OPTIMAL SELECTION INDICES IN EARLY STAGE TRIALS IN SUGARCANE BREEDING PROGRAMS

By

PHILLIP JACKSON[1], XIAN MING WEI[2], FELICITY ATKIN[3]
[1]CSIRO, Townsville, [2]Sugar Research Australia, Mackay, [3]Sugar Research Australia, Meringa
Phillip.Jackson@csiro.au

KEYWORDS: Selection, Selection Index, Economic Weightings, Fibre.

Abstract

SELECTION IN SUGARCANE breeding programs is complicated by measured traits being affected greatly by non-genetic effects, especially in early stage selection trials, and because clones vary for multiple traits affecting industry profitability which need to be selected for simultaneously. Selection index theory is a method which accounts for these issues, and has been progressively applied within the Australian sugarcane breeding program. In this paper, the selection index framework is briefly reviewed, and this is used to consider two issues in early stage selection trials. These are (i) adjusting index coefficients for presence of competition effects and genotype × environment interactions and (ii) incorporating traits which have no direct economic impact by themselves (e.g. DNA marker predictions or physiological traits) but which are correlated with other traits of high value (e.g. yield). Hypothetical but realistic examples are given to illustrate the methods. In addition, the economic weighting for fibre is considered. Economic weightings for fibre have recently been reviewed and the impact of different realistic weightings on selection in diverse populations typical of those in early stage trials is considered. Selection indices reflecting value of fibre based on some contrasting prices for electricity were derived and an example set of data presented in this paper. An overall conclusion from analysis of several data sets was that future progress in breeding programs for a range of realistic production systems and assumptions about product prices (e.g. for electricity or biofuel) is unlikely to be sensitive to the relative value of fibre, even at upper ends of likely potential value. Recommendations for dealing with uncertain future fibre values in breeding programs are made.

Introduction

The effectiveness of sugarcane breeding programs is fundamentally dependent on how good selection is. The key aim in each of three or more stages of selection is to select the best genotypes (clones) for progression to the next stage, up to a decision being made on whether a variety is good enough to be released to farmers for commercial production. Selection of elite clones is also critical for identifying parents for the next generation of seedlings to be produced from crossing.

Although the basic concept of selection may at first seem simple, it is complicated by two things. First, observations or measurements of clones for all traits are affected greatly by non-genetic effects. These non-genetic effects may be caused by factors such as heterogeneity in soil or water availability within field trials, random 'error' variation in how measurements are made, competition effects (where growth in seedlings or small plots are affected by growth of neighbouring plants), and genotype × environment interactions (where relative performance of genotypes changes in different environments). Different traits will be affected to different degrees by each of these components of variation, and these will also be affected by trial designs (e.g. number of replicates) and specific field and environmental conditions of the trial.

A second complication is that sugarcane clones vary for many traits which affect industry profitability (e.g. cane yield, sugar content, disease resistance, lodging propensity etc.), and generally no clones will be superior for every trait. Therefore, even if all traits could be measured precisely, ranking for overall economic value is still difficult.

Selection index theory provides a way to deal with these two complications and rank clones objectively based on best available knowledge. This method provides a single numeric value, called an optimal selection index, which maximises the correlation between the index and the true genetic worth of genotypes given available data and knowledge. This method was basically developed over 70 years ago (Smith, 1936; Hazel, 1943) but is arguably under-utilised in many sugarcane breeding programs.

The Australian sugarcane breeding program gradually began using the Hazel-Smith selection index approach from around 2006 through the use of best linear unbiased prediction to determine genetic values of clones, with economic weighting coefficients for each trait which depict relative economic impact to industry on changes to trait performance (Wei *et al*., 2008). The economic weightings were derived through analysis of trait impact on costs and revenue, in consultation with sugar industry stakeholders and experts. However there is ongoing development of selection index application, including ongoing updating of economic values of traits.

In this paper we provide an introduction to selection index theory and use this as a basis for considering further developments in existing methods as follows:

(i) Current methodology to determine selection indices in early stage trials may not account for some variation which may be confounded with estimates of genetic variance including genotype × environment interaction (because trials are usually at a single site) and competition effects (because single row plots are usually used).

