View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Repository

Applying data mining tools to improve grain quality for growers

Dean Diepeveen[1], Leisa Armstrong[2], Peter Clarke[1], Doug Abrecht[1], Rudi Appels[2] and Matthew Bellgard[3]
[1]Department of Agriculture and Food, Western Australia
[2]Edith Cowan University, Western Australia
[3]Centre of Comparative Genomics, Murdoch University

KEY MESSAGES

Most of the crop variety information from breeding organisations promotes the positive aspects of varieties included, rather than a grower specific view.

The grower needs to take into account across all the new varieties improved germplasm traits when making cropping decisions.

Data mining techniques offer opportunities for summarising diverse data into a coherent format which is understandable by growers.

Data mining offers opportunities to improve the process of identifying and recommending new varieties as it can facilitate the analysis of multiple information sources.

Data mining can provide both the crop breeder and grower with a more translucent view of the general trends in the performance and other plant traits of new varieties.

AIMS

This research uses published information from several Western Australian information sources, which is available to growers and applies data mining techniques in order to determine if improvements can be made in the identification of crop variety performance. Furthermore, the research addresses the idea that data mining may be one approach that can be used to address concerns relating to crop variety recommendations and other factors that affect the variety decisions.

The research aims to assess whether data mining could improve on the current traditional statistical analysis methods used to choose new crop varieties. An examination will be made on the application of multivariate algorithms to a sample wheat dataset and inferences will be made as to whether this could further assist wheat grower's decision-making on crop variety choices.

METHOD

Information on various plant traits such as Grain Yield (GY), Grain Protein (WP), Seivings (SV03) and Grain Weight (HWT) were collected from three field sources from trials carried out in the Western Australian wheatbelt (Ref. 1, 2, 3). Information was collected on the performance of four recently released variety 'Westonia', 'Wyalkatchem', 'Carnamah' and 'Calingiri'. Data mining techniques were used to develop a series of classifying variables such as location, year , trial-type, soil in order to collate the available information. Furthermore, data selection was then done with the outcome variables GY, WP, SVO3, and HWT to ensure some balance between these traits. Results were then collated into a table for analyses using the 'R' software (Ref. 4). Data from 2005 and 2006 was chosen in order to minimise the effect of seasonal climatic conditions across datasets. The data mining technique of multivariate analysis was carried out using Asreml-R (Ref. 5) library within R.

RESULTS

Differences were found in the analysis between the four varieties when using just one trait as apposed to variety yield estimates with four traits. Table 1 displays information published by DAFWA on suggested yields from trials conducted by DAFWA. Table 2 shows the estimated yields from NVT results when using Residual Maximum Likelihood (REML) with variety modelled as fixed and trials as random. Table 3 displays variety estimates using a multivariate model when using all four traits GY, WP, SV03 and HWT. The multivariate model used a random trait by variety term and a sparse trait by trial term.

Table 1. Suggested variety yields (Data source 1)

	Agzone1	Rank	Agzone2	Rank	Agzone3	Rank	Agzone4	Rank	Agzone5	Rank	Agzone6	Rank
Calingiri	2.78	1	2.44	4	3.37	1	1.9	1	1.94	4		
Carnamah	2.63	2	2.76	3	3.15	3	1.67	4	2.1	3	2.86	3
Wyalkatchem	2.55	3	2.82	2	3.08	4	1.85	2	2.23	1	2.98	1
Westonia	2.48	4	2.88	1	3.28	2	1.8	3	2.19	2	2.95	2

Table 2. Estimated variety yields (Data source 2)

