Strategies to interpret yield maps: Predicting grain protein using yield

S. Norng[1], A. N. Pettitt[1], W. M. Strong[2], D. Butler[2]

[1] Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology
[2] Farming Systems Institute, Department of Primary Industries, Toowoomba

Abstract

Protein is a precious commodity for grain farmers with premiums being paid for grain protein above 12.5%. Information on protein content is costly to obtain at present. In this paper, we attempt to predict proteins using yield (a variable that is readily available). We have outlined two methods of regression to predict protein using the agronomic relationship that exists between proteins and yield. The first method is linear regression (nonspatial) using a global neighbourhood (whole field). The second is local regression using local neighbourhoods (data points that lies within discs of a given radius). Different neighbourhood sizes were compared and the optimum radius was selected using the cross-validated mean squared error criterion. The two methods were applied to two different sets of data. The result showed that reasonable coincidental protein maps could be produced.

Key words

linear regression, cross-validation, protein mapping, yield mapping

Introduction

Research works in precision agriculture to date have concentrated on methods to manage the observed variation in yield [3]. This practice is commonly known as site-specific management. The major emphasis for this type of research is to increase farmers' profits by maximising production while at the same time applying minimal input at minimal risk to the environment. Spatial management is still in its infancy, and further research is needed in all areas especially in decision support systems (DSS) before farmers can apply it to their farming system.

In the Australian grain industry premiums are paid according to grain protein content, and particular grades require a minimum and/or a maximum protein content. In certain years, protein premiums make substantial contribution to profits for grain farmers. Therefore it pays to understand the variation in both yield and protein in order to make better management decisions.

An approach to better understand the variation in protein and yield is to look at maps of protein and yield and then relate it to causal variables using agronomic theory. A yield map can be easily produced but a protein map on the other hand is not as easily obtainable.

To produce a protein map on the same resolution as yield, information on protein content from harvested grain must be collected. Yield measurements are readily available through the Global Positing Systems (GPS), and computer-based Geographical Information Systems (GIS). There is a wide range of yield sensors available on the market, from volumetric to mass detection sensors [2].

Despite the rapid progress in sensor technology there are no on-the-go protein sensors on the market as yet. At present, protein sampling is done manually. The practical difficulty involving time and sampling cost can produce spatially sparse protein and spatially dense yield measurements. To produce a more precise protein map, protein samples are needed at a higher spatial resolution.

The purpose of this study is to develop methods for protein prediction such that coincidental protein and yield maps can be produced using a limited number of protein samples.

Materials and method

Data acquisition-The first set of data was collected at Jimbour, southern Queensland, on the 30th of April 1999. The field was planted with sorghum on controlled-traffic lines. Yield was collected in 1 second intervals using a differentially corrected OmniSTAR? GPS system mounted on a harvester. The GPS had a root mean square error of ~ 1m in the horizontal direction. The yield sensor used was an AgLeader? mass flow sensor (www.ag-leader.com). This kind of data collection process has resulted in a large, dense yield data set. Protein was collected every 40-50m along the path of the harvester, noting the location. The protein data is therefore relatively sparse compared with yield.

The second set of data comes from Gurley, northern New South Wales, collected on the 23rd to 24th of November 1999. It was planted with wheat on controlled traffic lines but harvested in any-which-way. The protein data were collected in the same manner as the sorghum data except that the field was large and provided more data. Yield was captured using an Advanced Farming System mass flow sensor (www.casecorp.com). The GPS used was a Motorola Viper, differentially corrected by FM broadcasts with a root mean square error of ~ 1m horizontally.

Method - We develop methods that predict protein using a relationship between yield and protein, and use these predicted and original proteins to produce protein maps. Agronomic research has shown that protein and yield is negatively associated [4, 5]. The first protein prediction method assumes a constant linear relationship between protein and yield. Simple linear regression with a global neighbourhood using the original yields, yields smoothed by a weighted moving average (MA) and yields smoothed by block kriging is considered. The second protein prediction assumes the relationship between protein and yield varies across the paddock. Therefore, local regression using neighbourhoods based on varying disc sizes was applied.

