13th Australasian Plant Breeding Conference
Breeding for Success: Diversity in Action
Christchurch, New Zealand, 18-21 April 2006

Benefits from Data and Pedigree Integration in Genetic Evaluation

G.W. Dutkowski[1], T.A. McRae[12], M.B. Powell[2], D.J. Pilbeam[2], K. Joyce[3], B. Tier[4], R.J. Kerr[1]

[1]PLANTPLAN Genetics Pty Ltd, Hobart, TAS, and Mount Gambier, SA, Australia
[2]Southern Tree Breeding Association Inc. Mount Gambier, SA, and Bunbury, WA, Australia
[3]Gunns Plantations Limited, Ridgley TAS, Australia
[4]Animal Genetics and Breeding Unit, University of New England, Armidale, NSW, Australia

ABSTRACT

The efficiency of plant breeding programs can be improved by utilising all performance data and pedigree relationships in genetic evaluation. The individual additive genetic model can be applied to industry wide programs with data spread across generations, locations, years and age classes in multi-trait analyses to produce more accurate breeding values for a given species. TREEPLAN® is being used routinely for genetic evaluation in tree species. Examples from operational breeding programs indicate that selection of improved genotypes in the past has been inefficient. TREEPLAN® breeding values predict realised gain trial performance better than historic models used. The TREEPLAN® approach to data standardisation provides a balance between computational efficiency and model complexity. Spatial analysis can also be used to increase within trial selection efficiency. Integrated analysis accounts for differences in genetic composition between trials and allows selections for different site types by taking into account selection history and correcting for poor connection between populations in previous generations.

INTRODUCTION

The Southern Tree Breeding Association (STBA) established national cooperative breeding programs for *Pinus radiata* (1983) and *Eucalyptus globulus* (1994) from the amalgamation of existing breeding programs of member companies. The amount and complexity of the amalgamated data from hundred of progeny trials precluded the use of much of it, and large across site analyses were only carried out infrequently (Jarvis *et al.* 1995) or with simple family models (White *et al.* 1992). The TREEPLAN® genetic evaluation system (Kerr *et al.* 2002; Kerr *et al.* 2001; McRae *et al.* 2004), and a supporting data management system (STBA-DMS^{TM}), were developed so that joint analysis of all trial data was no longer a limiting step in breeding. TREEPLAN® is now being used to routinely update genetic values in *E. globulus*, *P. radiata* and *E. nitens* on a program wide basis, and is easily adapted for other species. This paper briefly outlines the models used by TREEPLAN® - further detail can be found in the works cited above. This paper concentrates on examples of the benefit of the application of such models to tree breeding programs. We think the approach and examples will be of general interest to plant breeders. The software is available as the PLANTPLAN® software, customised to the needs of individual breeders.

MATERIALS AND METHODS

TREEPLAN® fits the model:

$$\mathbf{y} = \mathbf{Wf} + \mathbf{Xr} + \mathbf{Yu} + \mathbf{Zs} + \mathbf{e}$$

where: **y** is the vector of observations on one or more traits; **f** is the vector of fixed site and design effects and any other site specific covariates, with its incidence matrix **W**; **r** is the vector of random design effects, with its incidence matrix **X**; **u** is the vector of random additive genetic effects (breeding values) with its incidence matrix **Y**; **s** is the vector of random specific combining ability effects (SCA) with its incidence matrix **Z**; and **e** is the vector of residuals.

The estimates of the fixed and random design and genetic effects are obtained by solving the mixed model equations (MME's) (Henderson 1984):

$$\begin{bmatrix} \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{Y} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\left[\mathbf{I}\otimes\mathbf{G}_r\right] & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Y}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{Y}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Y}'\mathbf{R}^{-1}\mathbf{Y}+\left[\mathbf{A}\otimes\mathbf{G}_a\right] & \mathbf{Y}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\left[\mathbf{I}\otimes\mathbf{G}_s\right] \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \hat{\mathbf{r}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{X}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Y}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

where the new terms represent variance-covariance matrices of the error (**R**), random design effects ($\mathbf{G_r}$), additive genetic effects ($\mathbf{G_a}$), and specific combining effects ($\mathbf{G_s}$) and the relationships between the additive genetic effects (**A**, the additive (or numerator) relationship matrix) and independent random effects (**I**), and $\otimes$ is the Kronecker product. The **R** and $\mathbf{G_r}$ matrices are specific to each site. The solutions to the mixed model equations give the highest correlation between true and estimated values, provided that the variances and covariances are known (Mrode 1998).