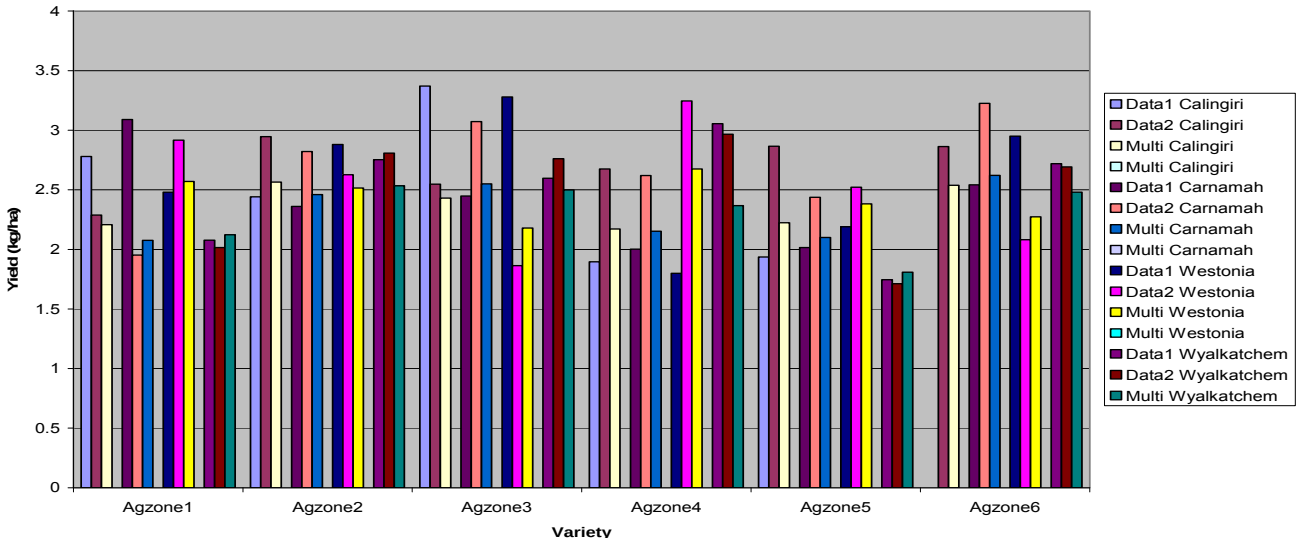
	Agzone1	Rank	Agzone2	Rank	Agzone3	Rank	Agzone4	Rank	Agzone5	Rank	Agzone6	Rank
Calingiri	2.29	2	2.95	1	2.55	3	2.67	3	2.86	1	2.86	2
Carnamah	1.95	4	2.82	2	3.07	1	2.62	4	2.44	3	3.23	1
Wyalkatchem	2.92	1	2.63	4	1.86	4	3.25	1	2.52	2	2.08	4
Westonia	2.02	3	2.81	3	2.76	2	2.97	2	1.71	4	2.69	3

Table 3. Multivariate variety yields using traits GrainYield, Protein, Seivings, and GrainWeight

	Agzone1	Rank	Agzone2	Rank	Agzone3	Rank	Agzone4	Rank	Agzone5	Rank	Agzone6	Rank
Calingiri	2.21	2	2.56	1	2.43	3	2.17	3	2.22	2	2.54	2
Carnamah	2.08	4	2.46	4	2.55	1	2.15	4	2.1	3	2.62	1
Wyalkatchem	2.57	1	2.52	3	2.18	4	2.67	1	2.38	1	2.27	4
Westonia	2.12	3	2.53	2	2.5	2	2.37	2	1.81	4	2.48	3

The results show differences up to 1.1 tonnes/hectare between the different data sources and multivariate estimates. Using all three datasets in the multivariate model results in show much lesser differences between varieties for each agzone when. Figure 1 illustrates these differences as a graph for each variety within agzone combination.

Figure 1: Variety Yield Estimates



CONCLUSION

This research paper shows that growers can make better decisions based on current variety information when techniques are used to combine the data. Data mining techniques, as applied in this study, suggest that when the data is integrated, the yield differences available to growers may not be reflected in the grower paddocks.

KEY WORDS

data mining, crop breeding, grain quality and wheat yield

REFERENCES

South East Premium Wheat Growers Association, 'SEPWA trials', http://www.sepwa.org.au/agronomy.html (accessed 24/08/2007).

Department of Agriculture and Food, Western Australia, 'Wheat variety guide 2007 Western Australia' http://www.agric.wa.gov.au/content/fcp/cer/wh/wheatvarguide07.pdf (accessed: 24/08/2007).

National Variety Trials and Australian Crop Accreditation System Limited, 'Trial report search' http://www.acasnvt.com.au/ACAS/TrialReport.aspx (accessed 24/08/2007).

R Development Core Team (2007). 'R: A language and environment for statistical computing. R Foundation for Statistical Computing', Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Gilmour, A.R, Gogel, B.J., Cullis, B.R. and Thompson, R. (2006), 'Asreml User Guide Release 2.0' ISBN 1-904375-23-5 VSN International Ltd, Hernel Hempstead, HP1 1ES, UK http://www.asreml.com.

Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.

Witten, I.H. and Frank, E. (2005). Data Mining. Practical Machine Learning and Techniques. San Francisco: Morgan Kaufman.

Paper reviewed by: Steve Penny