Integration of precision farming data and spatial statistical modelling to interpret field-scale maize grain yield variability in New Zealand

G. Jiang*[1], M. Grafton[1], D. Pearson[1], M. Bretherton[1], and A. Holmes[2]

[1]Massey University, Palmerston North, New Zealand
[2]Foundation for Arable Research, Christchurch, New Zealand
*Email: g.jiang@massey.ac.nz

Abstract

Spatial variability in soil, crop, and topographic features, combined with temporal variability in weather can result in variable annual yield patterns within a paddock. The complexity of interactions between these yield-limiting factors requires specialist statistical processing to be able to quantify spatial and temporal variability, and thus inform crop management practices.

This paper evaluates the role of multivariate linear regression and a Cubist regression model to predict spatial variability of maize-grain yield at two sites in the Waikato Region, New Zealand. The variables considered were: crop reflectance data from satellite imagery (Sentinel 2 and Landsat 8), soil electrical conductivity (EC), soil organic matter (OM), elevation, rainfall, temperature, solar radiation, and seeding density. The datasets were split into training and validation sets, proportionally 75% and 25% respectively. Both models learn using 10-fold cross-validation. Statistical performance was evaluated by leaving out one year of yield data as the validation set for each iteration, with all remaining years included in the training set for building the prediction models.

In the multiple-year analysis, the Cubist model (RMSE=1.47 and R^2=0.82 for site 1; RMSE=2.13 and R^2=0.72 for site 2) produced a better statistical prediction than the MLR model (RMSE=2.41 and R^2=0.51 for site 1; RMSE=3.37 and R^2=0.30 for site 2) for the prediction of the validation set. However, for the leave-one-year-out analyses, the MLR model provided better statistical predictions (RMSE=1.57 to 4.93; R^2 = 0.15 to 0.31) than the Cubist model (RMSE = 2.62 to 5.9; R^2 = 0.05 to 0.14) for Site 1. For Site 2, both models produced poor results.

Yield data for additional years and inclusion of more independent variables (e.g. soil fertility and texture) may improve the models. This analysis demonstrates that there is potential to use statistical modelling of spatial and temporal data to assist farm management decisions (e.g. variable rate application, precision land levelling, irrigation, and drainage). Once the functional relationship between within-paddock yield potential and complementary variables is established, it should be possible to provide an accurate management prescription, enabling variable rates of an input (e.g. plant density, fertiliser) to be applied automatically across the paddock based on the "yield-input" response curve.

Keywords: spatial yield prediction, precision farming, satellite imagery.

1. Introduction

The practice of precision farming in the New Zealand (NZ) arable sector began in the early 1990s. Since then, there has been wide-scale uptake of precision farming tools such as guidance systems and variable-rate irrigation. However, the commercial uptake of variable rate applications (VRA) (which have the potential to improve farming efficiency) has been limited due to the lack of available information to estimate yield response (Holmes and Jiang, 2018).

With the increasing availability of regularly captured spatial data and publicly-available satellite imagery and climatic records, there is potential to use this information to inform farm management decisions. However, appropriate spatial analysis techniques are still limited for this type of application. Progress has been made on delineating management zones (MZs) within-paddocks to represent similar yield-limiting factors based on a variety of spatial information (e.g. historical yield data, geo-referenced aerial photographs, soil and topography features) using spatial classification techniques (Khosla et al., 2010; Hedley et al., 2017; Holmes and Jiang, 2018). From this, a single rate of an input (e.g. fertiliser, seeding rate) can be applied to each MZ. However, it is difficult to quantify spatial yield and temporal variability without a detailed level of understanding of yield potential and crop response to specific variables (e.g. climate, crop type, management practices) (Kitchen et al., 2003; Guastaferro et al., 2010).

