COVARIATE SELECTION USING PENALISED REGRESSION ANALYSIS IN MILL MUD DATA ANALYSIS

Sugar Research
Australia

M.E. Olayemi and J.K. Stringer
Sugar Research Australia, PO Box 86, 50 Meiers Road, Indooroopilly, QLD 4068, Australia

ABSTRACT

Cane yield data from an experiment designed to investigate the application of mill mud-ash treatments were analysed together with soils apparent electrical conductivity (ECa) measured at four soil depths. The apparent soil electrical conductivity measurements are highly dependent on one another and were used as covariates in the analysis. The variance inflation factor (VIF) indicates that there is multi-collinearity amongst these covariates. A penalised regression approach using lasso with SBC as the tuning method was used to select soil ECa covariates that are most relevant to cane yield. A linear mixed model was then fitted to the data with treatment and selected covariates from the lasso regression as fixed effects and replicate as the random effect. There was significant difference in treatment effects for cane yield (t/ha).

INTRODUCTION

Mill mud is the material remaining after cane juice is clarified and filtered in combination with other beneficial by-products like ash from sugarcane mills. It is a valuable soil conditioner and an important source of plant nutrients. Sub-surface application of mill mud is being trialled to improve productivity in the Herbert region of Queensland, Australia. For precise and variable application rate of mill mud to farm plots soil apparent electrical conductivity (ECa) mapping is needed. Soil ECa mapping allows farm block fertility management zones to be defined and create variable nutrient application rates. Due to changes in technology, soil ECa measurements are moving from a two-level soil depth to four-level soil depth measurements. These measurements are highly correlated. Fitting all the covariates in the model could lead to biased coefficient estimation, high standard errors and may result in incorrect conclusions about the relationship between outcome and predictor variables. Therefore, an appropriate methodology to account for collinearity while fitting these covariates in the analytical model should be used.

OBJECTIVE

- Aim is to assess the ECa covariates for collinearity, explore penalised regression (LASSO) for mitigating the collinearity and for feature selection.
- Compare mill mud treatment means for cane yield tonnes/ha to determine the best treatment.

METHOD

- Data from a sugarcane trial from Herbert region, Queensland, Australia.
- Seven treatments were evaluated:
  - *Control = No mill mud/ash applied*
  - *ABa50 = Ash Banded 50 t/ha*
  - *ABa100 = Ash Banded 100 t/ha*
  - *ABr200 = Ash Broadcast 200 t/ha*
  - *MBa50 = Mill Mud Banded 50 t/ha*
  - *MBa100 = Mill Mud Banded 100 t/ha*
  - *MBr200 = Mill Mud Broadcast 200 t/ha*
- Covariates: apparent ECa measured at the soil depths of 0.5, 1.0, 1.6 and 3.2m.
- Covariates were accessed for collinearity by their variance inflation factor (VIF).
  - *If > 10 there is collinearity*
- Penalised regression approach using LASSO (Tibshirani, 1996), with SBC as the tuning method was applied to the data for variable selection, using PROC GLMSELECT of SAS Analytic software (SAS Institute, 2013).

  $\beta\hat{}^{LASSO}(\lambda) = \text{arg imin\_}_{\beta}(||Y - X\beta||^2 + \lambda||\beta||/_1)$

  where λ is a tuning parameter and $||.||\ /_1$ stands for the vector $/_1$-norm.
- This approach placed a constraint on the size of the regression coefficients in the model by shrinking them toward zero.
- Variables with a regression coefficient equal to zero after the shrinkage process were excluded from the model.
- Covariates with a non-zero coefficients were the most strongly associated with the response variable and were kept in the model.
- The selected variables were then used to fit a mixed model to access the effect of treatment on cane yield.
- Treatments were considered as fixed, replicate as random effects and the selected soil EC as covariates.