The traits (termed selection criteria) defined in the model are an amalgamation of data from many sites, where these data can reasonably be expected to represent the same trait. Genotype by environment interaction (GxE) is accommodated by treating different site types as different traits (Falconer and MacKay 1996) and we use the same principle to deal with indirect selection based on measurements early in the growth cycle. In comparison with treating each trait on each site as a separate trait, this approach reduces computational complexity and ensures that estimates of variances and correlations are available for $\mathbf{G}_a$. It also ensures that giving the breeder a comprehensible number of selection criteria to deal with. For the breeder, the results are further simplified by combining selection criteria into values at harvest (breeding objective traits) using the method of Schneeberger *et al.* (1992), and combining these into an economic index using economic weights derived for each client's production system using the method of Ponzoni and Newman (1989).

Data for each site is standardised to a unit additive variance to ensure that all data is on the same scale. This removes variance heterogeneity due to differences in productivity or units of measurement. Where trials are too small to yield reliable estimates of variances and correlations, models or standardised values are used instead. The use of **R** and $\mathbf{G}_r$ matrices specific to each site allow for differences in heritability while using design models specific to each trait on each site. The **A** matrix links all genotypes in the pedigree by the proportion of genes identical by descent (IBD) and allows predictions of breeding value for each genotype for each trait using the correlations in $\mathbf{G}_a$. The inverse of *A* derived from simple rules (Henderson 1976) is used in the MME's, and these can be modified for fixed group effects due to origin or selection history (Westell *et al.* 1988), partial selfing (Dutkowski *et al.* 2001), and pollen mixes (Perez-Enciso and Fernando 1992). The software is very efficient as it uses an equivalent gametic model for trees without offspring (the majority) (Quaas and Pollak 1980), running jobs with 20 selection criteria and over 200,000 genotypes in less than 15 minutes.

RESULTS

Increased selection efficiency

Historically, in common with many breeding programs, progeny trials have been analysed individually using univariate family models. Retrospective analysis of 30 years of trials of *P. radiata* using TREEPLAN® indicates that efficiency of selection in the past has been as low as 30% (Table 1) when ranked on an index adopted in the 1990s (White *et al.* 1992). The efficiency is calculated from the average breeding value of the selections that were crossed relative to a notional breeding population of 300 selected from the top 1000 genotypes planted in each decade. While selections from the 1990s trials have not yet been selected, it will presumably be much closer to the best 1000 than had previously been achieved. The extra gain will come from both using the integrated data, as well as appropriately weighting each trait. On this index it seems that there has been historically high emphasis on stem quality (confirmed by anecdotal evidence) and very little on branch quality, with negative efficiency reflecting crosses less than the trial averages.

Table 1. - Historic selection efficiency in *Pinus radiata*.

Trait	Decade	Trial Average	Crossed	Top 1000	Efficiency
Index	1970s	19.1	58.6	117	41%
	1980s	*10.8	40.2	118	27%
	1990s	49.1		178	
Volume	1970s	14.2	47.5	61.2	71%
	1980s	12.0	41.7	59.0	63%
	1990s	26.7		103	
Stem Quality	1970s	-0.04	0.15	0.10	134%
	1980s	0.02	0.08	0.18	40%
	1990s	0.11		0.21	
Branch Quality	1970s	0.07	0.01	0.42	-17%
	1980s	-0.02	-0.06	0.41	-9%
	1990s	0.12		0.52	

*Trial average declines in the 1980s due to new infusions.

Better correlation with large plot performance

Results from a large plot realised gains trial of *P. radiata* indicates that the performance of the TREEPLAN® breeding values is in the main good, but with poor prediction for some families (Figure 1). Importantly, projection of the stand growth measurement to age 20 showed that the predictions from the breeding values based on individual tree measurement in small plots were not inflated (not shown). The performance of the families was better predicted than from historic family model breeding values derived using BLP (White *et al.* 1992) (Figure 1).

