DEVELOPING SUGAR CANE YIELD PREDICTION ALGORITHMS FROM SATELLITE IMAGERY

By

ANDREW ROBSON¹, CHRIS ABBOTT¹, DAVID LAMB², ROB BRAMLEY³

¹Department of Employment Economic Development and Innovation, Kingaroy
²University of New England, Armidale
³CSIRO, Adelaide, SA

andrew.robson@deedi.qld.gov.au

KEYWORDS: Yield Forecasting, Satellite Imagery, SPOT5, GNDVI.

Abstract

THE RESEARCH PRESENTED in this paper discusses the accuracies of remote sensing and GIS as yield prediction tools at both a regional and crop scale over three Australian cane growing regions; Bundaberg, Burdekin and the Herbert. For the Burdekin region, the prediction of total tonnes of cane per hectare (TCH) produced from 4999 crops during the 2011 season was 99% using an algorithm derived from 2010 imagery (green normalised difference vegetation index) and average yield (TCH) data extracted from 4573 crops. Similar accuracies were produced for the Bundaberg region during 2010 (95.5% from 3544 blocks) and 2011 (91.5% for 3824 crops) using a Bundaberg specific algorithm derived from 2008–2010 imagery and yield data. The Bundaberg algorithm was also accurate in predicting yield at specific in-crop locations (91.5% accuracy; SE = 0.028).

Introduction

Accurate in-season predictions of regional yield are of vital importance for formulating harvesting, milling and forward selling decisions, while at a block scale they provide growers with an understanding of both in-crop variability and total production. Currently, annual cane production estimates are made by quantifying the area of cane grown within a region by visual in-season yield assessments.

Although this method can produce accuracies of up to 95% (A Pitt pers. comm. 2011) it can be influenced by variable climatic conditions such as those experienced in 2010. As such, geographic information systems (GIS) and remote sensing (RS) may offer an additional tool for validating these predictions as well as potentially providing a more accurate seasonally sensitive method of prediction.

GIS and remote sensing in the sugar industry

Geographic information systems (GIS) have been widely adopted by the Australian sugar industry as an essential tool for the recording and managing spatial data (Davis et al., 2007). One such system developed for the Mackay and Burdekin region has greatly increased the integration of mill and productivity datasets, thus enabling greater efficiencies in data retrieval and analysis of client information (Markley et al., 2008).

Similarly, the development of a whole-of-community GIS system by the Herbert River sugar district has created the capacity to record real-time cane harvester operations via GPS, enabling improvements in the coordination and planning of the cane harvest, efficient reporting of
harvest performance and the identification of consignment errors. This information has also been used to improve rail transport infrastructure safety and efficiency (De Lai et al., 2011).

Globally, satellite imagery has been identified as an effective tool for predicting sugar cane yield (Fernandes et al., 2011; Benvenuti and Weill, 2010; Bégué et al., 2010; Simões et al., 2009; Abdel-Rahman and Ahmed, 2008; Bégué et al., 2008; Almeida et al., 2006; Simões et al., 2005; Krishna Rao et al., 2002; and Rudorff and Batista, 1990), although such research has been limited in Australia (Noonan, 1999; Markley et al., 2003; Robson et al., 2011; Robson et al., 2010; Lee-Lovick and Kirchner, 1991).

For the past decade, Mackay Sugar Ltd has been the predominant adopter of satellite imagery as a commercial yield forecasting tool for the Mackay region, utilising yield prediction algorithms derived from SPOT imagery (Markley et al., 2003). The research presented in this paper investigates the development and validation of similar algorithms over three additional Australian growing regions including Bundaberg, Burdekin and Herbert.