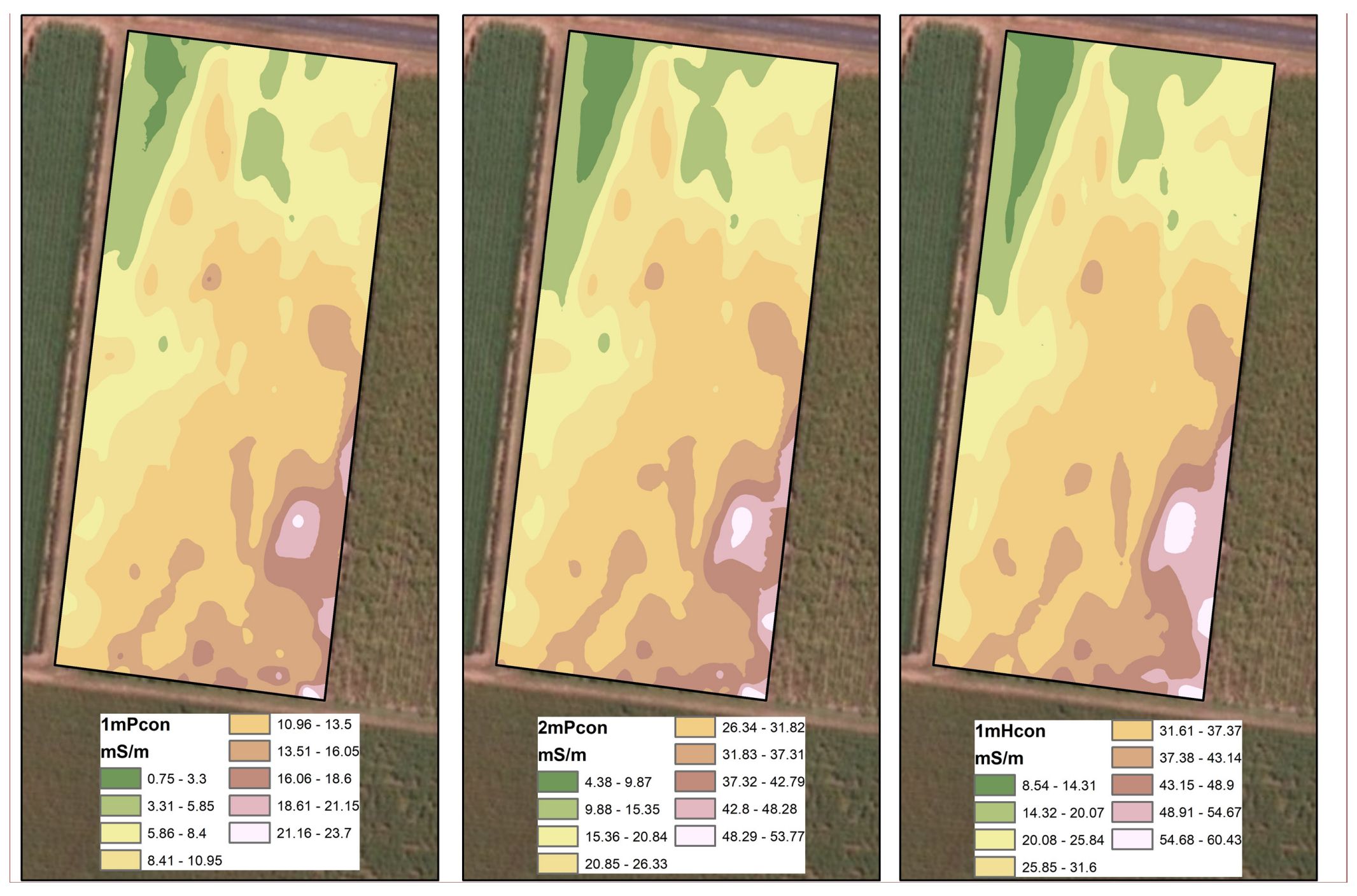

Fig1. *Example of soil apparent electricity conductivity (ECa) measurement at soil depth 0.6m (1mPcon), 1.2m (2mPcon) and 1.5m (1mHcon)*

RESULTS

- Covariates were highly correlated with r between 0.96 to 0.99.
- VIF for all the variables were > 10 (412 to 2029)
- Soil ECa measurement at soil depths of 0.5 and 1.0 m were the most strongly associated covariates with cane sugar yield