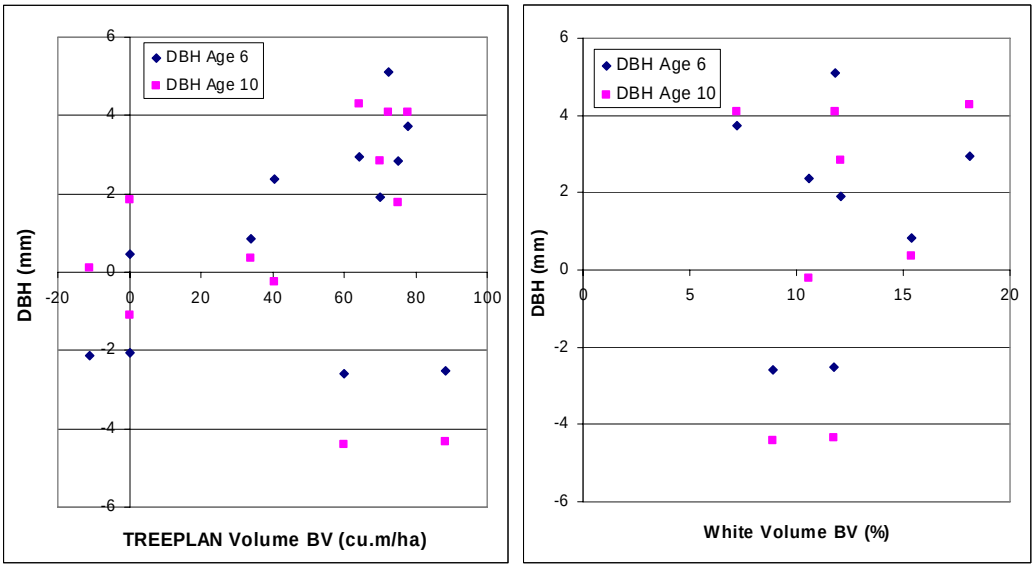

Figure 1 - Realised gain trial family growth performance for DBH at ages 6 and 10 years predicted from TREEPLAN® mid-parent harvest volume breeding values from mid-parent family model volume breeding values by White *et al*. (1992) using BLP.

Additive standardisation is a good approach

The TREEPLAN® system amalgamates data across sites and standardises it to a unit additive variance, but it allows each site to have its own error variance, and thus its own heritability. A test data set of three trials of *E. globulus* with little GxE (r_g>0.9) measured for growth at age 4 was examined to look at the efficacy of this approach. Constraining the inter-site genetic correlation to be one (model r_g=1) resulted in only a small (but significant) decrease in likelihood (6.4) when compared to the full multivariate model (Full), confirming low GxE (Table 2).

Table 2 - Relative log-likelihood for different models and data standardisations.

			Data standardisation			
Model	Traits	Equations	None	$1/SD_a$	$1/SD_e$	$1/SD_p$
Full	3	72050	6.4	6.4	6.4	6.4
r_g=1	3	72050	0	0	0	0
V_a=V_a,V_e=V_e	1	40104	-499	-407	-55	-43
V_a=V_a	1	40104	-71	-0.8	-27	-24
V_e=V_e	3	72050	-51	-151	-0.4	-1.7

This likelihood difference was not affected by the data standardisation used: standardising by the additive ($1/SD_a$), error ($1/SD_e$) or phenotypic standard deviation ($1/SD_p$). Various simpler models were tried by setting the error or additive variances to be the same across sites (V_x=V_x), or allowing them to be different. Models with a single additive variance (V_a=V_a) can be treated as a univariate genetic analysis, dramatically reducing the number of equations needed to solve the MME's, and thus making large problems computationally feasible. In general, where the data standardisation matched the model, there were small changes in the likelihood (Table 2). The TREEPLAN® approach of standardising by the additive variance and allowing a separate error variance for each site gave a likelihood not very different from the r_g=1 model (-0.8), but did so with a much reduced number of equations, indicating a good balance between model complexity and computational efficiency.

Spatial adjustment increases gain

While spatial models as advocated by Gilmour *et al.* (1997) and Dutkowski *et al.* (2002) are soon to be incorporated into TREEPLAN®, data adjusted for such models has been used for a large *E. nitens* breeding value prediction project. Where such models are used, most design features are eliminated, reducing model complexity. In common with the work of Dutkowski *et al.* (2005), most of the selection gains are small, but in some instances they can be large (Figure 2), giving more gain without the need for extra replication or trials.

