TREEPLAN® - A Genetic Evaluation System for Forest Trees

[1]McRae T. A., [3,4]Apiolaza L. A., [3]Dutkowski G. W., [2]Kerr R. J.,
[1]Pilbeam D. J., [1]Powell M. B. and [2]Tier B.

Abstract: -- The TREEPLAN® genetic evaluation system is designed specifically for the efficient and accurate prediction of breeding and other genetic values in trees. TREEPLAN® uses the preferred statistical method of best linear unbiased prediction (BLUP) using an individual tree additive genetic effect. Although BLUP methods are well developed theoretically, other software is suitable only for breeding value estimation and prediction on small and/or highly structured (balanced) data sets. Packages such as ASREML and SAS have hardware and software limitations that make them unsuitable for routine prediction on large data sets with complex pedigree structures and overlapping generations. TREEPLAN® fits a reduced individual tree model for purposes of efficiency. TREEPLAN® can model multiple genetic groups, handle clonal data, fit multi-trait models with more than 50 traits, accommodate heterogeneous variances, fit site specific statistical and genetic models, and also weights information across environments (accounts for genotype by environment interaction) and time (allows for age:age correlations).

The Southern Tree Breeding Association (STBA) is routinely using TREEPLAN® for genetic evaluation in Australian tree improvement programs for *Pinus radiata, Eucalyptus globulus* and *E. nitens.* TREEPLAN® has allowed data across generations and years to be combined in a multi-trait analysis to produce single lists of breeding values for each trait and environment combination. TREEPLAN® is easy to use and has the 'industrial strength' to handle large amounts of unbalanced data with the complex pedigree structures that are usually associated with national or regional tree improvement programs. TREEPLAN® is fully integrated with a web based data management system that efficiently handles data and pedigree information. The analytical power and flexibility of the TREEPLAN® system has made routine genetic evaluation in trees a straightforward process.

INTRODUCTION

The total plantation estate in Australia is 1.63 million hectares (National Plantation Inventory 2003). The Southern Tree Breeding Association (STBA) runs the national breeding cooperatives for *Pinus radiata* and *Eucalyptus globulus.* These two species comprise about two-thirds of the national estate, and are mostly used for solid wood products and pulp and paper production.

Tree improvement programs fundamentally consist of (i) defining a breeding objective,

---

[1] Southern Tree Breeding Association Inc., Mount Gambier, SA, Australia. [2] Animal Genetics and Breeding Unit, University of New England, Armidale, NSW, Australia. [3] Cooperative Research Centre for Sustainable Production Forestry, University of Tasmania, Hobart, Australia. [4] Current address: Forestry Tasmania, Hobart, TAS, Australia.

(ii) mating among parents, (iii) testing offspring in field trials, (iv) analysing performance data and genetic evaluation, (v) selecting trees for deployment and further breeding with elite parents. In general, tree breeders have been proficient at handling the biological aspects of tree breeding and trial establishment. However, they have often failed to optimise in a timely manner genetic evaluation using pedigree and correlated performance information. That is, it is relatively easy to plant and assess trees in progeny trials to generate data. However, it is often much more difficult to process the data collected in an efficient and comprehensive manner. As a consequence, the STBA, like many other tree improvement programs, had access to many records (unprocessed data) from research and breeding trials that did not meet the usual restrictive requirements of a simple analysis.

Historically, tree breeding has emphasised experimental design features (replicates, plots and incomplete blocks in increasingly complex designs) in trees to account for local environmental effects, compared to more complete modelling of the genetic components. Single-generation, single-site and single-trait mixed models have thus been the norm in tree breeding. This has allowed the use of straightforward methods of analysis, including best linear prediction (BLP), without a numerator relationship matrix. Family models have largely been used with a second stage to predict within-family values.

The STBA adopted the individual tree additive genetic model (ITM) BLUP in its tree improvement programs during the 1990s (Jarvis *et al*. 1995). However, its application was limited to relatively small and uncomplicated data sets until the development of the TREEPLAN® system. The application of such a model occurred later in tree breeding, and is much less common than in animal breeding. This situation has arisen because breeding programs for trees are usually in their early generations, with simple shallow pedigrees, and trees are evaluated in large designed trials. Families are often the result of open-pollination, such that simpler family models are possible for the prediction of parental breeding values. Unlike animals, trees are often not subject to culling, so that data sets are more balanced. The magnitude of genotype by environment interactions (GxE) is often unknown, except in a large environmental range. The number and type of traits measured is usually limited, but is rapidly evolving as wood quality traits assume greater importance.