Statistical modelling techniques (e.g. stepwise multiple linear regression) have been used to help understand the relationship between crop yield and measured soil and site parameters, using large, spatial, multivariate datasets. Improved results have also been reported with more complex machine learning techniques such as neural networks (Kitchen et al., 2003; Drummond, 2003). However, the implementation of neural networks can be computationally time-expensive. The time used for training the model could vary from hours to weeks depending on the structure of the neural network (e.g. the number of hidden layers, neurons) and the optimisation methods, which increases the difficulty of processing spatially-dense precision farming data.

This paper presents an approach that attempts to estimate yield by integrating spatial yield data, seeding density, high-precision elevation points, multispectral satellite imagery (i.e. NASA's Landsat-8 and Sentinel-2 ESA's missions), soil EC, and meteorological data. The approach evaluated was the application of a multivariate linear regression (MLR) and a Cubist regression model to predict within-paddock maize-grain yield potential. It is hypothesised that such a modelling approach can help farmers modify crop management practices to maximise yield and minimise costs.

2. Method

2.1. Sites

Two sites in Waikato were chosen for this study because of their consistent within-site management histories. Site 1 (175.372 E, -37.835 S) is located at Tamahere, 10 km south of Hamilton. It is a 10-ha paddock which has been dedicated to growing maize. Spatial yield data of maize (Zea mays) grain was collected over four years (2014, 2015, 2017, and 2018) by a yield monitor fitted on an 8-row (6 m swath) combine harvester with a GPS receiver. Site 2 (175.487 E, -37.676 S) is a 23-ha paddock located 5 km southwest of Morrinsville. Three years of maize-grain yield data (2014, 2017

and 2018) were collected. For both sites, spatial data points were recorded at 1-second intervals during harvest.

2.2. Data modelling

Models used

The two statistical methods used for the analysis were multivariate linear regression (MLR) and Cubist tree regression. The Cubist regression model is a technique primarily used in remote sensing studies for handling large datasets and has, in the past, reported promising results when predicting continuous variables. It has a faster training speed than other computationally intensive machine learning methods such as random forest and neural networks (Aviv and Lundsgaard- Nielsen, 2017; Noi et al., 2017). For MLR, the dependent variable (yield) needs to be normally distributed. To achieve a normal distribution, the historic yield was transformed using natural logarithm.

Multiple-year analysis

The "split-sample" approach was used to measure prediction accuracy, in which a subset (validation set) of the data is withheld from training. A measure of the accuracy of prediction on this validation set is then reported. In the multiple-year analysis, data subsets were created from the maize data for all available years (2014, 2015, 2017 and 2018). Each training set consisted 75% of the data, randomly sampled (with no replacement), and each validation set contained the remaining 25%. Ten-fold cross-validation was then performed to select the best model parameters in order to optimise model performance. In this procedure, the data are divided into 10 subsets of equal size. The regression technique is then repeated 15 times, with each repetition leaving out one of the validation subsets, and using only that subset to compute the root mean square error (RMSE).

Leave-out-one-year analysis

Because of the spatially sparse meteorological data, multiple-year data was cross-validated by withholding one year of data as the validation set for each iteration, with all remaining years included in the training set. The training set was used to predict yields for the year that was held out as a validation set. This process was iterated over the data for all the years and RMSEs were computed. This will provide an indication of the ability of the trained model to handle new information (i.e. yield data collected from an additional harvest).

3. Results

3.1. Model outputs

Multiple-year analysis

For both sites, both models demonstrated reasonable accuracy for predicting yield (Table 1), since the prediction errors for the validation set (RMSEs) were smaller than the standard deviation (SD) of the multiple-year predictions. For both sites, the cubist model showed that it was able to explain 70% - 80% of yield variation. This is better than MLR which explained 30% - 50% of the variation.