(ii) Traits which don't directly have an economic impact, but correlate with economically important traits may be included in future selection indices. Examples include genomic prediction functions using DNA marker information and physiological measurements. A framework is needed to determine such indices and quantify predicted extra genetic gains so that costs of screening for these traits can be compared with the expected benefits.

(iii) Economic weightings for fibre have recently been reviewed and the impact of this trait on selection is considered.

Deriving an optimal selection index

When considering phenotypic values of traits, a selection index is constructed in the form of a linear equation of observed trait values for each genotype with a coefficient for each trait value:

$$SI_i = b_1*X_{1i} + b_2*X_{2i} + ... + b_n*X_{ni} \quad \text{(Equation 1)}$$

where SI_i = the selection index of genotype i ; b_1, b_2, …, b_n are the index coefficients to be estimated (below) for trait 1, trait 2, …, trait n; and X_{1i}, $X_{2i,}$ …, X_{2n} are the deviations of measurements trait 1, trait 2, …, trait n on genotype i from the mean of the population being selected from.

The main requirement in applying a selection index is therefore to determine the index coefficients (b values) which maximise the correlation between the true genetic worth of the individuals considered for selection and the index. It can be shown (e.g. White and Hodge, 1989) that the index coefficients can be determined from:

$$\mathbf{b} = [\mathbf{V}]^{-1}.\,[\mathbf{C}]\,.\mathbf{a} \quad \text{(Equation 2)}$$

where: **b** = n × 1 vector of index coefficients to be estimated (i.e. there are n traits measured), **V** = n × n matrix of the phenotypic variances and covariances among the measured traits, **C** = an n × q

matrix of covariances between the measured traits and genetic values of traits of economic value (i.e. there are q traits affecting economic value which are being considered in assessing economic worth of the genotype) and **a** = q × 1 vector of the economic weightings of traits affecting economic value.

Equation 1 may be written in a slightly different way as:

$$SI = \mathbf{b}'(\mathbf{y} - \boldsymbol{\alpha}) \text{ (Equation 3)},$$

where SI is the selection index of a genotype, **b'** is the transpose of the n × 1 b vector (above) y is an n × 1 vector of phenotype measurements for the genotype, and **α** is an n × 1 vector of the means of the phenotype measurements for the population being considered for selection and against which the genotype is being assessed. Based on equation 2, and considering that $\mathbf{V}^{-1}$ is symmetric, equation 3 may also be written as:

$$SI = \mathbf{a'\ C'V^{-1}(y - \alpha)} \text{ (Equation 4)}$$

The value $\mathbf{C'V}^{-1}(y - \boldsymbol{\alpha})$ is equivalent to the q × 1 vector of *best linear predictors* (BLPs) for the genetic traits of economic value of the genotype. Where ordinary least squares estimates of **α** are not satisfactory, such as if there is unbalance in the selection trial, the well known statistic *best linear unbiased predictor* (BLUP) is more appropriate than the BLP. This substitutes **α** with an alternative and more appropriate value. Statistical software may be readily used to directly determine BLUPs from selection trial data being used to rank genotypes. In this way, the worth or index of a genotype may be determined from:

$$w = \mathbf{a'\ g} \text{ (Equation 5)},$$

where **g** is the vector of BLUPs for each trait of economic value being predicted for the genotype.

Current procedures employed in the Australian sugarcane breeding programs mainly use equation 5 with BLUPs derived from analysis of individual traits from trial data. However, this function does not deal with potentially important sources of variance which may be confounded with apparent genetic variance (e.g. some competition variance and GE variance), and does not include traits which have no economic impact but which may correlate with economically important traits. Also, if traits are correlated and independent analysis is done on each trait the index will not take account of these covariances.

Example of an optimal selection index

The following hypothetical example uses realistic assumptions based on observed genetic and statistical parameters in selection trials in the Burdekin region, and economic weightings of traits derived from past studies (Wei *et al.*, 2006; Jackson *et al.*, 2014). It is presented to illustrate an approach which addresses the potential limitations to past approaches indicated above, and also how an additional trait, stomatal conductance (g_s), may add value in selection despite having no direct economic value.