**Yield predictions using remote sensing techniques**

The amount of electro-magnetic radiation (EMR) reflected from a sugarcane canopy is positively correlated to the leaf area index (LAI), which in turn may correspond to the amount of biomass within the crop, and therefore yield (Bégué et al., 2010). However, this relationship can be influenced by variations in canopy architecture, foliar chemistry, agronomic parameters and sensor and atmospheric conditions (Abdel-Rahman and Ahmed, 2008). More specifically, variety, crop class (plant or ratoon), date of crop planting or ratooning, duration of harvest period and environmental variability are all factors that have been shown to influence the accuracies of yield prediction algorithms developed from remotely sensed imagery (Zhou et al., 2003; Singels et al., 2005; Inman-Bamber, 1994).

In an attempt to remove influences such as spectral interference or ‘noise’, previous researchers have investigated a number of vegetation indices. The most commonly used Normalised Difference Vegetation Index (NDVI), addresses some measurement errors associated with atmospheric attenuation and shading, however it can saturate in large biomass crops such as sugar cane with a LAI greater than three (Benvenuti and Weill, 2010; Bégué et al., 2010; Xiao, 2005; Xiao et al., 2004b; Xiao et al., 2004a; Huete et al., 2002; Huete et al., 1997). To reduce the effects of saturation, a number of additional indices have been employed including the Green Normalised Difference Vegetation Index (GNDVI) (Gitelson et al., 1996; Benvenuti and Weill, 2010).

Timing of image capture has also been identified to be an important consideration when predicting cane yield, especially when compared to the growth phase of the crop. Sugar cane undergoes three distinct growth phases including germination or establishment and tillering, vegetative development or stalk growth and stabilisation, senescence or maturation (Bégué et al., 2010; Simões et al., 2005; Fernandes et al., 2011; Krishna Rao et al., 2002). During the vegetative growth stage NDVI can increase from 0.15 to 0.7, before remaining relatively stable (if unstressed) during the maturation phase, until harvest (Bégué et al., 2010).

Almeida et al. (2006) identified this time period to be 3–6 months prior to harvest, while Simões et al. (2005) suggested 240 days after planting or ratooning. As well as a stabilisation period of NDVI, a ‘synchronisation’ of NDVI was also observed across various plant and ratoon ages due to climatic factors such as rain and temperature. This synchronisation and stabilisation of NDVI is important as it indicates that there is likely to be an extended window of image capture where variability in the canopies’ spectral response as well as differences across crops is minimalised (Bégué et al., 2010; Almeida et al., 2006; Krishna Rao et al., 2002; Rudorff and Batista, 1990).
Methodology

Study districts
Research was conducted in three climatically distinct Queensland cane growing regions of the Herbert (2107 mm of rainfall annually), the Burdekin (1005 mm) and Bundaberg (930 mm) during the 2010 and 2011 growing seasons.

Satellite imagery and spatial data
During the 2010 and 2011 cane growing seasons, full scene (3600 km²) SPOT 5 satellite images were captured over the Herbert (2 June 2011); the Bundaberg region (10 May 2010 and 27 March 2011); and over the Burdekin region (14 May 2010 and 22 April 2011).

The spectral resolution of SPOT5 imagery is green (0.5–0.59 µm), red (0.61–0.68 µm), near infrared (0.78–0.89 µm) and shortwave infrared (1.58–1.75 µm), with a spatial resolution of 10 m pixels. All SPOT5 imagery used for this research was corrected for top of atmosphere reflectance (TOA) (SPOT Image, 2008) and orthorectified to a corrected base layer.

Block boundary GIS vector layers detailing at tribute tables of agronomic data, including variety, class, total area harvested and tonnes cane harvested, were sourced from either milling or productivity services within each region.

Extraction of spectral information
For all cane blocks within the extent of each SPOT 5 image (Figure 1) spectral information was extracted using the open source software Starspan GUI (Rueda et al., 2005). A 20 m metre buffer was applied to each paddock boundary to ensure the extracted information did not include non cane-specific pixels. Spectral and agronomic information including mill data was exported to a single text file to enable additional analysis.

