TREEPLAN®- A GENETIC EVALUATION SYSTEM FOR FOREST TREE IMPROVEMENT

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

INTRODUCTION

TREEPLAN® is a genetic evaluation system for forest tree improvement. The system is designed specifically for the efficient and accurate prediction of genetic values of trees for breeding and deployment purposes. TREEPLAN® uses best linear unbiased prediction (BLUP). BLUP has important statistical advantages over the more traditional methods used in tree breeding as BLUP allows for the comparison of individuals across space and time, regardless of the environment in which the trees are grown. Although BLUP methods are well developed theoretically, software currently available is suitable only for breeding value estimation and prediction on ‘small’ and/or ‘well structured’ (balanced) data sets. Packages such as ASREML (Gilmour et al. 1999) and SAS/STATS (1991) have hardware and software limitations that make them unsuitable for the routine prediction of breeding values on very large data sets.

The Southern Tree Breeding Association (STBA) implements breeding programs in Australasia for *Pinus radiata* and *Eucalyptus globulus*. A feature of these programs is a ‘rolling front’ strategy with overlapping generations. Without suitable BLUP software it is difficult to combine information across locations, years and generations, particularly where the pedigree is complex. A lack of suitable BLUP software may have acted as an impediment to genetic progress in many tree improvement programs internationally. The STBA and the Animal Genetics and Breeding Unit (AGBU) have developed TREEPLAN® as an ‘industrial strength’ system for breeding value prediction using technologies developed and routinely used in animal breeding and livestock improvement.

BLUP IS BEST

Robinson (1991) describes BLUP as

> “a method for jointly estimating a set of random variables which are *linear* in the sense that they are linear functions of the data; *unbiased* in the sense that the average value of the estimate is equal to the average value of the quantity being estimated; *best* in the sense that they have minimum mean squared error within the class of linear unbiased estimators; and *predictors* to distinguish them from estimators of fixed effects”.

This definition was written for statisticians by a statistician. What is the significance of BLUP for practical tree breeders? Before this can be answered we need to define the problem.

The tree breeder wants to know the best trees (parents and offspring) for a suite of economically important traits. The breeder also needs to identify those families in which the parental genes combine well. A series of trials are established to provide phenotypic data on trees for partitioning genetic and environmental effects. Usually trials are established in more than one location to account for genotype by environment interaction and to ensure the security of the breeding population against catastrophic loss. Rolling front schemes (Borralho and Dutkowski, 1998) with overlapping generations have been introduced to reduce generation interval, even out work loads, and to provide new selections regularly, rather than episodically. For example, STBA has adopted rolling front strategies for *E. globulus* and *P. radiata*. As breeding schemes mature, trial data accumulate containing progeny information from more than one generation.

[1] **[1]Animal Genetics and Breeding Unit[*], University of New England, Armidale, New South Wales 2351, Australia**
[2]Southern Tree Breeding Association, PO Box 1811, Mount Gambier, South Australia 5290, Australia
[3]Cooperative Research Centre for Sustainable Production Forestry, GPO Box 252-12, Hobart, Tasmania 7001, Australia

Some traits are more expensive to measure than others and thus, only subsets of progeny are measured for particular traits. Data are generally unbalanced in that genes of some parents have greater representation in progeny trials than others, and there are varying amounts of information from different sites. Traits usually have a continuous scale, but occasionally are categorical or binary in nature. For some species, genotypes can be clonally replicated. Given the complex nature of forest tree trial data, BLUP is the best method of analysis.

BLUP predicts values for various genetic effects such as breeding values (general combining abilities) and family effects (specific combining abilities). Estimated breeding values (EBVs) and estimated family effects (SCAs) allow direct comparison of all trees and families in the population on a genetic basis, regardless of when and where the data were observed and whether or not the trees have been clonally replicated. BLUP also accounts for unequal information such as different numbers of progeny or trait records. The accuracy of each EBV and/or SCA will reflect the quality and quantity of information used to estimate them. For the EBVs to be accurate, it is important that all the information on which selection is based are included in the analysis (Sorenson and Kennedy, 1984).

BLUP can be described as the ultimate selection index as it uses the covariance between traits as well as the covariance between relatives. This facilitates the correct weighting of all the data to estimate breeding and genetic values. Jarvis *et al.* (1995) describe a multi-trait BLUP model in use for forest trees.