Modern tree improvement programs demand a greater use of BLUP to predict genetic values for several reasons. Breeding programs are progressing and now span several generations. Individual programs with different samples of the same base population are being consolidated into larger cooperatives. It is important to account for the effects of selection over time. Many programs are now making the transition to overlapping generations, where a proportion of all breeding activities is performed each year, and all families are not tested at all test sites at the same time. Finally, there is a need for integrating all data between trees and between traits, making it easier for selection and to monitor the genetic progress of breeding programs.

Currently, the STBA is collecting performance data in trials on third-generation progeny in *P. radiata* and second-generation progeny in *E. globulus*. In the past, breeding values

were estimated using BLP for *P. radiata* (White *et al.* 1992 ab) and BLUP for *E. globulus* (Jarvis *et al.* 1995). Due to a lack of suitable BLUP software, multiple and independent lists of breeding values made it difficult to compare trees for genetic merit across a population. Despite the existence of good genetic linkage, pedigrees were too complex to be accommodated. Large quantities of data were also excluded because trial assessments were incomplete or done at different ages. That is, the data were 'messy' or did not fully satisfy other restrictive requirements of 'balance'.

This inefficient use of data and information is clearly undesirable, particularly for large national breeding cooperatives. In order to overcome this weakness, the STBA designed TREEPLAN® to apply 'industrial strength' individual tree model BLUP on a program wide basis. Although the STBA and AGBU initially developed the TREEPLAN® system for use in the Australian tree improvement programs for *P. radiata* and *E. globulus*, it has been designed with flexibility for much wider application.

This paper discusses some of the key features of TREEPLAN® and its routine application of BLUP in forestry.

The Genetic and Statistical Models

The statistical approach used in TREEPLAN is designed for maximal efficiency as it includes all the design effects used in simpler analyses, but can incorporate all of the data that has been collected in a single analysis – combining different traits and across all pedigrees. It fits a linear mixed model of the form:

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

where: y is the vector of observations on one or more traits; f is the vector of fixed site and design effects, 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 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) using Gauss-Seidel iteration:

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

where, the new terms represent variance-covariance matrices of the error (R), random design effects (G_r), additive genetic effects (G_a), and specific combining effects (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.

This model offers substantial advantages over the models usually used in forest genetic trial analysis. Breeding values (and other genetic effects) are estimated for all traits, for all trees in the pedigree – both parents and offspring, in a single analysis. Where a trait has not been measured on a tree then the best prediction is made of its breeding value using information from relatives and from traits correlated at the genetic, design or error levels. If there is no such information, then the estimate is at the population mean, but the variance of the estimates grows as the amount of information, and thus its reliability, increases. The use of correlated traits allows correction for the effect of selection in measurement, as long as the data used for selection is included. The solutions give the highest correlation between true and estimated values, provided that the variances and covariances are known. This is a substantial improvement over BLP, where the fixed effects are assumed to be known. The mixed model equations are extremely robust, and can be readily extended to more complex models.

The model uses the A matrix to track the proportion of genes in common between trees in the pedigree and gives solutions for all of them without any secondary process of the data in what has been called an individual tree model (ITM). It easily handles half-sib and full-sib pedigrees, and simple rules have been worked out (Henderson 1976) to create the inverse that is used in the MME's. The matrix can be modified for the types of pedigrees that are common in forest genetic trials: fixed provenance or selected parentage (such as seed orchard) effects (Quaas 1988), partial selfing (Dutkowski and Gilmour 2001), and even pollen mixes (Perez-Enciso and Fernando 1992).