Vegetation indices
To identify the best correlations between satellite imagery and crop yield (TCH), a number of vegetation indices were examined including the Normalised Difference Vegetation Index (NDVI), Green Normalised Difference Vegetation Index (GNDVI) (equation 1), the Soil Adjusted Vegetation Index (SAVI) and the Two-band Enhance Vegetation Index (EVI_2).

These indices were calculated for every cane block defined by a GIS paddock boundary within each image capture area. Using harvested tonnes of cane per hectare (TCH) supplied by the respective mills, the indices that provided the highest correlation coefficient were identified.

For all regions, the GIS attribute data were used to separate the spectral information on the basis of variety, crop class (plant or ratoons) and age of crop, in an attempt to improve the correlations.

\[ \text{GNDVI} = \frac{P_{\text{NIR}} - P_{\text{GREEN}}}{P_{\text{NIR}} + P_{\text{GREEN}}} \]  \hspace{1cm} (1)
where \( P_{\text{GREEN}} \) and \( P_{\text{NIR}} \) are the TOA reflectance values measured in the green and near infrared spectral bands.

Additionally, predictions of average yield were made for 3544 (2010) and 3,824 (2011) cane crops within the Bundaberg region using an algorithm derived from the linear relationship between 2008 and 2010 crop yield and corresponding SPOT5 data (Robson et al., 2011) (equation 2).

The accuracy of this algorithm was also evaluated against point source locations within a single crop and validated within field measurements. Sampling coincided with the commercial
harvest of the crop and consisted of 5 m linear cane rows hand cut at replicated locations representing high, medium and low GNDVI values, located with a non-differential GPS unit.

GNDVI yield prediction algorithm (Bundaberg)  
\[ y = 3.1528 \times \exp(5.6973 \times x) \]  
(2)

where \( y \) = predicted average yield (TCH) and \( x \) = average GNDVI value extracted from TOA SPOT5 image. (n = 150 crops)

A similar prediction was also undertaken for 4999 cane crops grown within the Burdekin region (2011 season) using an algorithm derived from the correlation between 2010 Burdekin crop yields and corresponding 2010 imagery (equation 3).

GNDVI yield prediction algorithm (Burdekin)  
\[ y = 12.691 \times \exp(3.8928 \times x) \]  
(3)

where \( y \) = predicted average yield (TCH) and \( x \) = average GNDVI value extracted from TOA SPOT5 image. (n = 4573 crops)

Fig. 1—SPOT5 images captured over each growing region (a) Burdekin, (b) Herbert and (c) Bundaberg and (d) closer view of agronomic information provided within the GIS attribute table.
Results

The initial aim of this research was to develop a generic image-based yield algorithm for all Queensland growing regions that was non-specific to variety, growth stage, and even seasonal variability.

However, it was quickly identified that one algorithm would be insufficient due to the large range of varieties planted as well as variation in growing and climate conditions across each region. As such, each growing region was evaluated separately.

Bundaberg

The correlation between TCH and spectral data extracted for 3824 cane crops grown within the Bundaberg region during 2011 (including 26 varieties with nine ratoon stages, plant, replant and standover classes) was promising with all vegetation indices producing correlation coefficients above 0.6, with GNDVI producing the highest \( r = 0.63 \) (Table 1). This correlation was further improved by segregating the data into plant and ratoon classes.

Table 1—Correlation coefficients \( (r) \) identified between TCH and individual spectral bands/vegetation indices for the Bundaberg district, 2011 growing season.

