Predictive models incorporating environmental covariates for genotype × environment × management (G×E×M) interactions applied to sorghum agronomy trials

Michael Mumford[1], Clayton Forknall[1], Daniel Rodriguez[2], Joseph Eyre[2], Loretta Serafin[3], Darren Aisthorpe[4], Kerry Bell[1], Alison Kelly[1,5]

[1] Department of Agriculture and Fisheries, Leslie Research Facility, Toowoomba, QLD 4350, Email: Michael.Mumford@daf.qld.gov.au

[2] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Gatton Campus, QLD 4343

[3] New South Wales Department of Primary Industries, Tamworth, NSW, 2340

[4] Department of Agriculture and Fisheries, Emerald, QLD, 4720

[5] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Hermitage Research Facility, Warwick, QLD 4370

Abstract

Researching the management (M) of genotypes (G) in agronomic experimentation is essential to help farmers maximise grain yield, though the approach is complicated by interactions emerging from changing environmental (E) factors across sites and seasons. Available statistical methods for modelling the G×E interaction are limited as they do not provide a functional understanding of how environmental factors influence the G×E interaction, nor assess how different management practices (M) influence the G×E interaction.

A predictive linear mixed model is proposed that incorporates site/season-specific environmental covariates into a standard G×E interaction framework. The model is extended to include continuously varying agronomic management practices whilst allowing for non-linear trait responses and complex variance structures. The methodology was applied to a multi-environment data set associated with GRDC's optimising sorghum agronomy program. The analysis identified key environmental drivers and management strategies that explained the G×E×M interaction, enhancing the biological understanding of the analysis results and allowing for the development of more robust recommendations for agronomic practices.

Keywords

cross validation, plant population, residual maximum likelihood

Introduction

It is well established that genotype (G) performance is strongly influenced by the environment (E), resulting in the common practice of assessing genotype adaptation in trials across different seasons and geographic locations (Cooper and Hammer 1996). Data collected from such a set of trials is referred to as a multi-environment trial (MET) data set and facilitates the modelling of the genotype by environment (G×E) interaction. Statistical methods for the estimation and exploration of the G×E interaction are well documented (Smith et al. 2001), however these approaches are limited due to their inability to provide a functional understanding of which environmental factors are driving the interaction pattern.

Predictive models that provide a better understanding of the environmental drivers influencing the G×E interaction can be built through the inclusion of environmental covariate (ECs) in the modelling framework (Heslot et al. 2014). These approaches also facilitate the prediction of genotype performance in a 'new', untested environments (Malosetti et al. 2016).

However, the development of statistical methods to model the G×E interaction using ECs has mostly occurred within the context of plant improvement (Crossa et al. 2010). When working in an agronomic setting, measuring the impact of management practice (M) on genotype performance is also a key objective (Rodriguez et al. 2018). This induces additional potential sources of interaction; genotype by management, management by environment and genotype by environment by management (G×E×M).

The objective of this study is to present a statistical method for modelling the G×E×M interaction with an extension to enable the incorporation of ECs into this framework. This will enable the identification of the key environmental drivers influencing the effectiveness of management practices on genotype performance.

Moreover, it facilitates the prediction of genotype performance for untested combinations of management practice and environment.

Methods

Motivating data

The motivating MET data set consists of six sorghum agronomy trial sites, with the locations ranging from northern New South Wales to central Queensland. All sites contained a consistent set of three treatments in factorial combination; time of sowing (TOS), genotype (G) and target plant population (M). All sites had a total of three replicate blocks and were planted in the 2018/19 sorghum growing season, with time of sowing treatments spanning the winter, spring and summer growing periods (Table 1). For the purposes of this study, 'environment' is defined as the combination of site and TOS, resulting in a total of 17 environments.

Table 1. Summary of the treatment structure and experimental design for each of the sorghum agronomy sites.

Site	TOS 1	TOS 2	TOS 3	# Hybrids	Target plant population (plants/m^2)	Experimental Design
Breeza (Dryland)	6/09/2018	17/09/2018	23/10/2018	6	3,6,9 & 12	Split-split plot
Breeza (Irrigated)	3/09/2018	18/09/2018	16/10/2018	8	3,6,9 & 12	Split-split plot
Ponjola	8/08/2018	12/09/2018	27/09/2018	8	3,6,9 & 12	Split-split plot
Warra	27/07/2018	19/10/2018	9/11/2018	9	3,6,9 & 12	Split-plot
Surat	8/08/2018	28/08/2018	24/01/2019	9	3,6,9 & 12	Split-plot
Emerald	26/07/2018	16/08/2018		8	3,6,9 & 12	Split-split plot

Environmental covariates

Weather data for each site was obtained from local weather stations and consisted of information on rainfall, radiation, air temperature, soil temperature and evapotranspiration. Measurements related to the phenology of each genotype, being days to i) emergence ii) flowering and iii) maturity, were also recorded. By combining the phenology measurements and weather data, a total of 19 ECs were derived for each environment and 17 of these ECs varied for each genotype in each environment. The 2 remaining ECs did not vary between genotypes or sowing times within a site.