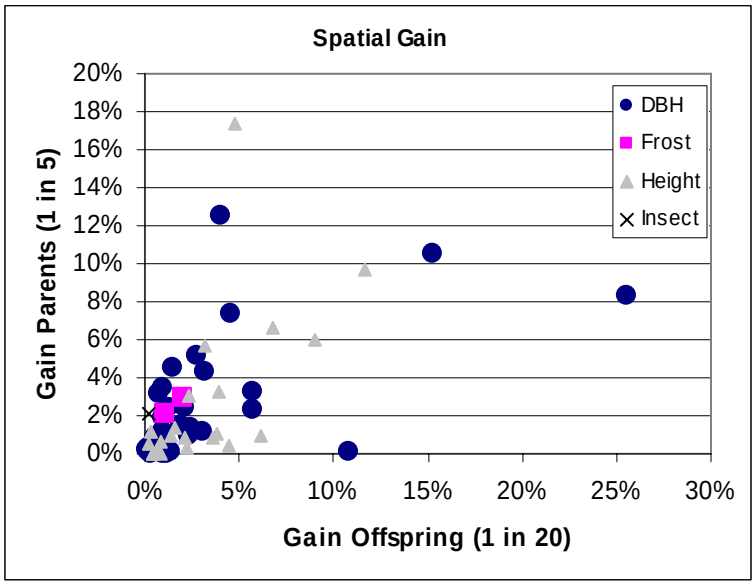

Figure 2 - Relative selection gain from spatial analysis for *Eucalyptus nitens* trials. DBH and Height are growth measures. Frost and Insect are damage measures.

Selection for different site types

In a *Eucalyptus nitens* breeding population with 2 generations of trials (3 generations including base parents), sites were stratified into cold and normal site types, with frost damage scores on some of the cold sites. In the first generation trials, the majority of trials were of the normal site type (and these had higher survival) and most selections for growth on normal sites came from six of the eight normal trials (Table 3). There was a clear difference in growth on cold sites between races, and most selections for cold sites came from one race.

However, while all cold site trials had selections made in them, the majority of cold site selections actually came from this cold tolerant race and its families on the normal sites, where there were more trees. The best selections with two generations of data for the cold site types include many (35%) from the first generation trials on cold sites. In contrast, selections for the normal site types are dominated by the second generation trials (Table 3). This situation is not unexpected as no specific crosses had been made between cold site type selections – the families planted in the second generation trials were not improved for growth on cold sites and therefore many of the good selections were still in the first generation trials. Using all of the data together allowed appropriate selections to be made, taking into account the crossing history.

Table 3 - Selections for different site types from trials on different site types with one or two generations of data.

Trials indicates the proportion of trials with selections in the top 100 for that site type. Trees indicates the proportion of the top 100 selected for that site type from trials on that site type.

				Selections for site types			
Generations of data	Trial site type	Generation	Trials (n)	Cold		Normal	
				Trials	Trees	Trials	Trees
1	Cold	1	3	100%	38%	100%	9%
	Normal	1	8	75%	62%	75%	91%
2	Cold	1	3	83%	35%	0%	0%
		2	8	25%	2%	38%	8%
	Normal	1	8	60%	20%	10%	1%
		2	28	32%	43%	50%	91%

Linking trials through offspring

Having different subsets of genotypes in different trials is usual in breeding programs and it makes comparisons between genotypes difficult. For base population trials, differences in composition may make it very hard to estimate groups differences, and such comparisons are essential for making good selections. For a *E. globulus* breeding program in South America (Sanhueza and Griffin 2001), local land race material had been tested in one set of trials, and introductions of Australian material had been tested in another set of trials. There were only four Australian family seedlots in common, and these were raised in a separate nursery from the land race material but planted in those trials. The Australian seedlings were of relatively poor quality and had grown very poorly in the land race trials, leading to relatively poor

performance and only 20% overlap between growth breeding values of the two populations. It had long been suspected from operational plantings that these two populations had more similar growth and that these first generation results were biased. Subsequently, crosses were made from selections in both sets of trials and their offspring planted together in second generation trials. Joint analysis of first and second generation trials showed that these populations were not as different as first generation analysis had shown, with around 80% overlap of growth breeding values now apparent. The linkage between these genetic groups through their offspring was strong enough to overcome the effect of the poor controls and give more realistic group difference estimates.

DISCUSSION

While the models used by TREEPLAN® are arguably the most comprehensive currently used in tree breeding, it is the potential gains from using them that is of real interest to plant breeders. In a number of operational breeding programs we have shown that there are gains to be made through data integration, spatial models, and application of economic values to the derived breeding values. Data integration using the strategies in TREEPLAN® allows appropriate selections to be made across different trials. Whether such gains can be made in all breeding programs will depend on the history of selection, crossing and testing in each program. However, the problems that we have encountered and, apparently, overcome are probably common in many plant breeding programs.

Conclusive evidence about the benefits of using these models is hard to find in operational breeding programs. However, simulation studies usually lack the nuances and imbalances found in the complex selection and testing histories of operational programs. Retrospective analysis of past selections shows that efficiency has been lower than was anticipated. Evidence of increased rates of gain will only be demonstrated after several generations. This will probably not be done initially in trees because of the long generation interval in these breeding programs. However, the current selections being made in *Pinus radiata*, and their subsequent crossing and testing (with appropriate controls) may well provide such definitive evidence, and we look forward to presenting this evidence in years to come at a subsequent plant breeding conference.