<table>
<thead>
<tr>
<th>Band/VI</th>
<th>All blocks</th>
<th>Plant cane</th>
<th>1st Ratoon</th>
<th>2nd Ratoon</th>
<th>3rd Ratoon</th>
<th>Variety Q208</th>
<th>Variety KQ228</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>0.20</td>
<td>0.23</td>
<td>0.17</td>
<td>0.15</td>
<td>0.16</td>
<td>0.51</td>
<td>0.50</td>
</tr>
<tr>
<td>Red</td>
<td>0.42</td>
<td>0.40</td>
<td>0.44</td>
<td>0.47</td>
<td>0.48</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>NIR</td>
<td>0.58</td>
<td>0.70</td>
<td>0.62</td>
<td>0.60</td>
<td>0.57</td>
<td>0.64</td>
<td>0.60</td>
</tr>
<tr>
<td>SWIR</td>
<td>0.37</td>
<td>0.28</td>
<td>0.37</td>
<td>0.38</td>
<td>0.38</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.61</td>
<td>0.67</td>
<td>0.68</td>
<td>0.69</td>
<td>0.66</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>GNDVI</td>
<td>0.63</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
<td>0.66</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>SAVI</td>
<td>0.61</td>
<td>0.71</td>
<td>0.66</td>
<td>0.65</td>
<td>0.62</td>
<td>0.68</td>
<td>0.64</td>
</tr>
<tr>
<td>EVI 2</td>
<td>0.61</td>
<td>0.72</td>
<td>0.66</td>
<td>0.65</td>
<td>0.62</td>
<td>0.68</td>
<td>0.64</td>
</tr>
</tbody>
</table>

The stability of correlation across varieties for both GNDVI and NDVI is important as it indicates that a ‘generic’ algorithm which is not cultivar specific may be possible for the Bundaberg region, a finding that supports initial results presented by Robson et al. (2011).

To further investigate the consistency of GNDVI values across varying classes, variety and seasons, the 2011 data (n = 3824) were overlayed with similar data used to develop the 2008–2010 algorithm (n= 150) (Figure 2). From Figure 2 it can be seen that although there is variance around the line of best fit, the overall trend between GNDVI and TCH is relatively consistent across the two data sets.

The calculation and then subsequent substitution of average GNDVI value (0.567) from 3544 crops grown during 2010 into the 2008–2010 algorithm produced an estimated average yield of 78.1 TCH, highly comparable to the actual milled yield of 81.8 TCH (95.5% accurate).

For the 2011 harvest season, an average yield of 80.1 TCH was predicted following the substitution of average GNDVI value (0.57) sourced from 3824 crops into 2008–2010 algorithm. This prediction was within 9% of actual milled harvest yield of 73.3 TCH (91.5%).

The accuracy of overall prediction, and the fact that the data was not segregated into variety or growth stage, indicates that this technology has the potential to predict regional cane yield within the Bundaberg growing region, a result that differs from previous findings by Lee-Lovick and Kirchner (1991).
Fig. 2—Correlation between GNDVI (SPOT 5) and TCH from Bundaberg cane blocks during the 2008–2010 (black points) and 2011 (grey points) seasons.

To coincide with regional forecasting, the development of such an algorithm offers the potential for predicting individual crop yield as well as the derivation of surrogate yield maps, prior to harvest. To test this, the accuracy of the GNDVI yield algorithm was also evaluated over point source locations within an individual Bundaberg cane crop (area 18.7 ha, var. KQ228) harvested 25 July 2011 (Figure 3). This analysis identified a strong relationship between predicted yield from a SPOT5 image captured on the 27 March 2011, and final yield measured on the 25 July 2011.

Fig. 3—(a) False colour image of Bundaberg cane crop (area 18.7 ha, var. KQ228) harvested 25 July 2011, with yellow markers indicating field sampling locations, (b) measured verse predicted cane yield at the locations identified in (a), (c) Classified yield map generated by applying the 2008–2010 GNDVI yield algorithm to the SPOT5 (27 March 2011) pixel values.
The generation of a classified yield map (Figure 3c) and subsequent accurate prediction of total crop yield from the average crop GNDVI value (predicted of 92 TCH, actual delivered yield of 88.7 TCH) further supports the potential of this technology for producing in-season yield variability maps.