	MLR model		Cubist model		Observed yield	
				Site 1		
	RMSE	R^2	RMSE	R^2	Mean	SD
Training	0.27	0.47	0.16	0.81	10.06	3.36
Validation	2.41	0.51	1.47	0.82	10.05	3.36
				Site 2		
	RMSE	R^2	RMSE	R^2	Mean	SD
Training	0.34	0.29	0.22	0.69	11.03	4.02
Validation	3.37	0.31	2.13	0.72	11.03	4

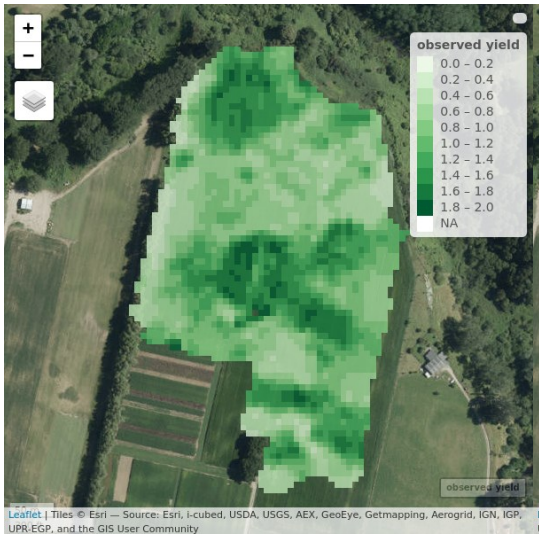
Table 1: Prediction results of the multiple-year analysis in training and validation.

Leave-out-one-year analysis

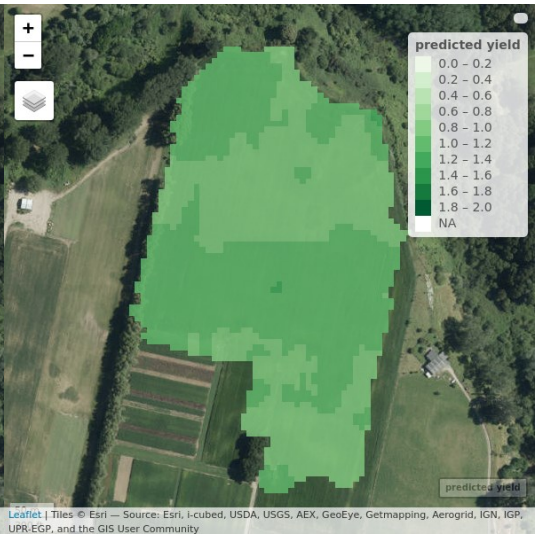
In the leave-out-one-year analysis, the results provided by the Cubist model are less accurate. The MLR model produced lower RMSEs and higher R^2 than the Cubist model for all individual years for Site 1 (Table 2). The higher RMSEs in the Cubist model may be a result of skewed data and suggests that the more complex machine learning models do not necessarily perform better at predicting within-paddock yield potential for a new harvest than a simple linear model as the data distribution is often unknown. The predicted yield maps are visually presented in Figure 1, contrasting observed with predicted maize yield for all available years.

	MLR model		Cubist model		Observed yield	
				Site 1		
Leave out	RMSE	R^2	RMSE	R^2	Mean	SD
2014	4.93	0.28	5.9	0.14	8.54	3.6
2015	3	0.18	3.51	0.05	13.5	3.18
2017	1.57	0.31	2.92	0.08	7.93	1.61
2018	1.85	0.15	2.64	0.14	11.23	1.84
				Site 2		
Leave out	RMSE	R^2	RMSE	R^2	Mean	SD
2014	9.09	0.3	7.33	0.09	12.44	6.3
2017	3.91	0.15	3.04	0.03	10.15	2.72
2018	3.15	0	2.22	0.05	10.71	1.69

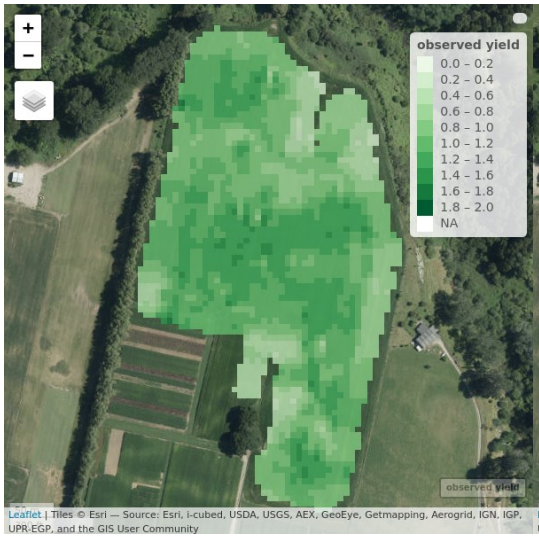
Table 2: Prediction results of the leave-out-one-year analysis.



