Novel approach to mid-season cotton yield forecasts

By Patrick Filippi – The University of Sydney

There has been a lot of interest in developing methods for forecasting cotton yield. Yield forecasts are useful for a variety of stakeholders, including growers, gins, and governments. Many studies focus on forecasting yield for large areas (e.g. regions), and late in the season. But growers need forecasts of yield at fine resolutions (within field), and earlier in the season in order to adaptively manage their cropping systems. This also aids growers in forward-selling and managing insurance and contracts.

Current approaches

There are currently a few methods that have been used to forecast cotton yield:

- Simulation models, such as APSIM-OZCOT (Agricultural Production Simulation Model). These simulation models can be accurate, but rely on numerous, and detailed specific input parameters. This information is usually only available at point locations measured in high detail, such as research stations. This makes it difficult to use simulation models to provide forecasts of yield for growers.

- Another popular approach is to use imagery from unmanned aerial vehicles (UAVs) to forecast yield. While relatively accurate predictions of yield can be made with UAVs, they can be expensive and time-consuming to operate, particularly over large areas. There are also issues with calibration, and most studies have only used data from a few fields or seasons, creating limitations on the applicability in the real world.

An opportunity to create a novel approach

At the University of Sydney (USYD), we have been developing a novel approach to forecast cotton yield to try and overcome the pitfalls and limitations of other current approaches. This involves using yield monitor data over many fields and farms, combining this with other publicly available datasets, and then creating machine learning models to forecast yield mid-season.

Many cotton growers have a good store of yield monitor data across many fields, and for many seasons, which is a fantastic asset that has been typically underutilised. We are also currently in the ‘digital revolution of agriculture’, meaning that we now have access to an enormous amount of data that not only describes crop yield, but also the factors that relate to yield (e.g. weather, satellite imagery, soil data). This provides a promising opportunity to combine all this to forecast cotton yield mid-season for cotton growers.

Aims of the study

- Develop a novel and cost-effective approach to forecast cotton yield mid-season (start of January) using historical yield mapping datasets and public data; and,
- Assess how the accuracy of cotton yield forecasts change as the spatial resolution of predictions change: from 30 metres, to fields, to the farm scale

Study area and period

A collection of large irrigated cotton farms in the Gwydir Catchment were used in this study. The total area encompasses about 13,000 hectares. Yield maps from yield monitors were available from 68 different fields, and 14 different seasons. In total, 253 yield maps of cotton were used in this study.

Yield modelling concept – topsaw

We know that crop yield is driven by a suite of different variables that vary in both space and time. These include:

- t = Topography (e.g. elevation, slope);
- o = Organisms (e.g. pests, disease, weeds);
- p = Plant measurements (e.g. reflectance, vegetation indices, canopy biomass);
- s = Soil (e.g. soil type, soil constraints, nutrient content);
- a = Agronomy/management (e.g. variety, sowing date, irrigation, previous crop); and,
- w = Weather (e.g. rainfall, temperature, solar radiation)

This can be summarised in the following formula, where crop yield (Y), is a function of all of these variables and the interaction between them:

$$ Y = f(t, o, p, s, a, w) $$

This topsaw framework is a simple conceptual tool which aims to describe the primary driving factors of crop yield. Combining the topsaw factors with yield data in a model allows forecasts of yield models to be made.

Datasets

As cotton yield is driven by a suite of controllable, and uncontrollable variables, many datasets were collated for use in this study. The specific variables used in the yield forecasting models are described in Table 1. The goal was to capture as many variables as possible to populate the topsaw modelling framework and represent the factors that control cotton yield.
Yield forecasting model and validation

The aim was to forecast cotton yield from the start of January using a random forest machine learning model. To do this, all the spatial and temporal data described in Table 1 was joined with the corresponding yield data at the relevant location and season. Many studies lack proper calibration and validation of yield models, but this study adopted a robust leave-one-year out cross-validation. This means that all data from one year is removed, and then the model is built on the remaining years of data. For example, all years from 2002 to 2016 are used to build the model, and then predictions are made on the 2017 year. This is then repeated for all combinations of seasons, and results in a realistic assessment of the accuracy of the model.

Results

The results showed that the most accurate forecasts of yield could be made at the farm scale, followed by the field scale, than the 30 metre resolution (Table 2). This makes sense, as it is easier

<table>
<thead>
<tr>
<th>Spatial resolution</th>
<th>LCCC</th>
<th>RMSE (b/ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 metre</td>
<td>0.42</td>
<td>2.11</td>
</tr>
<tr>
<td>Field</td>
<td>0.63</td>
<td>1.72</td>
</tr>
<tr>
<td>Farm</td>
<td>0.65</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Results were obtained with leave-one-year-out cross-validation. LCCC of 1 is a perfect fit. The lower the RMSE the better.