The linear model assumed for each trait was:

$$y_{ijk} = \mu + g_i + c_i + ge_{ij} + error_{ijk}$$

where

y_{ijk} = phenotypic value of genotype i in environment j and replicate k (i.e. the observed performance of the genotype, and upon which selection is based);

μ = mean of all observations;

g_i = true genetic effect of genotype i (i.e. the effect we would really like to estimate and use for selection but can never know precisely because of the other effects);

c_i = competition effect of genotype i (i.e. deviation of the performance of genotype i of expected value in small plots from its performance in pure stand; pure stand means if it is grown in a commercial field by itself);

ge_{ij} = genotype × environment interaction of genotype i in environment j;

$error_{ijk}$ = random experimental error effect of genotype i in environment j in replicate k.

Some statistical models may partition a proportion of competition effects from phenotypic variation depicted in the above model (e.g. Stringer and Cullis, 2002). This proportion is unclear at this stage, but for the purpose of this example and simplicity it is assumed that all competition effects remain as part of phenotypic variance.

Based on this model,

$$\sigma_p^2 = \sigma_g^2 + \sigma_{comp}^2 + \sigma_{gxe}^2/(n_e) + \sigma_{error}^2/(n_e.n_r)$$

where

σ_p^2 = phenotypic variance of the estimates of means of genotypes;

σ_g^2 = (true) genetic variance of genotypes under test;

σ_{comp}^2 = variance due to competition effects;

σ_{gxe}^2 = variance due to g × e interaction;

σ_{error}^2 = variance due to experimental error variance;

n_e = number of environments used for testing;

n_r = number of replicates per environment.

Based on results from past analyses of variance of a set of selection trials, assumptions for these components were made (Table 1), and it was assumed there is only one environment ($n_e = 1$) used for selection and two replicates per clone ($n_r = 2$).

Table 1—Assumptions for statistical parameters for three traits for the example selection index derived in the text. Phenotypic variances assume one environment and two replicates.

Statistic	Trait		
	CCS (%)	Cane yield (t/ha)	Stomatal conductance (mol m^{-2} sec^{-1})
Genetic variance (σ_g^2)	1.0	350	500
Competition var. (σ_{comp}^2)	0	1050	0
GE variance (σ_{gxe}^2)/n_e	0.4	52	0
Error variance (σ_{error}^2)/$n_e n_r$	1.0	900	400
Phenotypic variance (σ_p^2)	1.9	1902	700

It was assumed that the genetic correlation (r_g) between CCS and cane yield was equal to zero (in most trials this varies from slightly negative to zero), a genetic correlation of stomatal conductance (g_s, measured under optimal conditions) and cane yield of 0.8 (based on data obtained by Basnayake *et al.*, personal communication), and error correlations between all traits equals zero. Therefore the phenotypic covariance between CCS and cane yield (t/ha, TCH) was zero. For conductance (g_s) and TCH:

$$\text{since } r_{g(g_{s.TCH})} = cov_g/(\sigma_{g,gs} . \sigma_{g.TCH}) = 0.8$$

$$\text{then, } cov_{g(g_{s.TCH})} = 0.8 * (500*350)^{0.5} = 334$$

Based on the above assumptions, the following values were derived for the 3x3 **V** matrix (i.e. the variances and covariances between the phenotypes of the measurements in the single row plots):

	CCS	TCH	g_s
CCS	1.9	0	0
TCH	0	1902	334
g_s	0	334	700

For considering covariances between the phenotype measurements made in a single row plot and true genetic values in multi-row plots, the only effect in common is the true genetic variance, and hence this is the covariance. Therefore the 3 × 2 **C** matrix is as follows:

	CCS (pure stand)	TCH (pure stand)
CCS (single row)	1.0	0
TCH (single row)	0	350
g_s (single row)	0	334

Based on prior work, it was assumed that a 1% unit increase in CCS (e.g. going from 12% to 13%) results in a total industry benefit of $400/ha, and 1 tonne per hectare increase in cane yield (in a pure stand) results in a total industry benefit of $35/ha.