(a) Observed yield 2014

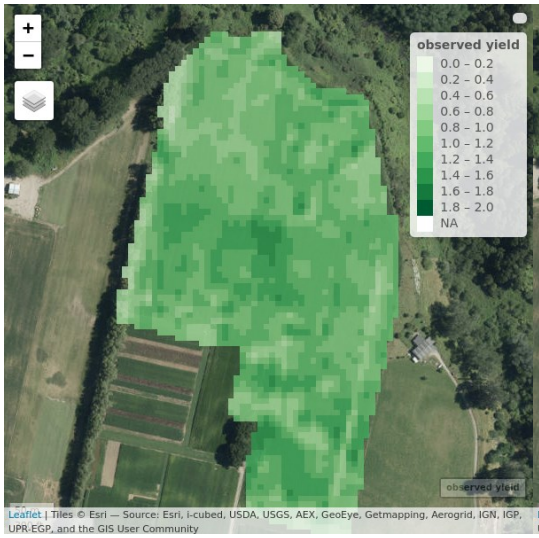


(b) Predicted yield 2014

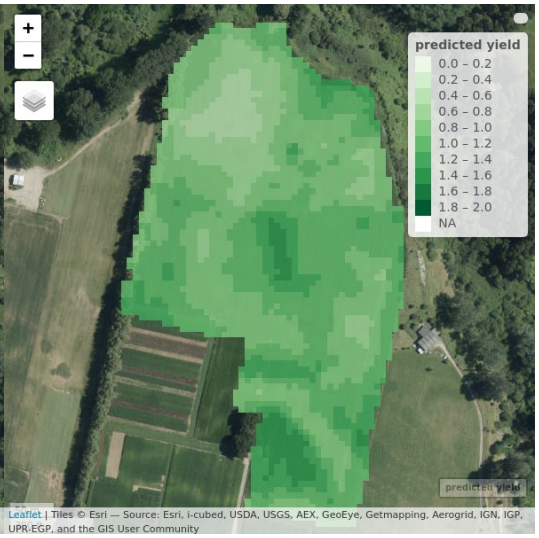


(c) Observed yield 2015

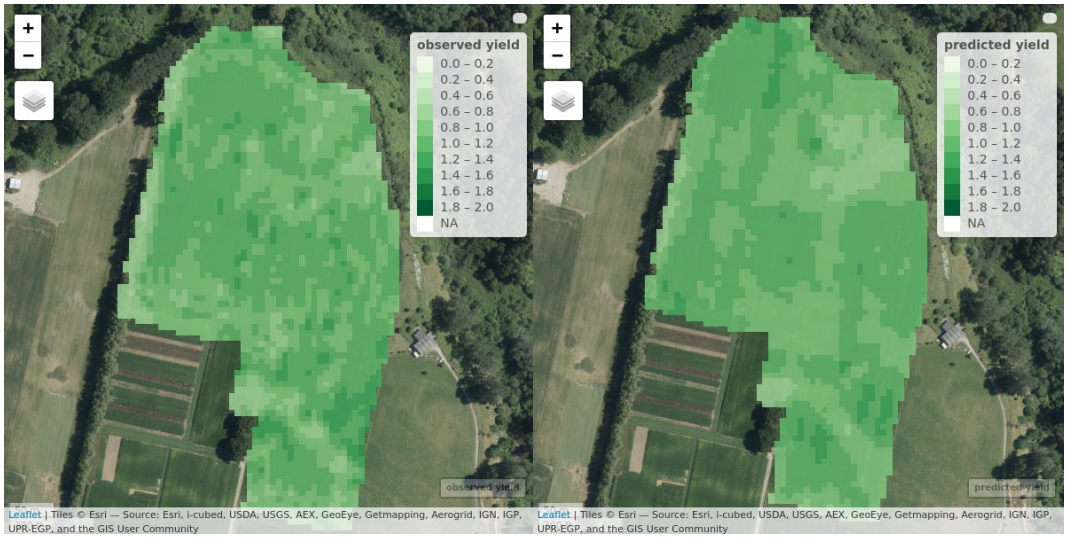
(d) Predicted yield 2015



(e) Observed yield 2017



(f) Predicted yield 2017



(g) Observed yield 2018 (h) Predicted yield 2018

Figure 1: Observed vs predicted maize yield maps (each yield map was normalised).

4. Discussion and conclusion

The results indicate that there is a potential to predict within-paddock crop yield using statistical modelling of spatial and temporal data. Given the model responses, yield data for additional years, and inclusion of further relevant variables may improve the model. Data consistency is a potential problem in the acquisition of useful remote sensing imagery at an appropriate growth stage for crop management in New Zealand, due to cloud coverage. Nevertheless, UAVs (unmanned aerial vehicles) are increasingly being used in agricultural applications and may offer alternatives to currently available satellite imagery by providing more relevant scales of data capture and the ability to capture information at more appropriate times of the year. Whilst acquiring better data to improve the model might remain a challenge in the near future, the application of the approach used in this study offers advantages over techniques that use spatial data collected from intensive and expensive grid sampling (Drummond et al., 2003; Liu et al., 2001). The minimal costs associated with the approach employed in this study are thus more likely to be of commercial interest to New Zealand farmers and may potentially inform crop management thereby contributing towards improved yield and farm input efficiencies.

5. Acknowledgements

I would like to express my thanks to the Foundation for Arable Research (FAR) for providing funding and data for conducting this research, and my supervisors for professional guidance.

