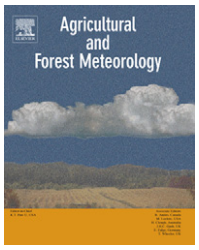
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/agrformet



Ensemble data mining approaches to forecast regional sugarcane crop production

Y.L. Everingham [a,*], C.W. Smyth [c], N.G. Inman-Bamber [b]

[a] School of Mathematics, Physics and Information Technology, James Cook University, Townsville, Qld 4814, Australia
[b] CSIRO Sustainable Ecosystems, Davies Laboratory, Townsville, Qld 4814, Australia
[c] Disaster Prevention Research Institute Kyoto University Gokasho, Uji, Kyoto 611-0011 Japan

ARTICLE INFO

Article history:
Received 4 December 2007
Received in revised form
27 October 2008
Accepted 28 October 2008

Keywords:
Predict
Forward stagewise
Simulation
Top-down
Machine learning
Lasso

ABSTRACT

Accurate yield forecasts are pivotal for the success of any agricultural industry that plans or sells ahead of the annual harvest. Biophysical models that integrate information about crop growing conditions can give early insight about the likely size of a crop. At a point scale, where highly detailed knowledge about environmental and management conditions are known, the performance of reputable crop modelling approaches like APSIM have been well established. However, regional growing conditions tend not to be homogenous. Heterogeneity is common in many agricultural systems, and particularly in sugarcane systems. To overcome this obstacle, hundreds of model settings ('models' for convenience) that represent different environmental and management conditions were created for Ayr, a major sugarcane growing region in north eastern Australia. Statistical data mining methods that used ensembles were used to select and assign weights to the best models. One technique, called a lasso approximation produced the best results. This procedure, produced a predictive correlation (r^{cv}) of 0.71 when predicting end of season sugarcane yields some 4 months prior to the start of the harvest season, and 10 months prior to harvest completion. This continuous forecasting methodology based on statistical ensembles represents a considerable improvement upon previous research where only categorical forecast predictions had been employed.

1. Introduction

Predicting crop production for agricultural industries is an important task. This is especially the case for agricultural industries that forward-sell the crop to customers well before the crop harvest commences. Overestimating crop production can lead to major shortfalls in meeting customer demands. Typically, this situation requires the seller to purchase the commodity from a competitor at a higher price to fulfil the sale and honour contractual arrangements with the customer. Adverse effects can also be associated with underestimating crop production. In years when pre-harvest crop prices are higher, profits can be lost by not securing the maximum amount of crop at the higher price. For many agricultural industries, underestimating crop production can also lead to difficulties in logistical management such as managing limited storage supplies and transporting arrangements. Early and accurate crop forecasts offer substantial benefits to industry through increased profitability, better logistical arrangements and improved customer satisfaction.

A diverse range of agricultural industries rely on accurate and timely crop forecasts. These industries extend to, but are not limited to, wheat, corn, maize, and cotton and the focus of this study—sugarcane (Bastiaanssen and Ali, 2003; Evering-

* *Corresponding author*. Tel.: +61 7 4781 5067; fax: +61 7 4781 5880.
E-mail address: yvette.everingham@jcu.edu.au (Y.L. Everingham).
0168-1923/$ – see front matter.
doi:10.1016/j.agrformet.2008.10.018

ham et al., 2003, 2007; Hansen and Indeje, 2004; Hansen et al., 2004; Zhang et al., 2005). The Australian sugar industry generates between one and two billion AUD dollars to the nation's economy annually. Sugarcane starts as a plant crop. The crop is harvested between June and November and regrows (ratoons) for harvesting approximately 12 months later, depending on the region. The Australian sugarcane industry makes initial crop forecasts, 12 months prior to knowing the exact size of the crop (December).