Statistical methods

The analysed variable was hand harvested total grain yield (t/ha) which was adjusted to 0% moisture for all sites. Established plant population was included in the model as a continuous variable, enabling prediction of grain yield across an observed range of established plant populations. All models were implemented in a linear mixed model framework via residual maximum likelihood, allowing for complex variances structures to be captured in the model including, but not limited to, heterogeneity of residual variance, experimental design terms and spatial field trend.

The ECs were introduced into the model as fixed effect covariates. Interactions between an EC and genotype or established plant population were also fitted as fixed effects. Random regression (Laird & Ware 1982) was employed to capture the genotype by established plant population interaction. A non-linear yield response to i) an EC and ii) established plant population was captured via natural cubic smoothing splines (Verbyla et al. 1999). Furthermore, the non-linear interaction effect between an EC and established plant population was captured through the use of tensor cubic-smoothing splines (Verbyla et al. 2018). Due to the small number of environments with respect to the 'population' of environments, the interaction between pair-wise combinations of ECs was not considered, thus the inclusion of additional ECs into the model is additive.

A forward selection procedure for the identification of ECs was completed such that at each iteration, the model was run 19 times (1 for each EC considered from the motivating data set). Once an important EC was identified, a backwards selection procedure testing for non-significant EC terms was performed to achieve a parsimonious model. The forward-backward selection procedure was then repeated until ECs were no longer significant.

Once the final model for the set of ECs was determined, a leave-one-out cross validation scheme was performed where the final model was re-run without one of the six sites and then predictions were obtained for the ‘left out’ site. The process was repeated for all six sites, to obtain a set of predictions in a ‘quasi-untested’ environment to assess for overfitting in the final model.

Results

From the analysis of the sorghum agronomy MET data set, three significant ECs were identified; i) post-flowering rainfall (mm) ii) pre-flowering rainfall (mm) and iii) post-flowering minimum temperature (°C). For post-flowering rainfall there was a significant linear interaction effect with genotype but no significant interaction with established plant population (Figure 1).

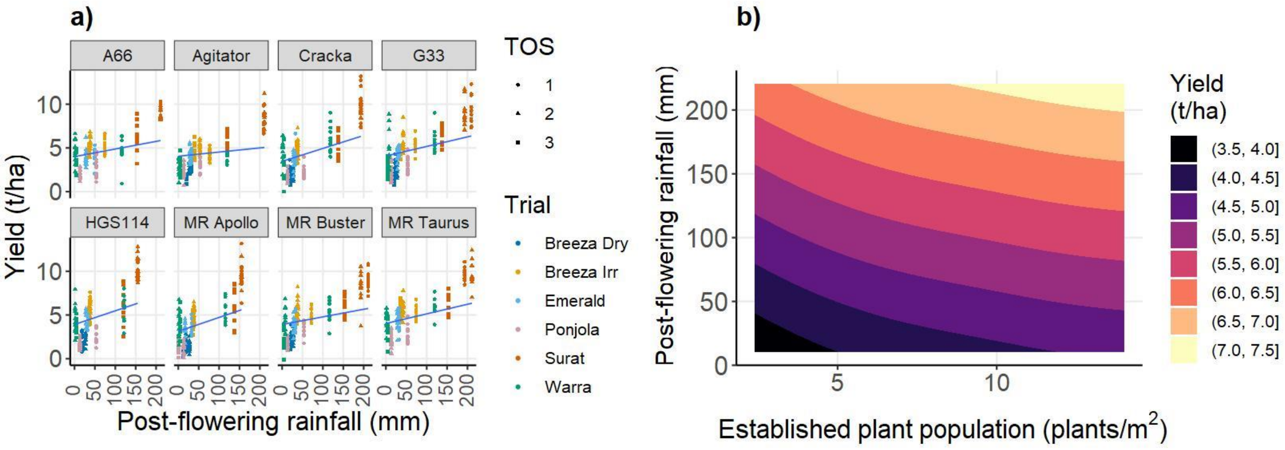

Figure 1) Yield predictions for the a) genotype × post-flowering rainfall interaction (significant) interaction and b) established plant population × post-flowering rainfall interaction (non-significant).

Additionally, pre-flowering rainfall had a significant linear interaction with established plant population, but no significant interaction with genotype. Finally, there was a significant non-linear interaction effect between post-flowering minimum temperature and established plant population. There were no significant three-way interaction effects for any of the three ECs.

Yield predictions from the final model that incorporates the three ECs as a surrogate for the environment term are displayed in Figure 2 for a subset of two genotypes and four environments. The predicted responses demonstrate that the three ECs do a reasonable job of explaining the differences between the high and low yielding environments (Figure 2a). Moreover, the model retains most of its predictive performance after leave-one-out cross validation (Figure 2b) providing evidence that the model is not overfitting and demonstrating how the proposed method can be used to predict genotype performance in an untested environment, under differential management practices.