The software uses an equivalent gametic model for computational efficiency in the prediction of breeding values for trees without offspring (the majority).

$$y_i = \mu_i + \frac{u_f + u_m}{2} + s_j + \phi_i + e_i$$

where: μ_I is the mean, y_i, s_j and e_i are as defined above, u_f and u_m represent the breeding values for the tree's female and male parents respectively, and ϕ_i represents Mendelian sampling in the formation of the tree's genotype. That is, $.5u_f + .5u_m$ represent "average" gametes from each parent, and ϕ_i represents the deviation from the average of the gametes received by the progeny. The genotypic and gametic models are equivalent models, in that the solutions to the unknowns will be exactly the same for both models. Their combined use is called a "reduced" individual tree model.

Trait Mapping to Selection Criteria

In theory, the MME's can handle all data by treating each measurement on each site as a separate trait, as long as all the variances and correlations are known. In practice, however, such and approach is computationally infeasible, not all variances and correlations are known and dealing with output would be very confusing to the breeder, because of the many traits. The mapping of multiple measured traits to a smaller meaningful number of selection criteria (SC) traits is a feature of TREEPLAN®. This allows a reduction (consolidation) in the number of traits for which breeding values are predicted in a multi-trait analysis. This mapping gives TREEPLAN® its flexibility and ease of use as the breeder can easily define the SC traits of interest. The mapping allows

us to consolidate data with different forms and scales of measurement, different ages and different sites, as long as it can be realistically assumed that all the measurements have a sufficiently high correlation to be treated as one. For example, if diameter at breast height (DBH) is measured between ages 3 to 12 years, then a sensible strategy is to propose three SC traits: DBH ≤4 yrs, DBH 5–8 yrs and DBH 9-12 years. We recommend only mapping traits displaying significant genetic variance in a single-site analysis.

Heterogeneous Variances

Breeding programs collect data from trials spread across a diverse range of site types and age classes. Some traits are or have been assessed using different protocols. For example, growth may have been measured as tree height, stem diameter or tree volume; and stem form using several scales with different levels of precision. The variance of performance traits such as growth usually increases with size, growth rate and age of trees. A linear transformation of the data such that the phenotypic variance is unity is an approach often used in plant and animal breeding to make variances homogeneous. A disadvantage of this approach for tree breeding is that a constant heritability would need to be assumed across all sites, despite some sites being more homogeneous. Tree breeders also have the benefit of large designed trials that provide good estimates of variances and spatial variability (replication and blocking), genetic and residual variances and correlations specific to each site. TREEPLAN takes advantage of the availability of these estimates to overcome these problems by: (i) transforming the data for each trait to unit additive variance on a site by site basis; and (ii) using the within site error (to allow for different heritabilities) and significant design factor (eg. rep, plot and incomplete block) variances in the BLUP analysis.

Genotype x Environment Interaction

As well as age differences, geographical location and/or site type are other possible criteria for proposing new SC traits out of the one generic trait such as growth. For example, it may be necessary to partition the SC trait, DBH ≤4 yrs, further in a multi-site run, according to province, state or soil type. GxE interaction is where different environments induce different kinds of genetic variance to be displayed. That is, GxE may result in a change of ranking of genotypes across environments. However, GxE due to scale effects is effectively removed by data transformation (standardisation). Flexibility in mapping of traits in TREEPLAN accommodates specific geographical and environmental combinations by creating environmental subclasses.

In practice, the best method to handle GxE is to consider the same character measured in two different environments as two different but correlated traits (Falconer and Mackay 1994). A trait measured at different locations can be considered biologically the same SC trait when the genetic correlation is high (for example, ≥0.8). A breeder can either define different production environments or ignore GxE (effectively selecting for general adaptation) if environmental effects are not repeatable. Past studies to quantify the magnitude and nature of GxE in Australia for *P. radiata* and *E. globulus* have been based

on limited data sets. Studies with more extensive data sets are currently under way to estimate across site correlations and better define the target production environments.

Genetic Groups

In forestry, parents of first-generation progeny are typically trees from native stands (or plantations) sampled from many different geographical regions that represent different provenances or races. Because provenances are quite genetically distinct it is important to assume that $E(g) \neq 0$, where g is the vector of genetic values. Male parents are usually unknown and female parents are assumed to be unrelated. Seeds from the female parents (founders) are collected from various localities spread across a wide geographical area. Thus, it is reasonable to consider that progeny are from more than one genetically divergent sub-population. TREEPLAN relates all foundation parents on the basis of their original provenance to genetic groups. In practise, data sets are likely to be far more complex. For example, a male parent (pollen) might be identified as belonging to a particular population, such as, a routine or an improved population. Founders introduced from another unrelated breeding program might also constitute a different genetic group. The modified mixed model equations of Quaas (1988) are used to derive solutions to g.