Crop forecasts can be generated by biophysical models that describe the interaction between the plant and the environment. Biophysical models use mathematical equations to derive accumulated biomass on the basis of observed and in some cases forecasted meteorological inputs such as daily temperatures, radiation and rainfall. Some biophysical models, particularly those which form the nucleus of decision support systems like APSIM (Keating et al., 1999) and DSSAT (Jones et al., 2003), allow for more detailed information about the environment and operational procedures to be incorporated into the modelling analysis procedure. Regional crop forecasting procedures that use biophysical models can be categorised into two strategies. We refer to these strategies as ''bottom-up'' or ''scaling-up'' (Hansen and Jones, 2000), and ''top-down'' or ''scaling-down'' (Potgieter et al., 2005; Shorter et al., 1991). Bottom-up approaches consider components that influence biomass production at very detailed levels of the system. Knowledge about these subsystems can be linked to gain knowledge about larger sub-systems. Akin to Shorter et al. who liken 'up' approaches by considering the detailed mechanistic process at the biochemical level through to the cell, plant and crop levels, our idea of a bottom-up approach is one where detailed knowledge about the biophysical conditions is used to predict yields on say a $1\ m \times 1\ m$ grid. As part of the bottom-up process, this knowledge would be gradually merged to predict yields across a larger spatial domain (e.g. block, farm or shire). Top-down yield forecasting methods reverse this approach. Top-down procedures consider the major system components that contribute to biomass accumulation and successively integrate information at more detailed levels of the system as required.

A disadvantage associated with bottom-up approaches is errors can aggregate through successive accumulation of inaccuracies which are often present in fine scale data across wide spatial domains (Everingham et al., 2007; Hansen and Jones, 2000). More often than not however, this highly detailed data is not available. Top-down approaches can oversimplify the problem at hand. Oversimplification occurs by assuming regional homogenous environmental and management conditions, when it is common knowledge these conditions can vary quite substantially for many cropping systems.

To compensate the invalid assumption of homogeneity, some authors consider many different environmental and management conditions that could be representative of the larger system. For example, Potgieter et al. (2005) estimated shire scaled sorghum yields in Australia by implementing a crop modelling procedure that considered many different settings of planting triggers, maximum number of sowings, soil water holding capacity and cropping stress periods. Potgieter et al. (2005) searched for the parameter combination that maximized cross-validated correlation with observed sorghum yields. The cross-validated correlations ranged from 0.5 to 0.9 for the different sorghum-growing shires in Australia. The authors applied a Monte Carlo permutation testing procedure to check the selected model was unlikely a consequence of chance. Everingham et al. (2007) implemented a similar strategy to Potgieter et al. (2005) and considered a procedure for optimising different input parameters to the APSIM sugarcane model when predicting regional sugarcane yields within the Australian sugarcane industry. Rather than using the single best model based on the optimal parameter settings, like the approach used in Potgieter et al. (2005), Everingham et al. (2007) selected a set of models that produced the highest cross-validated correlation measure and simply averaged the subsetted models to provide probabilistic categorical forecasts of ''high'', ''medium'' or ''low'' crop sizes for major sugarcane growing regions in Australia. These categorical forecasts were produced in December, 7 months prior to harvest commencement (approx. June). Their technique produced cross-validated correct classification rates between 55 and 72% for the different sugarcane growing regions. The correct classification rate was the number of years the predicted crop production category was equal to the observed category, divided by the total number of years tested, expressed as a percentage. These rates were substantially better than the chance rate of 33%. Categorical forecasts in December were valued by marketers who preferred a broad categorical forecast with a long lead time over a sharper forecast with a short lead time. Nevertheless sharper forecasts are required to optimise storage and shipping arrangements across the industry and to plan the duration of the harvest season as conditions become suitable for it to start.

Identifying and selecting a subset of quality models from a vast array of models (see for example Potgieter et al. (2005) and Everingham et al. (2007)) can be challenging. Ensembles offer one solution to this problem. Ensembles are a statistical framework for efficiently combining information obtained from various sources and models. The individual models included in the ensemble are commonly referred to as ensemble members. Each ensemble member tries to predict a response variable. The ensemble combines the predictions made by the members to predict the same response. By combining many models to give a single overall model, ensembles are more stable and often more accurate than any individual model (Breiman, 2001; Krogh and Vedelsby, 1995).

Ensembles are a data mining procedure that have been used to improve the predictive capability of models across a wide range of disciplines such as chemometrics, bioinformatics, ecological modelling and have been used extensively in climate modelling literature (Mevik et al., 2004; Knutti et al., 2002; Martelli et al., 2003). Consequently, most applications of ensembles for yield forecasting purposes have predominantly been integrated within climate forecasting systems. This approach takes into consideration the uncertainty associated with the simulation of the climate system (Doblas-Reyes et al., 2006). Ye et al. (2006) however take a different approach. They used an ensemble of models to successfully predict citrus yields using airborne hyperspectral imagery. There exists many unexplored opportunities for incorporating ensemble learning as part of yield prediction methods in agricultural enterprises.

Sugarcane yield forecasting efforts based on cropping systems simulators have largely been concentrated in South Africa (Bezuidenhout and Schulze, 2005; Bezuidenhout and