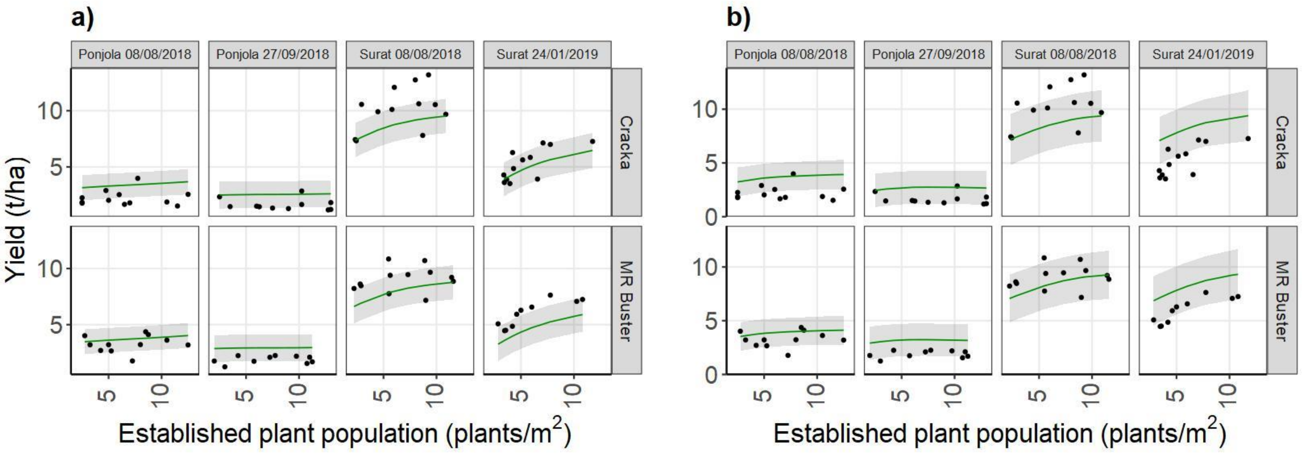

Figure 2) Predictions (coloured line) of the yield response to plant population for a subset of two genotypes and four environments in both a) a tested and b) an untested environment via 6-fold leave-one- -out cross validation. The shaded regions denote the 95% prediction intervals. Raw yield (black points) were adjusted for experimental design and field trend effects in the baseline model.

Conclusion

A new methodology has been presented that incorporates ECs into the analysis of G×E×M data, accounting for complex variance structures and non-linear trait responses to ECs. Results from the model indicated that grain yield performance of sorghum genotypes would be optimised in environments that have i) high pre-flowering rainfall, ii) high post-flowering rainfall and which saw genotypes flowering under iii) an optimal post-flowering temperature. Under this set of optimal G×E conditions, a high established plant population further optimised grain yield. This study is the first step towards developing one-stage statistical models that can identify key environmental drivers of G×E×M interactions and, by providing the ability to predict the performance of genotypes in an untested environment under differential management practices, can be utilised by crop modellers to refine model predictions.

Acknowledgements: The authors would like to acknowledge the Grains Research and Development Corporation for their financial support and investment into projects UOQ 1808-001RTX and DAQ00208.

References

Cooper, M. and Hammer, G.L. (1996). Plant adaptation and crop improvement. IRRI.

Crossa, J., Vargas, M. and Joshi, A.K. (2010). Linear, bilinear, and linear-bilinear fixed and mixed models for analyzing genotype × environment interaction in plant breeding and agronomy. Canadian Journal of Plant Science, 90(5), pp.561-574. (https://doi.org/10.4141/CJPS10003).

Heslot, N., Akdemir, D., Sorrells, M.E. and Jannink, J.L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics, 127(2), pp.463-480. (doi:10.1007/s00122-013-2231-5).

Laird, N.M. and Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics, pp.963-974. (https://doi.org/10.2307/2529876).

Malosetti, M., Bustos-Korts, D., Boer, M.P. and van Eeuwijk, F.A. (2016). Predicting responses in multiple environments: issues in relation to genotype × environment interactions. Crop Science, 56(5), pp.2210-2222. (https://doi.org/10.2135/cropsci2015.05.0311).

Rodriguez, D., De Voil, P., Hudson, D., Brown, J.N., Hayman, P., Marrou, H. and Meinke, H. (2018). Predicting optimum crop designs using crop models and seasonal climate forecasts. Scientific reports, 8(1), pp.1-13. (DOI:10.1038/s41598-018-20628-2).

Smith, A., Cullis, B. and Gilmour, A. (2001). Applications: the analysis of crop variety evaluation data in Australia. Australian & New Zealand Journal of Statistics, 43(2), pp.129-145. (doi:10.1111/1467-842X.00163).

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(3), pp.269-311. (https://doi.org/10.1111/1467-9876.00154).

Verbyla, A.P., De Faveri, J., Wilkie, J.D. and Lewis, T. (2018). Tensor cubic smoothing splines in designed experiments requiring residual modelling. Journal of Agricultural, Biological and Environmental Statistics, 23(4), pp.478-508. (https://doi.org/10.1007/s13253-018-0334-9).