Clonal Data

Individual trees can be replicated using various forms of vegetative propagation. Clonal tests are common in *P. radiata* and are also used in some Eucalypt breeding programs. TREEPLAN currently treats clones as the same individual and matches unique clone identities to a single genotype. Clonal replication can improve the precision of breeding values. Versions of TREEPLAN currently being developed will be capable of predicting genetic values, including additive and non-additive genetic effects, for individual clones, recognising the potential for somaclonal variation and propagation effects. This functionality is particularly important for deployment of clones.

Partial Selfing in Open Pollinated Seed

Trees can be partially self-fertile, generating pedigrees where two progeny may be selfed sibs (both progeny result from selfing), a selfed sib and an outcrossed sib, full-sibs or half-sibs. In the *E. globulus* breeding program most progeny tested in the first-generation are derived from open-pollinated seed collected from founder trees in native forest stands. Until many more second-generation progeny (from controlled pollination crosses) are included in the analysis, the accuracy of breeding value prediction is dependent on how well the relationship coefficients between sibs of open-pollinated trees can be defined. Dutkowski (2001) has outlined simple rules to modify the NRM when a selfing rate in native stands is assumed. These rules can be further extended to account for the equilibrium level of inbreeding in the stand and the level of coancestry in the trees local to the female parent from which seeds were collected. Sparse stands of trees are expected to have a higher level of inbreeding among the progeny than dense stands. This functionality is currently being implemented in TREEPLAN.

Running TREEPLAN®

An efficient data management system is critical for accessing data and pedigree information to produce breeding values quickly. The TREEPLAN® analytical system is fully integrated with a modern data management system (STBA-DMS) which operates via a web based interface. TREEPLAN® can be run independently of the STBA-DMS, but its interactive nature makes the process of genetic evaluation far more straightforward and efficient. It also facilitates data entry and analysis from various locations. The STBA-DMS is mainly designed for storage and retrieval of tree data for the purposes of genetic evaluation. It is flexible and accommodates different species of trees. User access is restricted and data is password protected to the level of traits within trials. This allows us to easily complete multiple TREEPLAN® runs for the membership, firstly using only generic data, but then also including data for traits belonging to a restricted group of clients. This provides the flexibility needed in large cooperative tree improvement programs to satisfy individual client needs and produce customised breeding values.

TREEPLAN® extracts genetic parameters, data and run specifications from the STBA-DMS. Making changes to specifications for a new TREEPLAN® run is a simple process. That is, it is a straightforward process to include (exclude) new trials and/or more traits in a multi-trait BLUP analysis. As new trials are assessed, the data is validated and entered. Multi-variate analyses are first done on a trial by trial basis using ASREML and the variances and correlations for all significant design and random genetic components are stored in the STBA-DMS. The system is designed to regularly update breeding values. That is, as quickly as a trait is measured, data entered and single site analysis completed, TREEPLAN® is then run with the complete database.

Genetic Evaluation in *E. globulus* and *P. radiata*

TREEPLAN® is being used routinely to predict genetic (breeding and deployment) values for trees included in the *E. globulus* and *P. radiata* databases. As new trials and traits are assessed, the data is entered into the database, analyses are done on a single site basis and parameters estimated, TREEPLAN® is run, and breeding values for all trees in the specified population are updated. Table 1 lists details of data sets used in recent runs of TREEPLAN®.

Pinus radiata. Breeding values were predicted for 117,778 genotypes (different trees) in the population. This included trials from the southern States of Australia (Powell *et al.* 2002). The inclusion of many (hundreds) outstanding historical first- and second-generation trials yet to be entered in the database, will be done as resources are made available. At this stage, breeding values are predicted for Selection Criteria targeting the different production regions defined in the National Plantation Inventory for Australia (Wood *et al.* 2001). Selection Criteria traits for growth include: six production regions by four age classes (0-5 yrs, 6-12 yrs, 13-24 yrs and >24 years). Branch angle, branch quality, branch size and stem straightness comprise the form traits. Basic density (0-12

yrs and >13 years) and Spiral Grain (0-6 yrs and ≥6 years) constitute wood quality traits. Data for disease and pest resistance/tolerance traits will be incorporated with time.