Results and discussion

Global regression results for sorghum (Table 1) and wheat (Table 2) indicated some variation in mean squared error (MSE). The MSE was calculated based on the predicted protein from the regression equation using the three different predictors (original yield, moving average yield and kriged yield). A smaller MSE indicates a prediction method with less error. In this case, the best predictor was kriged yield, which happened to be yield that has been smoothed most. However, the global regression method assumes a constant relationship between protein and yield across the paddock. The R^2 values for all three predictors are relatively small which indicates that the linear regression fit is not very good. A reason for this may be that the relationship between protein and yield maybe changing at the local level. Therefore prediction should not necessarily be made based on global neighbourhood basis.

Yield	Intercept	Gradient	R^2	MSE
Original	10.39	-0.159	0.1603	0.1935
MA	10.49	-0.181	0.1814	0.1885
Kriged	10.76	-0.239	0.2330	0.1767

Table 1: Global regression results for Sorghum

Yield	Intercept	Gradient	R^2	MSE
Original	11.86	0.143	0.0990	0.1778
MA	11.64	0.193	0.1389	0.1207
Kriged	11.26	0.275	0.1976	0.1123

Table 2: Global regression results for Wheat

The local regression improved the MSE values for sorghum (Table 3) and wheat (Table 4). In the local regression, we varied the size of the neighbourhood in order to find the optimal radius size i.e the radius that would give us the smallest error in predicting proteins. The cross-validated MSE is the criterion that we choose over the standard MSE because as we can see in the two tables that with decreasing

neighbourhood sizes (decrease in radius) the MSE approaches zero. This leads us to a false sense of security that we are achieving better predictions if we keep reducing the size of the neighbourhood. This is a result of the trade-off between bias and predictive variance in the calculation of the MSE. The standard MSE does not account for the variance as the number of parameters decreases, it only takes account of the bias. The cross-validated MSE on the other hand accounts for both. Therefore, the cross-validated MSE is more statistically sound as an optimum criterion.

Radius (m)	Constant weight	Weighted	CV-Weighted
200	0.146	0.135	0.148
150	0.140	0.124	0.143
100	0.122	0.109	0.146
80	0.117	0.097	0.150
60	0.101	0.078	0.162
50	0.096	0.060	NA

Table 3: Local regression MSE for Sorghum

Radius(m)	Constant weight	Weighted	CV-Weighted
200	0.083	0.072	0.076
150	0.073	0.063	0.069
100	0.061	0.053	0.063
80	0.056	0.047	0.062
60	0.048	0.040	NA

Table 4: Local regression MSE for Wheat

The optimum radius size for sorghum was 150m according to the cross-validated MSE. The optimum radius size for wheat, using the cross-validated MSE, was 80m. We note that the MSE for wheat did not conform to the cross-validated MSE theory like the sorghum. This may be an indication of the difference between two types of data set. The sorghum being harvested on controlled-traffic is better behaved than the wheat, which was collected from a conventional harvesting operation.

The map produced from the predicted protein (Figure 2) is very similar to the map produced from the original protein data (Figure 1) using block kriging [1]. We note that the kriged protein map, which is just an interpolation of the original protein data, is smoother than the map produced by the protein prediction using the local regression method. Kriging, despite being both a smoother and an interpolator can not predict protein from yield. If we want to predict protein from yield, we would have to use co-kriging [1], which is beyond the context of this paper. The local regression method, although is capable of predicting protein from yield, is still crude in its prediction because it take accounts for spatial dependence between protein and yield in a rather simplistic manner.