Therefore the (2 × 1) **a** vector is:

CCS (pure stand)	400
TCH (pure stand)	35

Given equation 2,

$$\mathbf{b} = [V]^{-1}.\,[C]\,.\mathbf{a}$$

then

$$b = -1\,.\,.$$
$$=$$

Therefore, the optimal selection index = 210.52*CCS + 3.82*TCH + 14.8*g_s. Using the same procedure but eliminating gs from consideration, the optimal selection index = 210.52*CCS + 6.44*TCH.

It may be shown (White and Hodge, 1989) that the expected correlation between the index prediction and the true genetic worth of genotypes (Correl (w, I)) is given by:

$$\text{Correl (w, I)} = (\mathbf{a' C V^{-1} C a / a' G a})^{0.5}$$

where **G** is the q × q matrix of genetic variances and covariances among the traits of economic value, and others are defined above. Substituting the above assumptions into this equation, the expected correlation between the genetic value and the index which includes stomatal conductance is equal to 0.71, and the correlation for the index which does not include stomatal conductance is equal to 0.52. Expected genetic gain from selection is proportional to this correlation (Falconer and Mackay, 1996) and therefore inclusion of stomatal conductance in the index would be expected to result in a 38% increased gain based on the assumptions above.

Adjusting BLUP coefficients for economic traits to deal with indirect traits

The derivation of BLUPs for individual traits being used for selection offers advantages such as availability of software and statistical models to deal with extraneous effects such as spatial variation in trials and unbalanced data. Selection indices may be derived by applying economic weightings to these traits (as in equation 5) and this is currently the way selection indices are generally derived in the Australia sugarcane breeding program. However this approach alone may not take account of some variance and covariance components such as competition and GE interaction, covariances between traits, and traits of no direct economic impact but which may be of value for selection because of correlated effects.

One approach to address these issues while still using BLUPs for fitting genetic values to each trait is through the following formula derived by Schneeberger *et al*. (1992):

$$\mathbf{b_{BLUP}} = \mathbf{G_{11}.G_{12}.a} \text{ (Equation 6)}$$

where $\mathbf{b_{BLUP}}$ is the vector of index weights for the BLUPs for each of the selection criteria, $\mathbf{G_{11}}$ is the genetic-covariance matrix of the selection criteria in the index, $\mathbf{G_{12}}$ is the genetic covariance matrix between the selection criteria in the index and the traits of economic value in the breeding objective, and **a** (defined as **v** in Schneeberger *et al.*, 1992) is defined as above.

The framework in equation 6 may be used for single row plots and trials at a single site, by inferring the traits measured are separate but correlated with the economically important traits of cane yield and CCS in a pure stand across multiple environments. This would need to assume variance components for both competition variance and GE variance. These assumptions may be based on past experience and research.

Effect of fibre value on selection index and selection outcomes

In the past, fibre content in cane has been weighted negatively in selection within breeding programs because of the adverse impact of high fibre on sugar extraction and/or milling rates. However, in some regions fibre may have positive value because of its contribution to electricity cogeneration capacity and revenue.

Recent research considered this question and determined economic weightings for production systems deriving significant revenue from cogeneration. The actual weightings (in units of $/ha per change in unit for the trait) will be dependent on region and mill specific factors. However, based on a realistic set of assumptions (Jackson *et al*., 2014) the following set of economic weightings were determined and shown in Table 2.

Table 2—Economic weightings for CCS, cane yield and fibre, based on realistic assumptions for the Australian sugarcane industry, for two prices received for electricity.

Scenario	Weighting for CCS ($/ha/% unit)	Weighting for cane yield ($/ha/t/ha)	Weighting for fibre ($/ha/% unit)
No cogeneration	427	34	–30
Cogeneration @ $100/MW-hr	440	48	62
Cogeneration @ $200/MW-hr	434	58	151

A set of stage 2 (clonal assessment trial) data from Meringa was examined with these alternative selection indices. In this case estimates of genetic and error variance were obtained from analysis of the trial, with the additional assumption that competition effects comprised two-thirds of the apparent genetic variance (in the single row plots) for cane yield, and GE variance was 50% of the size of genetic variance. These assumptions are reasonably robust and based on past research, and it is also noted that results are not highly sensitive to realistic variations in these assumptions (data not shown). The three selection indices derived were as follows: (i) for $0/MW-hr: 5.42*TCH + 237*CCS + –7.1*fibre, (ii) for $100/MW-hr: 7.65*TCH + 245*CCS + 14.8*fibre, (iii) for $200/MW-hr: 9.25*TCH + 241*CCS + 36.1*fibre.

