Identifying key crop performance traits using data mining

D. Diepeveen[1] and L. Armstrong [2]

1 Department of Agriculture and Food Western Australia, Australia.
ddiepeveen@agric.wa.gov.au
2 School of Computer and Information Science, Edith Cowan University, Australia

Abstract

A range of crop related information is distributed to farmers by public and private breeding organizations as well as seed merchants. Farmers use this information to make critical cropping choices which can improve their farm profitability. Frequently, this information highlights the advantages of a particular variety over other released varieties. However, such information is generic and may not be applicable for all farming situations. Better information processes exist that can improve the quality and reliability of this information for individual farming situations. The application of data mining techniques to crop research data enables the customization of information to each individual grower's farming situation.

The challenge from a research perspective is to identify the key attributes that determine crop performance across different farming situations such as geographic location, soil types, and seasonal conditions. The key attributes include nutrition and soil type, grain yield and quality, sowing and harvest dates and tolerance to environmental stresses. This paper applies data mining techniques to explain this crop performance variability. The results from which can be used by growers to identify particular combination of traits should be used to identify high performing varieties. This study used several techniques to identify differences in crop performance across the different geographic regions. Our research findings suggest that growers could use such data mining techniques to identify high performing varieties for their specific locations and farming practices through the adoption of particular varieties.

Keywords: decision making, cropping, farmer, data mining

Introduction

Information on new crop varieties is important to farmers when assessing whether to adopt these varieties. This information can be used as part of the farmer's decision-making process to help to improve crop production. Often changing to a newer crop variety will result in greater yields with little or no change in farm-resource outlays. Thereby, it is important to both the farmer and seed marketer that this variety information is accurate. The seed marketer aims to produce information to sell seed of the variety while the farmer wants to know how true this information when applied to his/her is farming situation. Farmer decision making is made more difficult, when there exists several potential varieties suitable for his/her farming situation.

Governments, and/or farmer organizations may provide unbiased crop variety comparison information for the agricultural industry. The Wheat Variety Guide from the Department of Agriculture and Food Western Australia provides key messages suggesting

varieties for each regions of Western Australia (Department of Agriculture and Food Western Australia 2008a). This information is collated from data obtained by undertaking research trials at various locations and the delivered to farmers in various formats. An example of an Australian wide program is the National Variety Testing program (National Variety Trials 2008).

The critical issue for the farmer is to integrate the information into a format that can then be applied to his/her specific farming location, practices and decision making. This process of decision making often is where the farmer may seeks advice and/or would use computer based tools. Several studies have described the benefits to farmers in using such computer decision aid tools (Donelly et al. 2002; Salmon et al. 2007); for example assisting farmers with sheep grazing.

Data mining is an automated prediction and analytical process which is involved in the transformation of data into useful information and applicable knowledge by having the particular characteristic of exposing hidden patterns. The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks. As such; the process of data mining involves sorting through large amounts of data and discovering patterns in the data (Witten and Frank 2005 p5.). Several agricultural studies have been reported using data mining technologies. (Abdullah and Ansari, 2005, Ekasingh et al., 2005, Holmes et al. 1998). Of these two studies by Abdullah and Ansari (2005) on cotton crops in Pakistan and Ekasingh et al. (2005) on crop modeling provide examples of studies carried out in an agricultural context using data mining.

Data mining has been previously been shown to empower farmers decision making process in terms of location specific crop variety choices (Armstrong et al. 2007). This study found that when various data mining cluster algorithms was used to identify cropping environments within Western Australia, unique cropping environments were identified that did not match the currently used agzones. These agzone form the basis of variety information from a number of organizations within Western Australia. This study suggested that the current variety information provided to farmers may not be indicative of crop performance at all crop growing locations. Armstrong et al. (2007) also proposed that data mining may provide a means to improving the quality of information used to make recommendations and may help farmers in their crop variety decision-making. The research presented in this paper extends the use of data mining tools by identifying key traits within a location or agzone to better identify variety performance and thereby enable the farmer to make better cropping decisions.

Background

The generation of crop variety information is often from research trials, where a new variety is compared with other varieties that are available to the farmer. These trials may be carried out at one or more locations under regional farmer practices. The data is then combined and averages produced for each trial measurement. The summarizing of the variety yield and performance is often affected by variability within and between the trials. With differing environmental and geographic condition at each trial location, the simple average yield may not necessarily capture an accurate measure of the trial measurement. By taking account of these differences, a better average can be obtained.

The critical issue faced by the farmer who is growing a crop at a particular location, is what impact does these plant performance traits have on the selected variety. The farmer needs to acquire information that will assist him to answer questions such as; "Does one variety do better than another for one or more of these conditions?", and/or "Do these conditions or events interact so that a particular variety is able to capitalize and achieve higher yields through a genetic advantage?". Variety performance information for these questions is often not available at the release of a new variety which may increase the difficulties farmers face with their variety selection decisions.

Data mining may alleviate these concerns as it can offer an ability to make generalized data more specific by identifying unique interactions that are common between multiple locations. In particular, these multivariate and meta-analysis approaches are able to identify components of variability across measurements, trials/locations and research groups. This process enables the formation of subsets of data that can then be summarized and analyzed for answering specific questions in the variety performance at a location.