**Burdekin**

For the Burdekin region, correlation coefficients produced between TCH and SPOT5 derived vegetation indices (captured 14 May 2010) for all 4573 crops were relatively consistent ranging from r=0.39 (NDVI) to r=0.44 (SAVI and EVI_2) (Table 2).

This correlation remained relatively unchanged when data was segregated into the different cultivars Q208 and KQ228, indicating that a yield prediction algorithm for this region may not be required to be cultivar specific.

<table>
<thead>
<tr>
<th>Band/VI</th>
<th>All blocks</th>
<th>Plant cane</th>
<th>1st Ratoon</th>
<th>2nd Ratoon</th>
<th>3rd Ratoon</th>
<th>Variety Q208</th>
<th>Variety KQ228</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>0.12</td>
<td>0.08</td>
<td>0.11</td>
<td>0.10</td>
<td>0.18</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>Red</td>
<td>0.18</td>
<td>0.19</td>
<td>0.15</td>
<td>0.17</td>
<td>0.11</td>
<td>0.27</td>
<td>0.23</td>
</tr>
<tr>
<td>NIR</td>
<td>0.41</td>
<td>0.40</td>
<td>0.39</td>
<td>0.28</td>
<td>0.22</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>SWIR</td>
<td>0.19</td>
<td>0.18</td>
<td>0.12</td>
<td>0.11</td>
<td>0.06</td>
<td>0.28</td>
<td>0.10</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.39</td>
<td>0.42</td>
<td>0.35</td>
<td>0.29</td>
<td>0.21</td>
<td>0.39</td>
<td>0.35</td>
</tr>
<tr>
<td>GNDVI</td>
<td>0.43</td>
<td>0.44</td>
<td>0.40</td>
<td>0.29</td>
<td>0.21</td>
<td>0.41</td>
<td>0.35</td>
</tr>
<tr>
<td>SAVI</td>
<td>0.44</td>
<td>0.44</td>
<td>0.41</td>
<td>0.31</td>
<td>0.23</td>
<td>0.39</td>
<td>0.37</td>
</tr>
<tr>
<td>EVI2</td>
<td>0.44</td>
<td>0.43</td>
<td>0.41</td>
<td>0.31</td>
<td>0.23</td>
<td>0.39</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Unlike the Bundaberg analysis however, there was a noticeable drop in the correlation coefficients with ratoon age, especially 2nd and 3rd ratoon (Table 2). This variation indicates that an algorithm that is not crop class specific may be inaccurate, especially when predicting point source yield within individual crops such as that displayed in Figure 3.

At a regional level the predicted average crop yield of 4999 crops grown during the 2011 season using the 2010 algorithm (equation 3) was 99% (actual average yield of 120 TCH, predicted 118.8 TCH). Although highly accurate, the result is not considered robust, due to the large spread of data (r^2 = 0.07) produced particularly with standover crops (grey markers in Figure 4). This predictive accuracy will however be further validated during the 2011–2012 season.

**Herbert**

The initial correlation between TCH and spectral data (SPOT 5 captured 2 June 2011) for 8596 cane crops grown in the Herbert region (including 53 varieties, multiple ratoon stages, plant, replant and standover) was poor (Table 3).

This result is believed to be attributed to severe climatic conditions experienced towards the end of 2010 and start of 2011. The Herbert region had around 25% of the 2011 crop as ‘stand over’ i.e. not harvested from 2010, with the remainder exhibiting various degrees of flood related damage. The removal of standover blocks did improve the coefficients.

The highest regression coefficients were identified by segregating the data into crop class and then variety, for example KQ228 plant crop r = 0.65 (GNDVI).
Fig. 4—Correlation between GNDVI (SPOT 5) and TCH from 2011 Burdekin cane blocks with black points indicating non-standover crops and grey points indicating standover crops.

Table 3—Correlation coefficients (r) identified between TCH and individual spectral bands/vegetation indices for the Herbert district.