A generally high correlation (>0.85) overall was observed for genotype value ranked on any of the indices (Figure 1). It is also noted that selection indices representing the same contrasting economic weightings for fibre have also been examined in final assessment trials (not shown here) with similar general results. This indicates that selection based on any of these indices within a broad population of genotypes will result in large selection gains for the other indices. Selection for superior genotypes according to one index results in a set of genotypes with much overlap with genotypes selected on the basis of other indices (i.e. many genotypes in common in top ranked

genotypes selected according to different indices). As expected the correlations were slightly less for those involving the most contrasting prices of electricity ($0/MWh versus $200/MWh), but the correlations were still generally high. Correlations were less (0.7–0.8) if only the top genotypes were considered, but these are considered still relatively high.

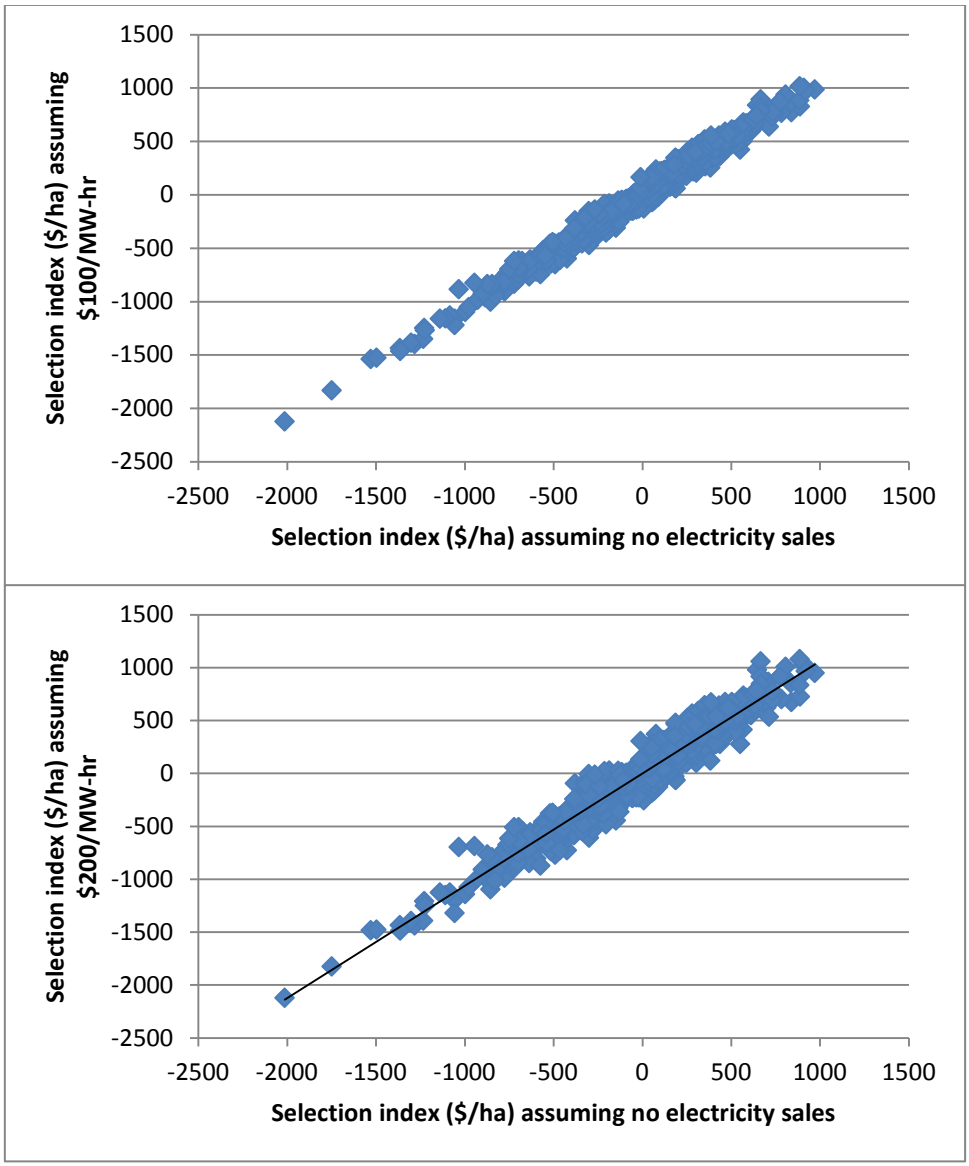

Fig. 1—Relationship between clone values in an early stage trial for selection indices assuming different values of electricity produced from cogeneration, as indicated.