APPLYING BLUP TO FOREST TREES

Animal geneticists have been instrumental in the development of BLUP technology in the last 30 years. In applying BLUP technology to forest trees it is important for the tree breeder to recognise the notable differences between animal and tree breeding data. Fecundity is generally much lower in animals, hence it is the norm to have a high proportion of parents to progeny. In tree breeding it is the opposite. There is a high proportion of progeny to parents. Animal breeders generally are not attentive to experimental design issues, but are more concerned with the definition of cohorts of animals treated alike. In tree breeding the opposite applies. Experimental design is important and cohort definition is not an issue. Hence, in tree breeding the statistical modeling of the experimental design in BLUP is a major concern. Other important differences are that trees can be more easily cloned and are able to self-fertilize. There are many similarities too. The use of genetic groups to account for significant founder effects is important to both animal and tree breeders. The problem of heterogeneous variances is also relevant to both types of breeders. However, the strategies for overcoming heterogeneous variances in genetic evaluation can be quite different. Tree breeders have the benefit of large designed trials that provide variance components specific to each site. Site specific variances can be used to adjust the data prior to the BLUP analysis. In animal breeding, where data suitable for variance component estimation is less abundant, it is more expedient to adjust for heterogeneous variances within the BLUP analysis itself.

TREEPLAN® - BLUP SOFTWARE CUSTOMISED FOR FOREST TREES

TREEPLAN® is BLUP software developed by the STBA and AGBU specifically for use in trees. TREEPLAN® meets the following “industrial strength” specifications.

- TREEPLAN® is able to model multiple genetic groups to take account of founder effects.
- Clonal replication is common for *P. radiata* in and is used in some other Eucalypt breeding programs. Thus, TREEPLAN® is designed to include information on unreplicated individuals and clonal replicates.
- Multi-trait models are likely to be complex, possibly requiring up to 50 traits to be fitted, some of which are categorical in nature. TREEPLAN® has been designed to operate efficiently when there are more than 50 traits fitted in the model.
- Because trial data comes from a diverse range of environments, TREEPLAN® can accommodate heterogeneous variances, both environmental and genetic.
- TREEPLAN® uses mixed models that are flexible: breeders can choose to treat design factors, such as REP, BLOCK and PLOT as fixed or random. The choice of which design factors are random or fixed is specific to the trial.
- The STBA have adopted rolling front schemes for its *E. globulus* and *P. radiata* breeding populations. Trial data accumulates regularly throughout the year and needs to be analysed quickly and efficiently without necessarily requiring the input of highly trained geneticists. For this reason TREEPLAN® is designed to make complex statistical methods accessible and easy to use. The TREEPLAN® system has also been integrated with the STBA

data management system operated via a web based interface. This facilitates the regular update of breeding and genetic values as information is collected.

TREEPLAN®s uses a reduced individual tree model

Because there is a high proportion of progeny in tree breeding data, relative to parents, it is appropriate that TREEPLAN® uses the reduced model of Quaas and Pollack (1980). In the reduced model individuals are separated into two classes – those with progeny (parents) and those without (non-parents). Different but equivalent models are used to analyse each class. The model used for parents is

$$y_{ij} = b_i + u_j + e_j$$

where, y_{ij} is the observation on the *j*th parent, b_i is the *i*th fixed effect (for example, *i*th REP effect), u_j is the breeding value of the *j*th parent and e_j is the residual. The equivalent gametic model used for non-parents is:

$$y_{ij} = b_i + 0.5(f_j + m_j) + \varepsilon_j$$

where: f_j and m_j are the breeding values of the female and male parents, respectively, of progeny *j;* and ε_j is the modified residual containing e_j and the (uncorrelated) sampling variance of the gametes. That is, the breeding values of non-parents are expressed in terms of their parental breeding values. This tactic can substantially reduce the number of equations to solve. Hence, the requirement for computing resources, namely time and random access memory can be substantially reduced. Breeding values for non-parents can be obtained once the solutions for parents have been computed, using a procedure known as back-solving (for details see Quaas and Pollack (1980)).

TREEPLAN®s uses site specific environmental variances

The use of trial-specific residual error variances provides a simple solution to the problem of heterogeneity of error variances across trials. Each trial/site to be analysed in a multi-trial TREEPLAN® analysis should be first analysed separately to configure the best model in terms of what design effects are significant, and whether significant design effects are treated as fixed or random, and to provide the relevant variances.

TREEPLAN®s assembles measured traits into a reduced set of selection criteria traits.

The use of trial-specific genetic variances implies that the same trait measured in each trial is considered a different biological trait. This approach is not efficient as the number of traits fitted in a TREEPLAN® run could be in the hundreds. The tactic used by TREEPLAN® is to consolidate (reduce) all measured traits into a practical number of selection criteria traits (SC). The breeder is free to decide on the suite of sensible SC traits to include in the TREEPLAN® analysis, and must map each measured trait in each trial to one of these SC traits. It is likely that a trait such as diameter at breast height (DBH) is measured over a continuous range of ages. For example, if DBH is measured between 3 to 12 years, then a sensible strategy is to propose three SC traits: DBH < 4, DBH 5 – 8, and DBH > 12 years. 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 DBH. For example, it is possible that it would be necessary to partition the SC trait, DBH < 4, further, according to each province, state or soil type.