---

Table 1: The data used, the spatial resolution, and the associated topsaw factor

<table>
<thead>
<tr>
<th>Data type</th>
<th>Data description</th>
<th>Resolution</th>
<th>Factor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield</td>
<td>Cotton lint yield monitor data</td>
<td>10 metre</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Satellite imagery</td>
<td>EVI – Landsat 7 (maximum Oct–Jan)</td>
<td>30 metre</td>
<td>p</td>
<td>Spatial-temporal</td>
</tr>
<tr>
<td></td>
<td>NDVI – Landsat 7 (maximum Oct–Jan)</td>
<td>30 metre</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ET – MODIS (cumulative Oct–Jan)</td>
<td>500 metre</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Rainfall</td>
<td>Aug–Sep (pre-sowing)</td>
<td>5000 metre</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oct–Jan (in-crop rainfall)</td>
<td>5000 metre</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>Oct–Jan (in-crop day degrees)</td>
<td>5000 metre</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>Topographic data</td>
<td>Digital elevation model</td>
<td>5 metre</td>
<td>t</td>
<td></td>
</tr>
<tr>
<td>Gamma radiometrics</td>
<td>Potassium (K)</td>
<td>100 metre</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td>Soil maps</td>
<td>Clay content (30–60 cm)</td>
<td>90 metre</td>
<td>s</td>
<td></td>
</tr>
</tbody>
</table>

---

Barcoo is a superbly appointed lodge at Dinner Plain in the heart of Victoria’s high country. This year round playground offers trout fishing, magnificent scenery, great restaurants, peace & quiet and other cool activities.

- 4 bedrooms (all with queen size beds)
- 3 bathrooms
- Spa pool
- Fully equipped with all mod cons
- Sleeps up to 16

GREAT VALUE FOR LARGE OR FAMILY GROUPS

Further details phone 03 5178 3088 or www.dinnerplain.com

---

ALL THE WINTER SPORTS
to predict yield across a large area, rather than at each individual point within a field. The Root Mean Square Error (RMSE) is the standard deviation of the prediction error – it essentially represents the average error of the model. This means the lower the RMSE, the better. The RMSE in this study ranged from 0.77–2.11 bales per hectare, but as the average yield is around 12 bales per hectare, this can be considered relatively good. A LCCC (correlation coefficient) of one is a perfect fit, and the models in this study obtained LCCC values from 0.42–0.65.

Figure 2 shows a plot of the observed and predicted yield at the farm-scale for each year. While cotton yield could be forecasted quite accurately in some years with most points sitting near the 1:1 line (e.g. 2005 and 2016), some years were relatively poorly predicted (e.g. 2017).

Poor prediction of some seasons could be due to extreme weather events. For example, in the 2017 cotton-growing season, the study area (Moree) experienced 54 days above 35°C in January/February, the longest period above this temperature in the history of any weather station in all of NSW. These extreme temperatures put the cotton crop under stress, and physiologically a considerable amount of water is simply used to cool the plant down, rather than for growth.

Conclusions and future directions

- Relatively accurate forecasts of yield could be made using the developed approach, which involved using large yield mapping datasets across many fields/farms and seasons, publicly-available datasets, and machine learning.
- The accuracy improved as the spatial resolution increased from 30 metre fields to the farm-scale.
- There is room to improve the accuracy of the models in the future. This will involve including more datasets, such as management information, better satellite imagery (e.g. CCCI), and better soil constraint maps.
- Yield forecasts at the within-field and field scale mid-season could help growers to adaptively manage according to the variability of cotton crops.
- Mid-season yield forecasts at the farm-scale also have considerable advantage in planning ahead, forward selling, harvesting logistics, and managing contracts and insurance.
- The topsaw framework developed is flexible, and uses satellite imagery of the crop while also taking advantage of the other factors that drive yield, such as weather and soil.
- A strong advantage of the approach implemented here is that it only uses readily available public datasets as predictor variables in the model. This means that no further data collection is required, and that it could be implemented at any cotton-growing field in any region of Australia. This would obviously require further calibration of the model to ensure that each region is appropriately represented.
- An operational tool for the industry could be developed if it is combined with other industry tools such as CottonMap and SataCrop.
- Overall, accurate mid-season forecasts of cotton yield could lead to improved production, input use efficiency, and profitability for growers and the Australian cotton industry as a whole.

For more information contact Patrick Filippi E: patrick.filippi@sydney.edu.au

Thanks to PCT (Precision Cropping Technologies) for providing access to the data, and to the farmers and agronomists who assisted in this project. Thanks to Guna Nachimuthu (NSW DPI) for his collaboration throughout the project.