IDENTIFICATION OF COMMERCIAL BLOCKS OF OUTSTANDING PERFORMANCE OF SUGARCANE USING DATA MINING


PAULO R. PELOIA†, LUIZ H. A. RODRIGUES1

ABSTRACT: In order to achieve more efficient agricultural production systems, studies relating to the patterns of influence factors on commercial blocks of outstanding performance can be performed to assist management practices. The performance is considered to be the difference between the yield of a given block and the average yield of the homogeneous group that it belongs to. The methods available to identify these outstanding blocks are usually subjective. The aim of this study was to propose an objective and repeatable approach to identify outstanding performance blocks. The proposed approach consisted of performance determination, using regression trees, and the classification of these blocks by k-means clustering. This approach was illustrated using a sugarcane model. The main factors influencing the tonnes of cane per hectare (TCH) and total recoverable sugar (TRS) yields were found to be crop age and water availability during ripening, respectively. These were used to create potential yield groups, and blocks with high and low performance were identified. The proposed approach was found to be valid in the identification of outstanding sugarcane blocks, and it can be applied to different crops or in the context of precision agriculture.

KEYWORDS: clustering, regression tree, yield variability.

INTRODUCTION

In the search for more efficient agricultural production systems, research studies have been performed with the aim to identify patterns in environmental and management factors that influence the yield variability from block-to-block in diverse crops (ANDRIANASOLO et al., 2014; RENAUD-GENTIÉ et al., 2014; ZHANG, 2012; TITTONELL et al., 2008).

This type of study, focusing on specific blocks of outstanding performance, can help in the adoption of more appropriate management procedures. The patterns associated with blocks of superior performance can be used to confirm that certain measures result in greater productivity, while low performance patterns indicate the need for improvement in management, or the identification of specific conditions that should be avoided (TANUSKA et al., 2012; VAZAN et al., 2011). The performance of a given block is considered to be the difference between the yield of the block and the average yield of the homogeneous group that it belongs to.

Despite the benefits of this type of analysis (LAWES & LAWN, 2005), the methods used to determine the potential yield groups and the identification of outstanding blocks are subjective, therefore, it is highly dependent of the interpretation of each specialist (TITTONELL et al., 2008).

Given the large number of variables submitted for analysis in this type of study (TANUSKA et al., 2012; VAZAN et al., 2011; LAWES & LAWN, 2005), data mining techniques are a promising alternative due to their ability to address complex databases with noise (HAN et al., 2012), a common feature of commercial block databases. In particular, the regression tree induction technique is able to identify and rank the factors that influence a production system (WITTEN & FRANK, 2011), and may be useful for establishing potential yield groups to determine the performance of blocks.

The same application can also be extended to precision agriculture, where instead of analysing the variability between specific block-to-block, the variability between different areas within the same block are considered. A study by SOUZA et al. (2010) demonstrated the use of a
decision tree to identify the main factors affecting the yield variability within a block. The main factor was identified to be the declivity of the land. Although this factor is not manageable, this approach defined the potential yield groups and identified outstanding points within the block.

Due to the benefits associated with analysing the patterns of blocks of outstanding performance, and the lack of an objective approach for their identification, the aim of this study was to propose an objective and repeatable approach for the identification of sugarcane blocks of outstanding performance.

MATERIAL AND METHODS

The computational package used in this research was JMP version 11.2 (SAS INSTITUTE INC., 2013).

The approach was conducted using a database which included yield values for sugarcane (*Saccharum officinarum*) blocks, in addition to the specific influencing factors related to this crop, provided by a production unit located in the western region of São Paulo, Brazil. This database contained a total of 2255 records, and each record corresponded to the data from one block, collected by the plant during normal working conditions. In addition to the goal attributes, tonnes of cane per hectare (TCH) and total recoverable sugar (TRS), there were another 65 descriptor attributes related to the environmental and management conditions. The approach described below independently addressed the two goal attributes, therefore, the whole process was conducted independently for TCH and TRS.

The proposed approach for the identification of blocks of outstanding performance consisted of several steps to determine the most relevant performance factors (attributes) for each goal (TCH and TRS) and the clustering of blocks, based on these factors.

The aim of determining the performance is to equalise the potential yield of the different blocks, such that they can be compared while taking into account the environment in which they were grown.

The performance of each block, related to the TCH and TRS attributes, is calculated by the difference between its yield and the average yield of the potential yield group from which the block belongs, divided by the standard deviation of the group, according to [eq. (1)]. The two performance variables for TCH and TRS were termed ZTCH and ZTRS, respectively. In this way, the average ZTCH or ZTRS of each group had a value of zero and the standard deviation was equal to one.

\[
Z_{PROD_{ij}} = \frac{PROD_{ij} - PROD_j}{S_j}
\]

where,

- \(Z_{PROD_{ij}}\) was the standard performance of block \(i\) in the potential yield group \(j\);
- \(PROD_{ij}\) was the yield of block \(i\) in the potential yield group \(j\);
- \(PROD_j\) was the average of the potential yield group, and
- \(S_j\) was the standard deviation of the group of blocks with the same potential yield group.