Investigation of the genetic correlation between similar traits measured in different trials is the best approach to constructing a sensible set of SC traits to include in TREEPLAN®. This implies that single-trait multi-site modeling work with a dedicated variance component estimation software package will be required. Once the mapping has been completed the information can be stored in a data base and accessed directly by TREEPLAN®.

There will still exist a considerable degree of heterogeneity of additive genetic variances among the measured traits that map to the same SC trait. TREEPLAN® uses an adjustment factor to correct each observation for any differences in additive genetic variances. A suitable adjustment factor is the inverse of the trial specific additive genetic standard deviation. However, the choice of any adjustment factor is left to the breeder.

A CASE STUDY

Consider the current STBA national *E. globulus* database. In a joint analysis of all trial data there will be a minimum of 22 SC traits (DBH in 4 states by 3 ages; DENSITY in 4 states by 2 ages; STEM and BRANCH). There are approximately 71,000 trees in the database consisting of 20 open pollinated, first generation progeny trials and 17 control pollinated, second generation progeny trials. There have been approximately 500 parents tested. If an individual tree model is used (software packages in common use such as ASREML and PEST use an individual tree model) the size of the partition of the mixed model equations relating to individual breeding values will be approximately 1,562,000. In TREEPLAN® the size of **this partition** is 11,000 because a reduced individual tree model is used and only parental breeding values are fitted. A section of the national *E. globulus* trial data was analysed using TREEPLAN® and ASREML (Gilmour et al. 1999). Approximately 12,000 progeny were analysed using a model containing only 7 SC traits (DBH in 4 states at one age, DENSITY at one age, STEM, BRANCH). Though ASREML is able to model heterogeneous error variances**,** homogenous residual error (co)variances across trials were used in the comparison for simplicity. Table 1 summarises the performance of the two programs.

	ASREML	TREEPLAN®
Memory usage (MBYTES)	139	113
Time (SECONDS)	185.1	5.7
Correlation between the EBVs predicted by the programs	.98	

The table shows that for this particular analysis TREEPLAN® uses 80% of the memory required by ASREML. The time taken to complete the analysis on a COMPAQ DS20E workstation is a fraction of the time needed by ASREML. These differences are likely to be magnified for larger, more complex models. The solutions obtained by both programs differed slightly, presumably due to rounding differences.

FUTURE ENHANCEMENTS

It is planned that version 2 of TREEPLAN® will have the following additional features:

1) Better modeling of intra-site environmental effects, (such as spatial analysis), that are particular to trees. This will better account for environmental heterogeneity;
2) Incorporation of information at the DNA level, that is, information on known quantitative trait loci (QTL); and
3) Modeling of dominance and epistatic effects to allow for the full exploitation of these effects by the STBA in breeding and deployment populations. Additionally, available clonal test data will be fully exploited in the model. Currently, in TREEPLAN® version 1, clonal data is used only to better estimate EBVs and SCA effects.

REFERENCES

Borralho, N.M.G. and Dutkowksi, G.W. (1996) Comparision of rolling front and discrete generation breeding strategies for trees. Canadian Journal of Forest Research 28: 987-993.

Gilmour A.R., Cullis B.R., Welham S.J. and Thompson R. (1999) ASREML Reference Manual. NSW Agriculture Biometric Bulletin No. 3.
Jarvis, S. F., Borralho, N.M.G. and Potts, B.M. (1995).Implementation of a multivariate BLUP model for genetic evaluation of Eucalyptus globulus in Australia. *In* ‘Eucalypt Plantations: Improving Fibre Yield and Quality’ (Eds. B.M. Potts, J.B. Reid, R.N. Cromer, W.N. Tibbits and C.A. Raymond). pp. 212-216. Proc. CRCTHF-IUFRO Conf., Hobart, 19-24 Feb. (CRC for Temperate Hardwood Forestry: Hobart).

Quaas R.L. and Pollak E.J. (1980) Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. 51:1277-1287.

Robinson, G.K. (1991) That BLUP is a good thing: the estimation of random effects. Statisitical Science 6:15-51.

SAS/STAT (1991) User’s Guide. Release 6.03 edition, SAS Institute Inc., Cary, NC.

Sorensen D., Kennedy B. (1984) Estimation of genetic variances from unselected and selected populations. J. Anim. Sci. 59:1213-1223.