Data mining as with most analytical techniques are dependent on the quality and quantity of the data. The better and more robust the methods used to measure the data, the more able analytical tools are able to identify real differences. Information on quality of the data is often not available but is implied based from the type of research trials undertaken. Knowledge of this measurement precision information can often help direct the type of data mining analyses that can be undertaken

Methods

Data for this study was taken from a subset of the information made available by the Department of Agriculture and Food Western Australia (Department of Agriculture and Food Western Australia 2008b). This variety comparison information included variety trials results for 8 nominated varieties since 1975 for 574 different trial locations. This information also included site specific information and trial management metadata. The locations of these trials are predominately undertaken on farms with a small percentage on government owned research stations.

The analyses use data mining tools from within R (R Development Core Team 2008) to develop a dataset suitable for combined analyses. A multivariate mixed model analysis using R and asreml-r (ref: ASReml 2008) to produce the predictions for all dimensions of the data. These predicted (or simulated) estimates formed a new dataset that was then analyzed with principal components analyses. If general trends were found across the simulated dataset that could be incorporated into the mixed model, then this was done and the simulated dataset re-generated. The predictions from these mixed models was then put into a data-cube implemented in Postgresql (ref: Postgresql 2008), OpenOffice.org Base development version 3 (ref: Openoffice.org) and the database driver (Postgresql-sdbc-driver 2008). The data-cube was then used for reporting and querying variety predictions.

Results and Discussion

The overall Western Australian south west average for several grain traits is displayed in Table 1. This table displays how the predictions have been effected through the averaging process across the 574 research trial locations across the south west agricultural area of Western Australia. The traits included in this table are: grain yield (GY); 100day develop score (DS); harvest height score (HH); straw strength (SS); harvested grain weight (XGWT);

harvested grain protein (XPRO); harvest grain over 2mm sieve (XS20) and 2.5mm sieve (XS25). The larger sieve size is only used for barley grain. Results showed that all traits averages were relatively uniform, even when taking into account differing measurement precision of each trait, This is partly do to the degrees of shrinkage from fitting trait by variety x location term as random in the multivariate mixed-model. The data subsets for specific locations and specific years, proved to be comparable with the actual data when it was available for comparison.

Table 1: Predicted values for the average location and season.

Predicted Values								
	GY	**DS**	**HH**	**SS**	**XGWT**	**XPRO**	**XS20**	**XS25**
Calingiri	2244	30.08	55.65	8.31	37.05	10.68	0.8	
Carnamah	2221.11	31.29	58.54	7.99	37.05	10.75	0.74	
Gairdner	2363.8	30.08	54.21	8.84	37.29	10.13	0.98	17.93
Spear	2091.89	29.29	59.96	7.7	37.05	10.54	1.36	
Stirling	2115.88	32.16	59.5	7.34	36.81	10.93	0.98	5.63
Tincurrin	2208.45	30.69	56.85	7.77	37.05	10.02	2.52	
Westonia	2295.19	33.33	56.48	8.07	37.05	10.49	0.85	
Wyalkatchem	2367.37	32.09	53.49	8.56	37.05	10.67	0.24	

Results from the principal component analyses using a prediction dataset for one Western Australian agricultural region is illustrated as a biplot in Figure 1. This plot displays the relationship between the varieties and the traits. The closer the trait is to the arrow point, the greater the correlation between that variety and trait. This analysis provides a tool for identifying relationships between traits and varieties specific for a location or region that the farmer can assess. In some cases, the farmer may be able to change his/her farmer practice to enable greater yield performance. For example, providing nitrogen fertilizer just prior to grain filling will alter the protein content of the grain given suitable environmental conditions. In other situations, the farmer could minimize the effect of an environmental impact through weed or disease spraying.

The application of data mining may enabled the farmer to drill down into the data and identify interactions that exist between a variety and measured traits. This could also provide an ideal tool for crop science researchers to investigate similar questions across site and locations and establish environmental adaptation of the variety. Firstly, by gaining this information and secondly having the means to alter a varieties performance is considered to be a key to improving plant breeding and developing better varieties in the future.

Conclusion

Farmers have a number of choices to make when selecting new varieties. Tools can provide some assistance to the farmer to enable them to explore the research and field trial data available on crop variety performance. These tools are available to farmers and are often freely available but require some knowledge of how they can be applied. Data mining and the various methodologies associated with it can reduce the complexity of the data enabling farmers to make decisions more easily.

With the increasing dependence of the grains industry on establishing new variety characteristics ensures that data mining and the related methodologies have a role to play in future decision making across the industry. For example, this may be best illustrated where

research and plant breeding trials result in large data sets and important decisions are made by farmers to ensure the appropriate choice of variety.