To determine the potential yield groups, we used the regression tree induction for the particular goal attribute, thereby identifying and ordering the most influential factors on the production system. We then selected the factors of greatest importance. Each combination of levels for the selected factors corresponded to the potential yield group.

The regression tree is a non-parametric modelling technique that can be used to explain the responses of a dependent variable (goal attribute) based on a series of continuous or categorical
Identification of commercial blocks of outstanding performance of sugarcane using data mining

independent attributes. This technique makes recursive divisions in the data based on the independent variables, in order to increase homogeneity in the response of the goal attribute (WITTEN & FRANK, 2011).

After the performance for each block had been determined, the blocks were grouped in order to identify the outstanding blocks. We used the k-means clustering method, which minimises the difference between points within the same group and maximises the difference between the k groups defined by the user (FORGEY, 1965). This process begins with determining the k centres of random groups, then each of the records is allocated to the centre group that is most closely located. The centres, representing the average records of the groups, are recalculated, and the records are then moved if the recalculated centre of another group is closer. This process continues until stable.

Three (k = 3) clusters were selected, and the algorithm was applied exclusively to the performance data, ignoring the descriptor attributes. The clusters formed were specified as either Low, Medium or High performance, according to their average value. The Low and High clusters were considered to have the most outstanding performance, while the blocks of the Average cluster were not considered to be outstanding.

Other values of k clusters could be used in the k-means method (MERTLER & VANNATTA, 2013), thereby identifying different groups of outstanding blocks, which can also help with the subsequent search for patterns. For example, two or five clusters, generating High/Low and Very High/High/Average/Low/Very Low groups, respectively. For a k value greater than 4, the use of the Kruskal-Wallis non-parametric test is recommended (GLANTZ, 2011) to verify if the different k values are indeed generating distinct clusters.

RESULTS AND DISCUSSION

The relative importance of the influence factors on TCH, obtained by the regression tree, showed an effect of crop age compared to the other attributes analysed (Table 1). This attribute was found to be responsible for 56% of the total variability explained by the regression tree, whereas the second most important factor, soil texture, was only responsibly for 4% of the variability. For this reason, crop age was selected for the preparation of potential yield groups for TCH.

<table>
<thead>
<tr>
<th>Influence factor</th>
<th>Unit</th>
<th>Relative importance</th>
<th>Accumulated importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crop age</td>
<td>Year</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>Soil texture</td>
<td>Clay (%)</td>
<td>0.04</td>
<td>0.60</td>
</tr>
<tr>
<td>Solar radiation (budding)</td>
<td>MJ m⁻² day⁻¹</td>
<td>0.04</td>
<td>0.64</td>
</tr>
<tr>
<td>Drought (maximum growth)</td>
<td>Days</td>
<td>0.03</td>
<td>0.67</td>
</tr>
<tr>
<td>Degree days (ripening)</td>
<td>°C day⁻¹</td>
<td>0.03</td>
<td>0.70</td>
</tr>
<tr>
<td>Borer (Diatrea saccharalis)</td>
<td>%</td>
<td>0.03</td>
<td>0.72</td>
</tr>
<tr>
<td>Water supply (maximum growth)</td>
<td>mm day⁻¹</td>
<td>0.02</td>
<td>0.75</td>
</tr>
<tr>
<td>Solar radiation (tillering)</td>
<td>MJ m⁻² day⁻¹</td>
<td>0.02</td>
<td>0.77</td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>0.23</td>
<td>1.00</td>
</tr>
</tbody>
</table>

There is known to be a declining trend in sugarcane yield with increasing crop age (INMAN-BAMBER, 2013). However, the magnitude of the difference of this factor in relation to the other ones is unexpected, which could have possibly remained undetected without the use of the regression tree or other more elaborate analysis tools (e.g. multiple regression).

As shown in Figure 1a, the performance results for sugarcane yield per hectare (ZTCH), and the clusters, illustrate the potential yield groups created for each level of crop age and its average (equal to zero), in addition to its classification as either Low, Medium or High. The Low cluster was found to represent 24.1% of the total number of blocks, while the High cluster represented 24.2%.
The ZTCH values that separated the Low and High performance blocks were -0.64 and 0.65, respectively. We noted that these values were valid for each of the potential yield groups. The participation percentage of both the Low and High clusters increased with increasing crop age, according to the results obtained from the Cochran-Armitage test for trend (Prob. > Z < 0.001 and Prob. > Z = 0.0386, respectively). In general, the participation in the initial years was found to be 18%, while for the more advanced years this value was 32%.

**FIGURE 1.** Values of (a) yield performance of sugarcane per hectare (ZTCH) and (b) yield of sugarcane per hectare (TCH), according to the levels of crop age and clusters.