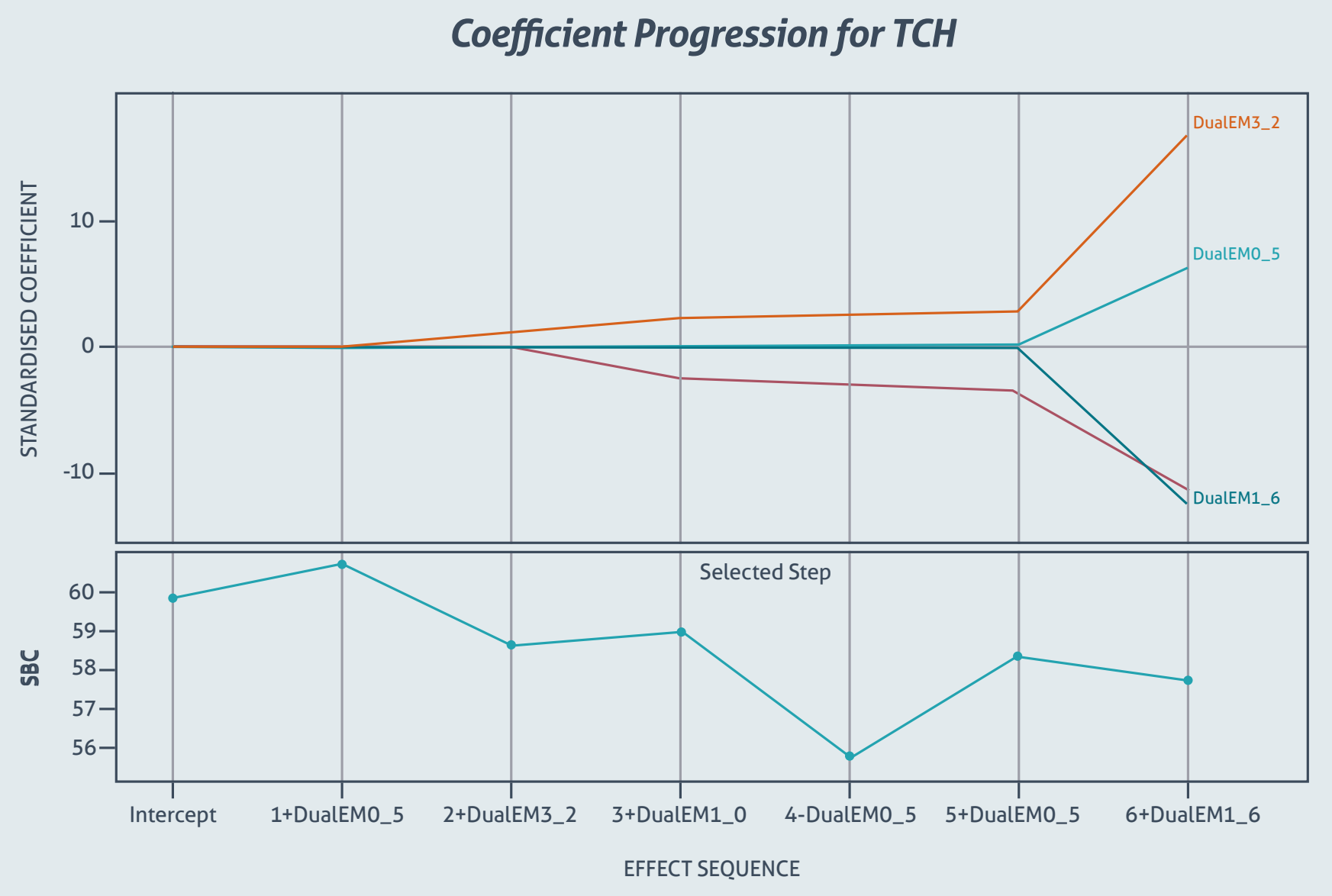

Fig2a. *Coefficient progression and model selection step, each curve represents a coefficient as labelled*

