiminen ATK- ohjelmistolla. Summary: Making a thinning plan for seed orchards using a computer program. Folia Forestalia 701. 26 p.

Oskarsson, O. 1995. Silmällä tehty savotta. Pluspuiden valinnan historia ja arki. Metsäntutkimuslaitoksen tiedonantoja 579. 68 p. (In Finnish)

Rusanen, M. 1992. Suomen metsänjalostuksen yleistilastoa 1.1.1992.. General statistics on forest tree breeding in Finland 1.1.1992. Metsäntutkimuslaitoksen tiedonantoja 421 (The Finnish Forest Research Institute, Research Papers). 18 p.

Sarvas, R. 1970. Establishment and registration of seed orchards. Folia Forestalia 89. 24 p.

White, T.L. & Hodge, G.R. 1989. Predicting breeding values with applications in forest tree improvement. Kluwer Academic Publishers. Dordrecht. 367 p.

Yrjänä, L., Karvinen, K. & Napola, J. 2000. Suomen metsänjalostuksen yleistilastoa 2000. General statistics on forest tree breeding in Finland 2000. Metsäntutkimuslaitoksen tiedonantoja 783 (The Finnish Forest Research Institute, Research Papers). 46 p.

Total of 16 references

Appendix. Three different kinds of mean for the performance level (PL) values of height (arithmetic mean, weighted mean of corrected values (PL´) and the same corrected further with the number of reliable tests (WMEAN)) and two other ranking variables, which are based on the lower confidence limit of the weighted mean of PL´ values. The example data, including some plus trees, are taken from the database 13 presented in Fig. 1.

Zone	Plus tree	*PL*		*w*	*PL´*		*WMEAN*		*WCONF*$_{0.50}$	*WCONF*$_{0.10}$
		Mean	n_{PL}	in average	Weighted mean	$n_{PL´}$		$n_{reliable}$		
5	P358	43.5	6	0.53	43.1	5	41.2	3	37.9	26.5
5	P652	54.5	4	0.59	62.9	4	65.1	2	55.6	38.4
5	P1679	42.0	4	0.59	42.7	4	40.6	3	39.1	30.8
5	P1680	44.7	4	0.59	48.6	4	48.2	3	46.9	42.9
5	P1697	70.1	2	0.57	58.2	2	58.2	1	56.0	46.0
5	P1701	64.1	4	0.55	60.8	4	62.6	2	54.6	39.6
5	P1702	49.0	4	0.55	44.0	4	43.0	2	39.0	27.0
5	P1703	25.4	3	0.62	27.6	3	23.8	2	18.9	–4.6
5	P1749	63.9	5	0.59	55.3	5	56.7	3	46.9	29.4
6	P765	68.9	2	0.57	61.0	2	61.0	1	59.2	51.3
6	P1592	65.3	5	0.49	53.2	5	53.7	2	48.1	36.6
6	P1667	39.8	4	0.61	35.0	3	32.5	2	29.7	14.9
6	P1669	60.8	4	0.61	48.6	3	48.3	2	43.8	30.8
6	P1673	61.9	3	0.62	63.1	3	65.3	2	57.7	43.2
6	P1675	20.0	2	0.57	24.4	2	24.4	1	11.6	–44.9
6	P1677	77.7	3	0.62	69.1	3	72.3	2	65.5	55.6
6	P1930	45.3	2	0.57	58.4	2	58.4	1	45.3	–12.4
6	P1933	37.7	2	0.57	49.9	2	49.9	1	42.5	9.5
6	P1936	35.7	2	0.57	39.7	2	39.7	1	35.8	18.9