Figure 1b shows the clusters obtained for ZTCH for the TCH attribute. We observed that there was no longer a fixed value for the segmentation of outstanding blocks, as identified previously. For example, planting blocks with an age of 18 months, classified as having Low performance (average yield of 63.1 tonnes ha\(^{-1}\)), has a TCH value above the blocks of the 5th year, which is beyond that classified as High (57.3 tonnes ha\(^{-1}\)). This is the main benefit of the proposed approach, as the objective and repeatable identification of outstanding blocks takes into account the environmental conditions in which they developed.

The observation that the High and Low clusters covered a similar number of records may be explained by the approximately normal distribution of TCH data (kurtosis of 0.09 and asymmetry of 0.05, where these values are equal to zero in a perfect normal distribution) and the clustering by k-means. As the data were found to be normally distributed, we expected the cut-off points between the clusters to be symmetrical, in addition to being close to that found previously using a box-plot method (GLANTZ, 2011).

Table 2 shows the results of the influence factors on TRS. The ‘water availability’ attribute during the ripening phase had the highest influence on TRS, with 51% of the variability explained by the model, also confirming a widely accepted concept in the literature (VAN HEERDEN et al., 2013). Therefore, this was the factor selected for preparation of the potential yield groups related to TRS.
TABLE 2. Relative and accumulated importance of the influence factors on the yield of total recoverable sugar (TRS) obtained by the regression tree.

<table>
<thead>
<tr>
<th>Influence factor</th>
<th>Unit</th>
<th>Relative importance</th>
<th>Accumulated importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water availability (ripening)</td>
<td>mm day(^{-1})</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Solar radiation (ripening)</td>
<td>MJ m(^{2}) day(^{-1})</td>
<td>0.07</td>
<td>0.58</td>
</tr>
<tr>
<td>Season</td>
<td>Year</td>
<td>0.06</td>
<td>0.64</td>
</tr>
<tr>
<td>Negative degree days (ripening)</td>
<td>°C day(^{-1})</td>
<td>0.05</td>
<td>0.69</td>
</tr>
<tr>
<td>Type of planting</td>
<td>Manual/mechanised</td>
<td>0.03</td>
<td>0.72</td>
</tr>
<tr>
<td>Degree days (ripening)</td>
<td>°C day(^{-1})</td>
<td>0.03</td>
<td>0.75</td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>0.25</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Unlike the ‘crop age’ attribute (selected for the formation of potential yield groups for TCH), which has nominal levels, the ‘water availability’ factor is continuous. Therefore, we considered that the number of groups was equal to the number of different values belonging to this attribute. Thus, the average TRS value for each group was determined by [eq. (2)] (Prob. > F < 0.0001).

\[
\overline{ATR}_j = 133.77 - 3.02 \overline{DHM}_j
\]  

where,

\(\overline{ATR}_j\) was the average TRS value for each group \(j\) of blocks (kg sugar per tonne of sugarcane), and

\(\overline{DHM}_j\) was the water availability in the ripening phase for the same group of blocks, (mm.day\(^{-1}\)).

ZTRS values lower than -1.0 and greater than 0.31 represented the groups of Low and High performance, respectively (Figure 2a). The High cluster represented 41.7% of the records and the Low cluster represented 14.4%. Unlike the cut-off points found for ZTCH, the cut-off points for ZTRS covered a distinct percentage of records, which may be explained by asymmetry to the left (kurtosis of 1.28 and asymmetry of -0.75). This confirms the ability of the k-means method to adjust itself to the distribution of the data (MERTLER & VANNATTA, 2013).

![FIGURE 2. Values of (a) yield performance of total recoverable sugars (ZTRS) and (b) yield of total recoverable sugars (TRS), according to the water availability during the ripening period and the clusters.](image)
The impact of the cluster formation on TRS (Figure 2b), considering the ZTRS, showed similar results to that observed for TCH, despite the selected influence factor being continuous.

The participation of clusters in the total records varied significantly. This highlights the importance for the method used to evaluate blocks of outstanding performance to consider the distribution of the data at the moment of classification (in the case of k-means) instead of a fixed arbitrary value. Using clusters of outstanding blocks that are more homogeneous is expected to facilitate the search for beneficial patterns (HAN et al., 2012; FERRARO et al., 2009).

CONCLUSIONS

The proposed approach was found to be valid for the identification of sugarcane blocks of outstanding performance in an objective and repeatable manner. Therefore, this approach should be tested in other crops and in the context of precision agriculture, in which, instead of investigating the cause of variability from one block to another, we could attempt to understand the variation in georeferenced points within a particular block.

The results from the regression trees for TCH and TRS showed a significant influence of factors which are not easily controlled or manageable, which were used for equalising the potential yield. If the regression trees had identified factors that are easily altered by the adoption of direct management or planning techniques, this analysis would assist in continuous improvement of the production system.

ACKNOWLEDGEMENTS

The authors thank the BioEn Fapesp and Odebrecht Agroindustrial companies for financial support (process Nº. 2012/50049-3) and the São Paulo Research Foundation (FAPESP; Brazil). The authors also thank the professionals who were interviewed for this study.

REFERENCES