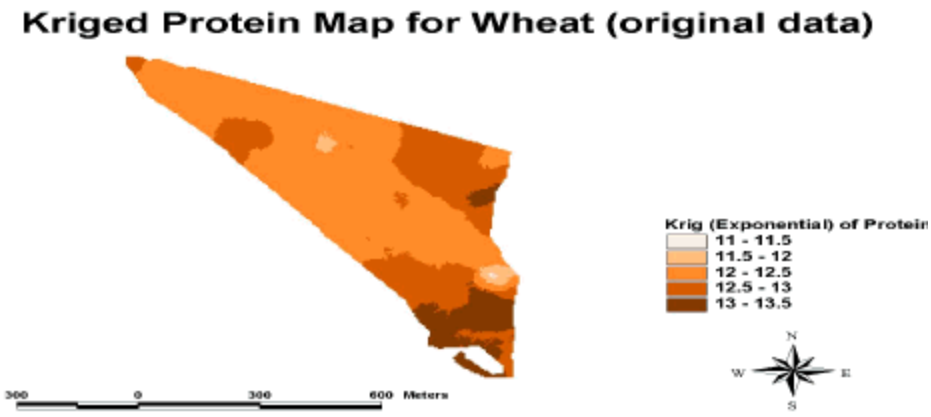

Figure 1: Protein map produced from original data using block kriging

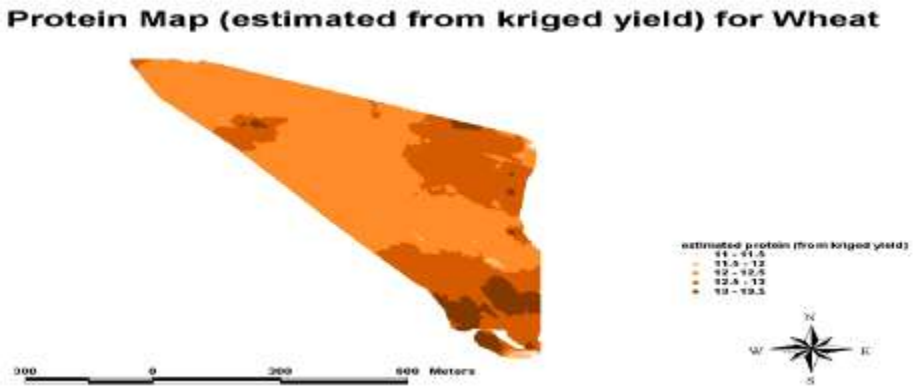

Figure 2: Protein map produced from the local regression of original protein and kriged yield

Conclusion

Prediction using local neighbourhoods gave better predictions than global neighbourhood reflecting that protein and yield co-varied across the paddock. The local regression showed that depending on the nature of the data, the optimum radius size varies. The sparser the data, the larger the optimum radius size and vice versa. The MSE for local regression was smaller than global regression for all radius sizes in both data sets, further supports the logic of local predictions rather than global. There was an indication that controlled-traffic fields are better behaved statistically than standard fields according to the conformity with the cross-validated MSE idea.

The maps produced from the local regression prediction were very similar to maps produced by kriging. Consequently, protein can be predicted from yield and reasonable coincidental maps can be produced from the predictions for management purposes.

Future work will incorporate statistical models that will take into account spatial dependence by modelling covariance function in terms of Euclidean distance. This kind of modelling should give better predictions and thus produce better more reliable maps.

Acknowledgments

We thank Troy Jensen and Rob Kelly of QDPI, and cooperators Mike Smith and Jamie Grant for providing us with the yield and protein data.

References

1. Cressie, N.A.C. 1993. Spatial Statistics. 2nd Ed. (J. Wiley, New York)

2. Moore, M. 1998. An investigation into the accuracy of yield maps and their subsequent use in crop management. PhD Thesis, Cranfield University, Silsoe, UK.

3. Robert, P. C., Rust, R. H., and Larson, W. E. 1996. Proceedings 3rd International Conference. Minneapolis, Minnesota.

4. Strong, W. M. 1981. Nitrogen requirements of irrigated wheat on the Darling Downs. *Australian Journal of Experimental Agriculture and Animal Husbandry* **21**, 424-31.

5. Strong, W.M., and Holford, I.C.R. 1992. Fertilisers and manures. In: Sustainable crop production in the sub-tropics: an Australian perspective. (Eds. A. L. Clarke and P. B. Wylie) (Department of Primary Industries, Queensland).