Conclusions and recommendations

Selection index theory can be used to address complications which arise in selection because of the need to deal with multiple traits of differing measurement accuracy and economic value. Selection indices are used in the Australian sugarcane breeding programs through deriving BLUPs for important traits independently and multiplying these by economic weightings.

Two extensions of this approach involve adjusting index coefficients for presence of competition effects and genotype × environment interactions and incorporating traits which have no direct economic impact by themselves but which are correlated with other traits of high value (e.g. yield).

An example is given for a physiological trait (stomatal conductance), but the same principles and procedures would apply for a DNA marker (or group of markers) where the presence or

absence (or degree of expression) has a significant covariance with economic traits. Incorporation of such traits may greatly improve selection gains if there are high genetic correlations with traits of economic value.

An overall conclusion in relation to the impact of contrasting values of fibre was that future progress in breeding programs for a range of realistic production systems and assumptions about product prices (e.g. for electricity or biofuel) is unlikely to be sensitive to the relative value of fibre, even at likely upper ends of potential value (e.g. $200/MWh electricity). This suggests that, even based on selection criteria assuming a low value of fibre as is the case currently, breeding programs will be at least reasonably on track to develop cultivars that are suitable for alternative production systems that may arise in the future.

It is recommended that the mainstream (core) Australian sugarcane breeding program use selection indices based on existing estimated value of fibre within individual regions (which may range from slightly negative to slightly positive based on the processing technologies available in that region). However, it is also suggested that clones undergoing selection in the breeding program should be assessed using a selection index which assumes a very high value for fibre (e.g. $200/MWh electricity). If any clones are identified which offer significantly higher value based on this index than clones selected using the mainstream program index, then a small to moderate number of the best should be retained through ongoing propagation in small plots.

This recommendation is based on the considerable uncertainty associated with predicting future fibre and energy (electricity or biofuel) values. Clones retained in propagation plots, at very low cost, based on selection indices assuming high fibre value would be immediately available for industry deployment, rather than needing to be identified through lengthy selection systems.

Acknowledgements

Data sets were obtained from project BSS250 (funded by SRDC) and from the Australian sugarcane breeding program database (SPIDNET) owned and maintained by Sugar Research Australia. Economic weightings for fibre were obtained from work done in SRDC/SRA funded project CPI017. We acknowledge the large contributions from James Joyce in assisting with the economic modelling in this project required to derive the economic weightings.

REFERENCES

Falconer DS, Mackay TFC (1996) ‘Introduction to quantitative genetics’. 4th edn (Longman Edinburgh).

Hazel LN (1943) The genetic basis for constructing selection indexes. *Genetics* **28**, 476–490.

Jackson P, Wei X, Hu F (2014) Final Report Project CPI017: Development of sugarcane for production systems utilising total biomass. Sugar Research Australia, Brisbane, Australia.

Smith H (1936) A discriminant function for plant selection. *Annals of Eugenics* **7**, 240–250.

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.

Wei X, Jackson P, Cox M, Stringer J (2006) Maximising economic benefits to the whole sugar industry from the BSES-CSIRO sugarcane improvement program. *Proceedings of the Australian Society of Sugar Cane Technologists* **28**, 181–186.

Wei X, Jackson P, Stringer J, Cox M (2008) Relative economic genetic value (rEGV)—an improved selection index to replace net merit grade (NMG) in the Australian sugarcane variety improvement program. *Proceedings of the Australian Society of Sugar Cane Technologists* **30**, 136–141.

White TL, Hodge GR (1989) ‘Predicting breeding values with applications in forest tree improvement’. (Kluwer academic publishers: Dordrecht, Boston, London).