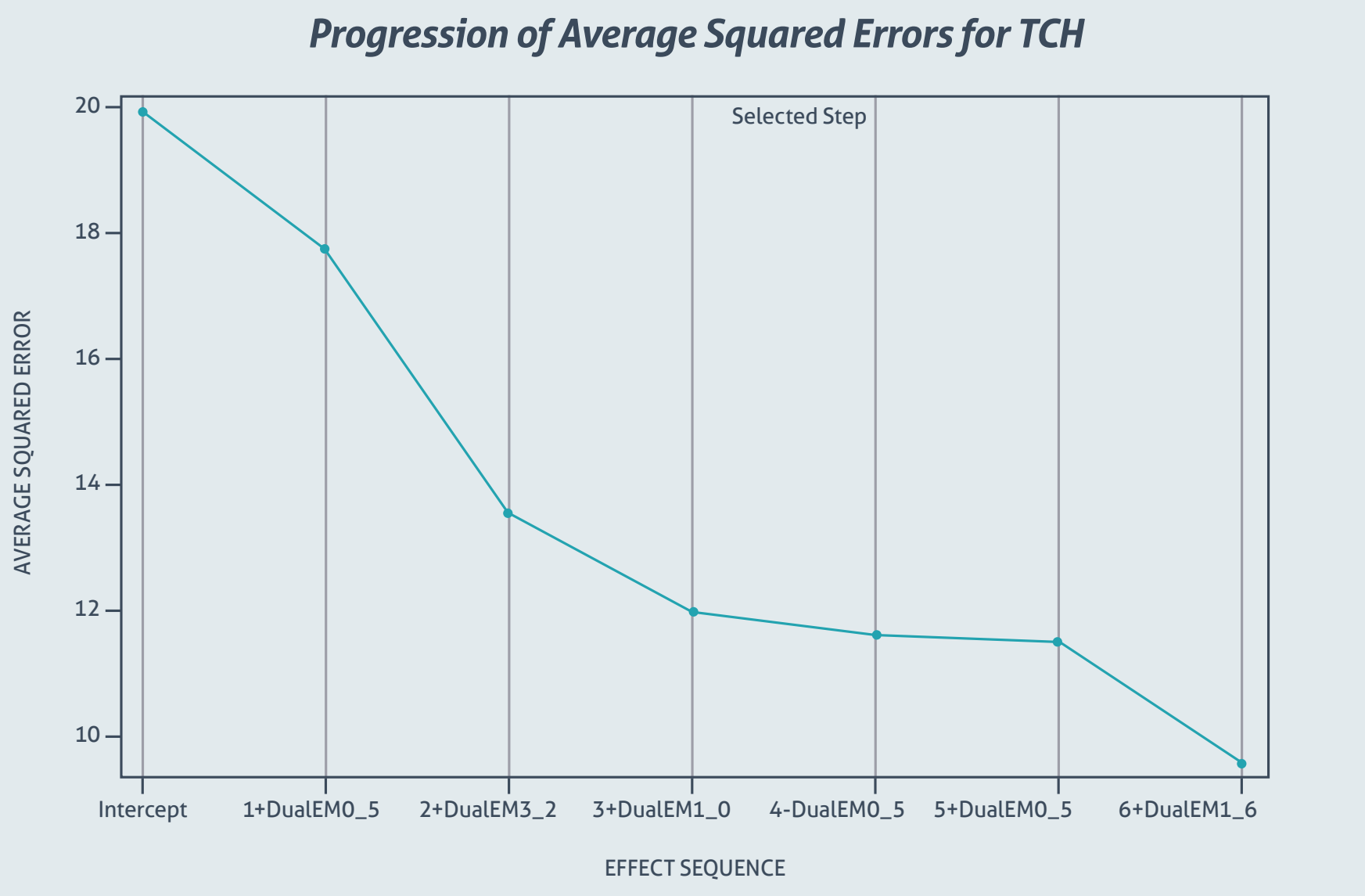

Fig2b. *Progression of average square errors for cane yield showing selected covariates*