Apart from the increased rates of gain in breeding programs that we think are likely with integrated analysis, our experience is that there are certainly operational efficiencies to be gained. Data management and analysis is no longer a bottleneck in the breeding programs. In the past, selection and crossing decisions have been delayed due to the inability of the breeders to carry out integrated analysis. Ad hoc selections were sometimes made from simple univariate analysis of each trial, without the ability to consider the relationships between traits and the relative quality of the genetic material in each trial for each trait. For those who use the TREEPLAN® system and its supporting database, there is no going back to the way that things used to be done.

ACKNOWLEDGEMENTS

We would like to thank Forestal Monte Aguila (Chile) for permission to show partial results from a breeding value prediction project for that company.

REFERENCES

Dutkowski GW, Costa e Silva J, Gilmour AR, Lopez GA (2002) Spatial analysis methods for forest genetic trials. *Canadian Journal of Forest Research* 32, 2201-2214.

Dutkowski GW, Costa e Silva J, Gilmour AR, Wellendorf H, Aguiar A (2005) Spatial analysis enhances modelling of a wide variety of traits in forest genetic trials. *Canadian Journal of Forest Research* accepted.

Dutkowski GW, Gilmour AR, Borralho NMG (2001) Modification of the additive relationship matrix for open pollinated trials. In 'Developing the Eucalypt of the Future'. Valdivia, Chile. (INFOR)

Falconer DS, MacKay TFC (1996) 'Introduction to Quantitative Genetics.' (Longman Scientific and Technical: New York)

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics* 2, 269-293.

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. *Biometrics* 32, 69-83.

Henderson CR (1984) 'Applications of Linear Models in Animal Breeding.' (University of Guelph: Guelph)

Jarvis SF, Borralho NMG, Potts BM (1995) Implementation of a multivariate BLUP model for genetic evaluation of *Eucalyptus globulus* in Australia. In 'Eucalypt Plantations: Improving Fibre Yield and Quality'. Hobart, Tasmania. (Eds BM Potts, NMG Borralho, JB Reid, RN Cromer, WN Tibbits and CA Raymond) pp. 212-216. (CRC for Temperate Hardwood Forestry)

Kerr RJ, Dutkowski GW, Apiolaza LA, McRae TA, Tier B (2002) Developing a genetic evaluation system for forest tree improvement - the making of TREEPLAN®. In '7th World Congress on Genetics Applied to Livestock Production'. Montpellier, France

Kerr RJ, McRae TA, Dutkowski GW, Apiolaza LA, Tier B (2001) TREEPLAN - a genetic evaluation system for forest tree improvement. In 'Developing the Eucalypt of the Future'. Valdivia, Chile. (INFOR)

McRae TA, Dutkowski GW, Pilbeam DJ, Powell MB, Tier B (2004) Genetic Evaluation Using the TREEPLAN® System. In 'Forest Genetics and Tree Breeding in the Age of Genomics: Progress and Future'. Charleston, South Carolina. (Eds B Li and S McKeand). (North Carolina State University)

Mrode RA (1998) 'Linear Models for the Prediction of Breeding Values.' (CAB International: Wallingford, U.K.)

Perez-Enciso M, Fernando RL (1992) Genetic evaluation of uncertain parentage: a comparison of methods. *Theoretical and Applied Genetics* 84, 173-179.

Ponzoni RW, Newman S (1989) Developing breeding objectives for Australian beef cattle production. *Animal Production* 49, 35-47.

Quaas RL, Pollak EJ (1980) Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science* 51, 1277-1287.

Sanhueza R, Griffin AR (2001) FAMASA Fiber Yield Improvement Programme (F.Y.I.P.): 10 years experience breeding *Eucalyptus globulus*. In 'Developing the Eucalypt of the Future. IUFRO International Symposium'. (INFOR, Chile)

Schneeberger M, Barwick SA, Crow GH, Hammond K (1992) Economic indices using breeding values predicted by BLUP. *Journal of Animal Breeding and Genetics* 109, 180-187.

Westell RA, Quaas RL, van Vleck LD (1988) Genetic groups in an animal model. *Journal of Dairy Science* 71, 1310-1318.

White TL, Rout AF, Boomsma DBB, Dutkowski GW (1992) 'Predicted breeding values of 1213 first generation parents.' Southern Tree Breeding Association, TR92-02, Mount Gambier.