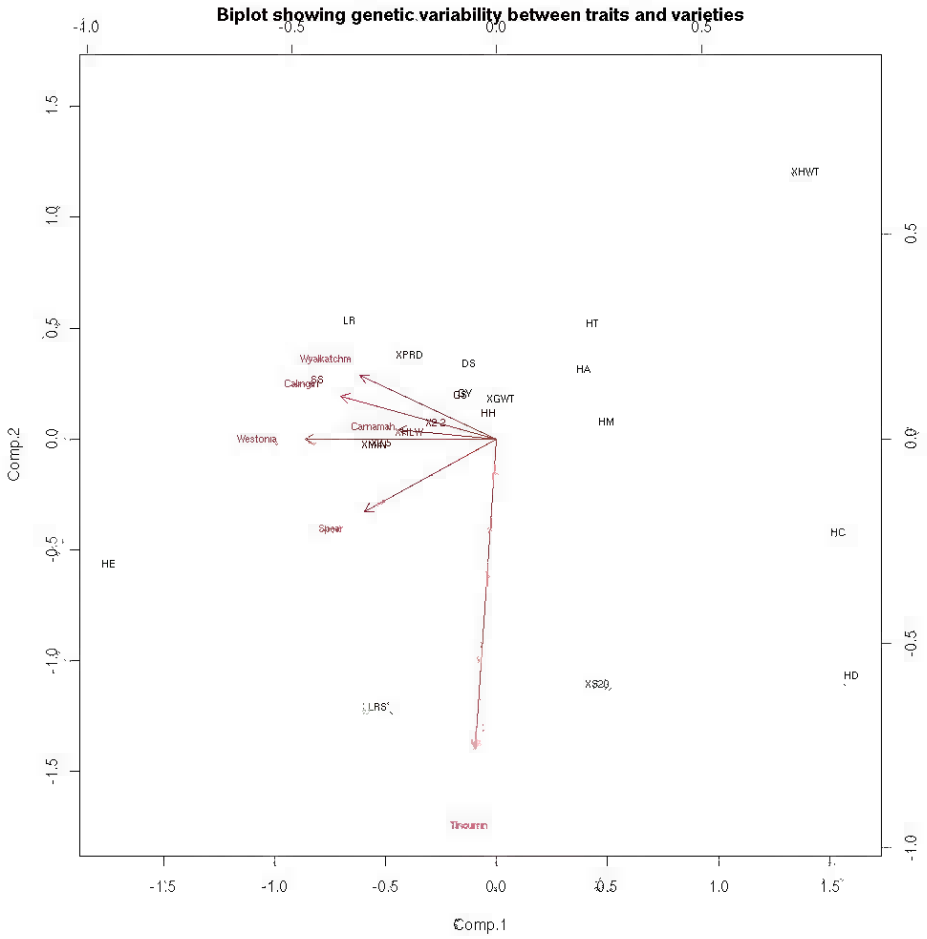

Figure 1. Multivariate *b*iplot displaying genetic variability traits and varieties.

References

Abdullah, A., & Ansari, I. A. (2005) Discovery of cropping regions due toGlobal Climatic Changes using Data Mining. Paper presented at the CAIR Publications, Beijing China.

ASReml (2008) Available from http://www.vsn-intl.com/products/asreml (Accessed 29 May 2008).

Department of Agriculture and Food Western Australia (2008) Wheat Variety Guide 2008 Western Australia (2008a) Available from http://www.agric.wa.gov.au/content/FCP/CER/2008wheatbulletin.pdf (Accessed 29 May 2008).

Department of Agriculture and Food Western Australia (2008b) Available from http://www.agric.wa.gov.au (Accessed 29 May 2008).

Donnelly, J. R., M. Freer, L. Salmon, A. D. Moore, R. J. Simpson, H. Dove, T. P. Bolger (2002) Evolution of the GRAZPLAN decision support tools and adoption by the grazing industry in temperate Australia. Agricultural Systems, Volume 74 (1) p115-

139.
Ekasingh, B. S., Ngamsomsuke, K., Letcher, R. A., & Spate, J. M. (2005) A Data Mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. Paper presented at the Proceedings of the 2005 International Conference on Simulation and Modelling, 17-19 January, Bangkok, Thailand.
Holmes, G., Hunt, L., & McQueen, R. J. (1998) User satisfaction with machine learning as a data analysis method in agricultural research. New Zealand Journal of Agricultural Research,, 41, p577-584.
National Variety Trials (2008) Available from http://www.nvtonline.com.au (Accessed 29 May 2008).
OpenOffice.org (2008) Available from http://www.openoffice.org (Accessed 29 May 2008).
Postgresql (2008) Available from http://www.postgresql.org (Accessed 29 May 2008).
Postresql-sdbc-driver (2008) Available from http://dba.openoffice.org/drivers/postgresql/index.html (Accessed 29 May 2008).
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from http://www.r-project.org. (Accessed 29 May 2008).
Salmon, L., Donnelly, J.R., (2007) Using grazing systems models to evaluate business options for fattening dairy bulls in a region with a highly variable feed supply, Anim. Feed Sci. Technol., doi:10.1016/j.anifeedsci.2007.05.016
Witten, I. H. and Frank E. (2007) “Weka 3: Data Mining Software in Java”. http://www.cs.waikato.ac.nz/~ml/weka/ (Accessed 29 May 2008).