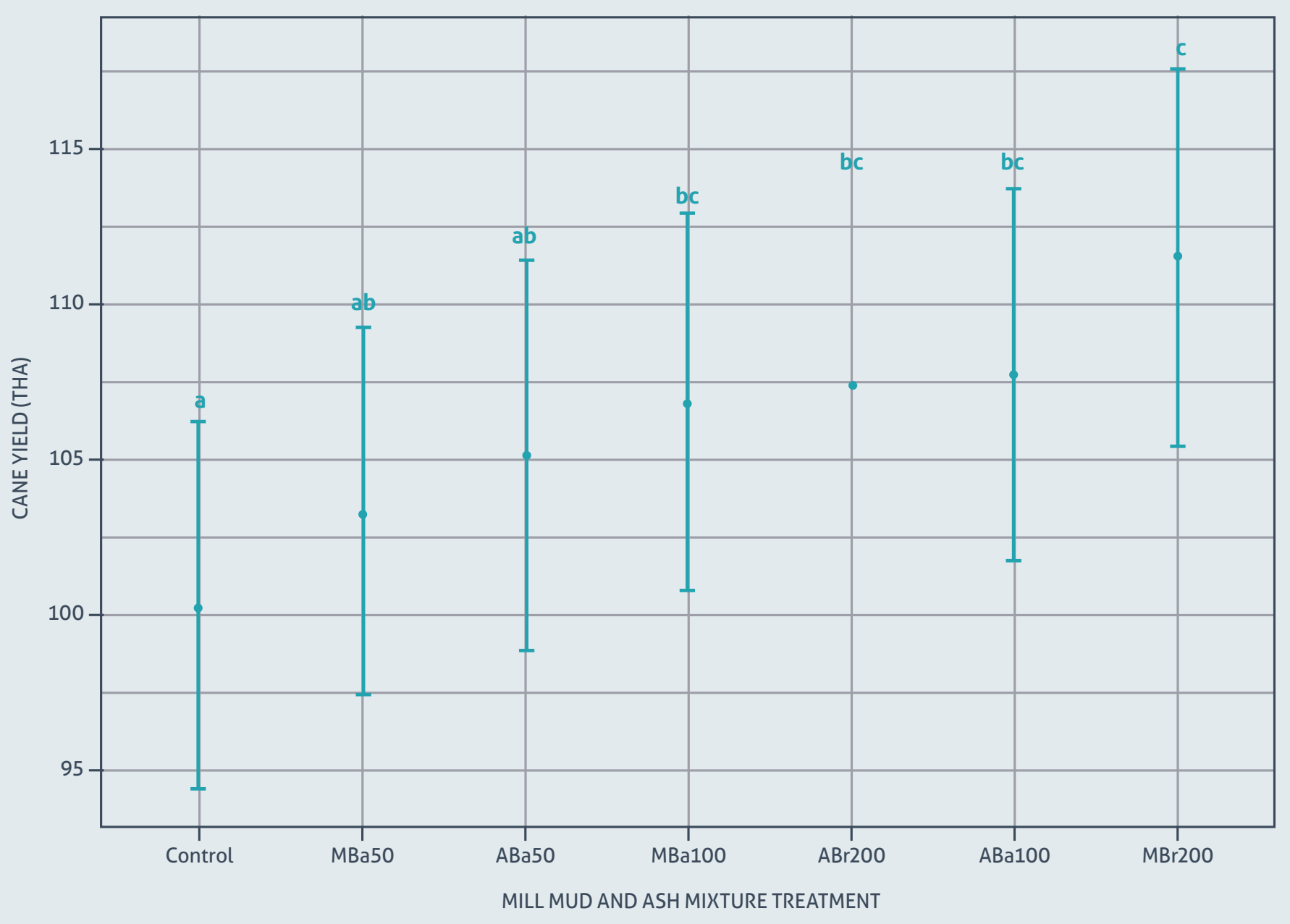

Figure 3. *Predicted cane yield t/ha, error bars are confidence interval at 95% level*

CONCLUSION

We were able to use LASSO regression to select the most relevant Eac covariates for our response variable, cane yield. The ash/mill mud application had significant impact on cane yield, with greater yield (tonnes/ha) from the mil mud broadcast at 200t/ha (MBr2000) than the banded application of either mill mud or ash at 50t/ha.

ACKNOWLEDGMENT

We wish to thank Peter Larsen and Carla Atkinson from Wilmar Sugar Australia for making the data available.

REFERENCES

SAS Institute. (2013). The SAS System for Windows. *Release 9.4. Copyright © 2013, SAS Institute Inc.,* (Cary, NC., USA.).

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288

Disclaimer: In this disclaimer a reference to 'we', 'us' or 'our' means SRA and our directors, officers, agents and employees. Although we do our best to present information that is correct and accurate, we make no warranties, guarantees or representations about the suitability, reliability, currency or accuracy of the information we present in this information sheet, for any purposes. Subject to any terms implied by law and which cannot be excluded, we accept no responsibility for any loss, damage, cost or expense incurred by you as a result of the use of, or reliance on, any materials and information appearing in this information sheet. You, the user, accept sole responsibility and risk associated with the use and results of the information appearing in this information sheet, and you agree that we will not be liable for any loss or damage whatsoever (including through negligence) arising out of, or in connection with the use of this information sheet. We recommend that you contact our staff before acting on any information provided in this information sheet. **Warning:** Our tests, inspections and recommendations should not be relied on without further, independent inquiries. They may not be accurate, complete or applicable for your particular needs for many reasons, including (for example) SRA being unaware of other matters relevant to individual crops, the analysis of unrepresentative samples or the influence of environmental, managerial or other factors on production.

sugarresearch.com.au