<table>
<thead>
<tr>
<th>Herbert District</th>
<th>Variety Q208 Plant</th>
<th>Variety Q208 1st Ratoon</th>
<th>Variety KQ228 Plant</th>
<th>Variety KQ228 1st Ratoon</th>
<th>Q200 Plant</th>
<th>Q200 1st Ratoon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band/VI</td>
<td>All blocks</td>
<td>Standover removed</td>
<td>Plant cane</td>
<td>1st Ratoon</td>
<td>2nd Ratoon</td>
<td>3rd Ratoon</td>
</tr>
<tr>
<td>Green</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.15</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>Red</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>0.05</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>NIR</td>
<td>0.23</td>
<td>0.46</td>
<td>0.54</td>
<td>0.45</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>SWIR</td>
<td>0.38</td>
<td>0.42</td>
<td>0.37</td>
<td>0.35</td>
<td>0.46</td>
<td>0.53</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.22</td>
<td>0.40</td>
<td>0.47</td>
<td>0.34</td>
<td>0.31</td>
<td>0.37</td>
</tr>
<tr>
<td>GNDVI</td>
<td>0.23</td>
<td>0.45</td>
<td>0.54</td>
<td>0.42</td>
<td>0.42</td>
<td>0.44</td>
</tr>
<tr>
<td>SAVI</td>
<td>0.23</td>
<td>0.45</td>
<td>0.54</td>
<td>0.42</td>
<td>0.37</td>
<td>0.46</td>
</tr>
<tr>
<td>EVI 2</td>
<td>0.24</td>
<td>0.46</td>
<td>0.54</td>
<td>0.42</td>
<td>0.37</td>
<td>0.46</td>
</tr>
</tbody>
</table>

These results indicate that for the accurate prediction of yield within the Herbert region a number of algorithms representing different growth stages and even varieties may be required. This hypothesis requires further validation over subsequent growing seasons, particularly seasons that are not influenced by extreme climatic conditions.

Discussion
The undertaking of this research over the three distinct growing regions was highly beneficial considering the array of success identified. Results from the Bundaberg region, and to a lesser extent the Burdekin, indicated that a ‘generic’ yield prediction algorithm may be developed and then used to accurately predict regional production and even within crop yield variability.
Although improved correlations were produced following the segregation of data into different groups such as crop class (Burdekin) and variety (Herbert) some consideration has to be made on the number of algorithms developed. In regards to variety, fifty-three were planted in the Herbert, twenty-six in Bundaberg and nineteen in the Burdekin in the years encompassed by this study. If other variables such as the segregation of regions into smaller climate-driven micro regions or crop class are also accounted for then the number of algorithms required would grow substantially.

One method to address this may be to develop algorithms for only the dominant varieties. For example, only three varieties (of nineteen) in the Burdekin accounted for 83% of the total number of planted blocks. Alternatively, varieties could be categorised into groups based on their spectral signatures. The use of multiple algorithms may increase the flexibility of the predictive models for the season upon which it is applied, allowing it to better compensate for changing percentages of varieties and classes throughout a district and the addition of new varieties.

In the past, the adoption of remote sensing as a yield prediction tool by the Australian sugar industry has been severely hampered by a number of limitations including: a lack of yield data from the mills due to privacy issues, an extended harvesting period resulting in a patchwork of different varieties, growth and ratoon stages in close proximity, seasonal or climatic variability, constant cloud clover, insufficient computational demands for image processing, a shortage of skilled analysts and concerns regarding the benefit-cost of adopting the technology.

Irrespective of these concerns the research presented in this paper identified satellite imagery and associated GIS data as useful tools for supporting current methods of yield forecasting, with the potential of improving both regional and in-crop yield predictions in the future following further validation.

Acknowledgements

The authors would like to acknowledge SRDC for providing funding for this research as well as those growers and industry partners whom have collaborated, particularly Bundaberg Sugar Pty Ltd, Herbert Resource Information Centre, Sucrogen and Farmacist Pty Ltd.

REFERENCES