Eucalyptus globulus. Beeding values were predicted for 174,369 genotypes in the population. This included trials from South Australia, Tasmania, Victoria and Western Australia (Pilbeam *et al.* 2002). A rolling front is used with some breeding, assessment and selection activities done on an annual basis. Prediction of breeding values is a dynamic process, such that TREEPLAN® breeding values are updated regularly as traits are measured, data compiled and validated. At this stage, breeding values for growth are predicted in four production regions by three age classes (0-4 yrs, 5-8 yrs and 9-12 years). Basic density, by two age classes, and pilodyn penetration comprise quality traits. Data for pest and disease resistances (defoliation), kraft pulp yield, NIRA pulp and cellulose content, collapse, shrinkage and tree form traits will be incorporated with time. Trees in the CSIRO collections (Gardner and Crawford 1987, 1988) will be used to establish a baseline for monitoring genetic improvement over time.

	Species	
	Pinus radiata	*Eucalyptus globulus*
Generations	3	2
Trials included in Analysis	68	87
Number of Selection Criteria Traits Analysed	19	10
Genetic (founder) Groups fitted	12	25
Families	3033	1550
Genotypes included in Analysis	117,778	174,369

Table 1. Data sets used in recent runs of TREEPLAN® for *P. radiata* and *E. globulus*.

Future Enhancements

In partnership with the Forest and Wood Products Research and Development Corporation (FWPRDC), STBA and AGBU plan to develop Version 2 of TREEPLAN®. Additional features will include: (1) Better modelling of intra-site environmental variation using spatial and competition models, (2) Incorporation of information at the DNA level (markers and candidate genes), (3) Modelling of dominance and epistatic effects to allow for the full exploitation of these non-additive genetic effects in clonal deployment populations, and (4) Development of a clearer understanding of GxE to better target different production environments.

CONCLUSIONS

Tree breeding programs have evolved to the stage where the adoption of BLUP is required to maximise return on investment through breeding. TREEPLAN® is a genetic evaluation system that facilitates the routine application of individual tree model BLUP

to forest tree data. TREEPLAN® can model multiple genetic groups, handle clonal data, fit multi-trait models with more than 50 traits, accommodate heterogeneous variances, fit site specific statistical and genetic models, and weight information to account for age-age correlations and genotype by environment interaction. TREEPLAN® has allowed data across generations and years to be combined in multi-trait analyses to produce breeding values for each trait and environment combination of interest on a program basis. TREEPLAN® is easy to use and has the 'industrial strength' and speed to handle large amounts of unbalanced data with complex pedigree structures. TREEPLAN® is fully integrated with a web based data management system that efficiently handles data and pedigree information. The TREEPLAN® system is being used routinely to update breeding values in the Australian tree improvement programs for *P. radiata* and *E. globulus*. TREEPLAN® also facilitates the adoption of efficient rolling front breeding programs with overlapping generations.

ACKNOWLEDGEMENTS

REFERENCES

Borralho, N. M. G. and Dutkowksi, G. W. (1998). Comparison of rolling front and discrete generation breeding strategies for trees. Canadian J. Forest Res. 28: 987-993.

Dutkowski, G. W. and Gilmour, A. R. (2001). Modification of the additive relationship matrix for open pollinated trials. In 'Developing the Eucalypt of the Future' 10-15 September 2001, Valdivia, Chile p. 71 (Instituto Forestal Chile).

Gardner, C. A. and Crawford, D. F. (1987). 1987 Seed Collections of Eucalyptus globulus subsp. globulus for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forest Research, Report, Canberra.

Gardner, C. A. and Crawford, D. F. (1988). 1988 seed collections of Eucalyptus globulus subsp. globulus for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forestry and Forest Products, Report, Canberra.

Gilmour, A. R., Cullis, B. R., Welham, S. J. and Thompson, R. (2001). ASREML Reference Manual. September 24, 2001.

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