6. References

Aviv, T and Lundsgaard-Nielsen V. 2017. *Ensemble of cubist models for soy yield prediction using soil features and remote sensing variables*. In 23rd Conference of Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada.

Drummond, S T, Sudduth, K A, Joshi, A, Birrell, S J, and Kitchen, N R. 2003. *Statistical and neural methods for site-specific yield prediction*. Transactions of the ASAE, 46(1):5.

Guastaferro, F, Castrignano, A, De Benedetto, D, Sollitto, D, Troccoli, A, and Cafarelli, B. 2010. *A comparison of different algorithms for the delineation of management zones*. Precision agriculture, 11(6):600-620.

Hedley, C, Ekanayake, J, McCarthy, A, et al. 2017. *Precision irrigation: trials to assess impacts on crop yield. In "Doing More with Less"*. Proceedings of the 18th Australian Agronomy Conference 2017, Ballarat, Victoria, Australia. 24-28 September 2017. Pp 1-4. Australian Society of Agronomy Inc.

Holmes, A, and Jiang G. 2018. *Increasing profitability & sustainability of maize using site-specific crop management in New Zealand*. In Proceedings of the 14th International Conference on Precision Agriculture, Montreal, Quebec, Canada.

Khosla, R, Westfall, D, Reich, R, Mahal, J, and Gangloff, W. 2010. *Spatial variation and site-specific management zones*. In Geostatistical applications for precision agriculture, Pp. 195-219. Springer.

Kitchen, N, Drummond, S, Lund, E, Sudduth, K, and Buchleiter, G. 2003. *Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems*. Agronomy journal, 95(3):483-495.

Liu, J, Goering, C, and Tian, L. 2001. *A neural network for setting target corn yields*. Transactions of the ASAE, 44(3):705.

Noi, P, Degener, J, and Kappas, M. 2017. *Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data*. Remote Sensing, 9(5):398.